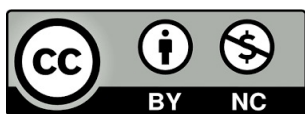Jorge Castillo Mateo

Modelos estocásticos para el análisis espaciotemporal de extremos. Aplicaciones al análisis de cambio climático

Director/es

Gelfand, Alan E.
Cebrián Guajardo, Ana C.

Universidad
Zaragoza
1542

Tesis Doctoral

MODELOS ESTOCÁSTICOS PARA EL ANÁLISIS
ESPACIOTEMPORAL DE EXTREMOS.
APLICACIONES AL ANÁLISIS DE CAMBIO
CLIMÁTICO

Autor

Jorge Castillo Mateo

Director/es

Gelfand, Alan E.
Cebrián Guajardo, Ana C.

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Matemáticas y Estadística

2023

Tesis Doctoral

Modelos estocásticos para el análisis espacio-temporal de extremos. Aplicaciones al análisis de cambio climático

Autor

Jorge Castillo Mateo

Directores

Ana C. Cebrián Guajardo
Alan E. Gelfand

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Matemáticas y Estadística
2023

Stochastic models for the spatio–temporal analysis of extremes

Applications to the analysis of climate change

Author

Jorge Castillo-Mateo

Department of Statistical Methods
University of Zaragoza

Supervisors

Ana C. Cebrián Guajardo

Alan E. Gelfand

Acknowledgements

Gracias.

A Ana, por acompañarme y dirigirme en este episodio llamado Tesis. Por ser en su momento mi principal tutora allá por el 2018, cuando yo aún era un alumno de grado preguntando por un *Trabajo fin de Grado*. Por tratarme como a uno más del equipo desde el primer día y por saber lo que se siente al exponerte por primera vez al público. Por su manera de tener los pies en la tierra, y por estar disponible en cualquier momento cuando es necesario.

A Jesús, también director de esta Tesis, a pesar de que no pueda ser reconocido como tal en los papeles. Porque su tutela ha sido constante y su paciencia infinita, tengo mis dudas de que hubiese sobrevivido a toda la burocracia académica sin su ayuda. Por abrirme las puertas a nuevos proyectos. Por sus consejos sobre la montaña. Por sus apretones de mano antes y después de mis viajes. Por sus ocurrencias propias de un ser de otro planeta—ojalá algún día poder visitarlo. Por su optimismo, interés, cercanía y generosidad, pues hacen que no existan los límites entre lo profesional y personal. En definitiva, por ser secretario, consejero y, sobretodo, amigo.

A Alan, por aventurarse a codirigir una Tesis en España aunque eso suponga no parar de firmar "autógrafos". Por enseñarme que el mundo es pequeño y que la otra parte del globo no está tan lejos como pensaba. Por darme la oportunidad de adentrarme en la Duke University en distintas estancias. Y al margen de lo brillante de sus ideas, por inspirarme con la ilusión desmedida que sigue demostrando por la investigación después de tantos años.

A Ana, Jesús, Alan, y Jesús Abaurrea—como el *Zaragoza team*—, por aportarme una visión tan diversa sobre la investigación y sobre el mundo. Por su guía, su implicación, sus ánimos, y su forma de hacer las cosas. Esta Tesis no hubiera sido posible sin su gran apoyo y ayuda, especialmente en los momentos menos buenos.

A Erin por acogerme en la NCSU, y por compartir conmigo su tiempo y conocimiento. A Sudipto por su excepcional trato en mi visita a la UCLA. A Sujit por su disposición para ayudar en distintos momentos de esta Tesis. A Lajos por sus

inesperados pero enriquecedores y divertidos correos. También a todos aquellos con los que he tenido la oportunidad de interactuar en mis distintos congresos y estancias, especialmente en los acogedores congresos de la SEB.

A mi tío Jesús, por iniciarme en la investigación y por hacerme ver que las matemáticas también se estudian. Por esos *paper*'s, *journal*'s y *ResearchGate*'s que me resultaban incomprensibles. Por permitirme nada más comenzar el grado colaborar en el que sería mi primer paper científico publicado en el journal *Engineering Geology* (Mateo-Lázaro et al., 2016) y que puede consultarse en ResearchGate. Por haber tenido la oportunidad de colaborar con él—por el momento—en otros cinco trabajos que no aparecen en esta Tesis (Mateo-Lázaro et al., 2018, 2019; Mateo-Lázaro et al., 2020; Mateo-Lázaro et al., 2022; Castillo-Mateo, Asín, Cebrián, Mateo-Lázaro et al., 2023).

A mis compañeros de departamento, y en especial, a todo el grupo de investigación de Modelos estocásticos, pues sin ellos, ninguna de mis estancias ni viajes a congresos habrían sido posibles. Especialmente a Jesús Abaurrea, con quien además de muchos trabajos he compartido despacho. A Tomás, porque si mi tío me hizo decantarme por las matemáticas, sus clases magistrales de *Técnicas de Regresión* me hicieron decantarme por la estadística, clases en las que durante tres años he tenido la gratificante oportunidad de compartir docencia. No puedo olvidarme de Miguel, porque el día que lo conocí me dió su número por si necesitaba algo, por mi primer congreso en Alcoy allá por el 2019, y por sus cautivadoras y elocuentes historias.

A mi familia, por su amor y apoyo incondicional. A mis amigos. En especial a Jorge, por ser mi otro hermano. A mis gatos, por hacerme compañía en las noches de desvelo. A todos mis tíos y primos, porque cuando os interesáis por mi trabajo me dáis fuerza para seguir adelante. A mis abuelos, que siempre me han querido. En especial a mi *yaya* Basi, porque no hay nada más valioso que el tiempo contigo. A mi hermano, por inculcarme desde pequeño y compartir conmigo el deporte como forma de vida. Junto a él también a Laura, por esas reuniones de cuarto en cuarto después de cada reunión familiar. A mi padre, por enseñarme a esforzarme y a trabajar duro.

A mi madre, que es mi madre y con eso lo digo todo. Por estar ahí cada uno de los 1095 días de escritura de esta memoria, especialmente en aquellos días en los que he estado más lejos. Por su manera de criarme, enseñarme, y por su amor infinito. A ella esta Tesis le ha robado mucho tiempo de estar conmigo.

<div align="right">Entre Carolina del Norte y Zaragoza, año 2023.</div>

Funding

Compendium of publications

This Thesis is a compendium of 8 research papers previously published or accepted for publication. Hereafter, the papers that constitute the main body of this Thesis are fully referenced:

1. **Castillo-Mateo, J.**, Lafuente, M., Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2022). Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain. *Journal of Agricultural, Biological and Environmental Statistics, 27*(3), 487–505. https://doi.org/10.1007/s13253-022-00493-3 [arXiv:2201.01687]

2. Cebrián, A. C., Asín, J., Gelfand, A. E., Schliep, E. M., **Castillo-Mateo, J.**, Beamonte, M. A., & Abaurrea, J. (2022). Spatio-temporal analysis of the extent of an extreme heat event. *Stochastic Environmental Research and Risk Assessment, 36*(9), 2737–2751. https://doi.org/10.1007/s00477-021-02157-z

3. **Castillo-Mateo, J.**, Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2023). Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics, 17*(3), 2305–2325. https://doi.org/10.1214/22-AOAS1719

4. **Castillo-Mateo, J.**, Gelfand, A. E., Asín, J., Cebrián, A. C., & Abaurrea, J. (*in press*). Bayesian joint quantile autoregression. *TEST*. [arXiv:2305.19080]

5. Cebrián, A. C., **Castillo-Mateo, J.**, & Asín, J. (2022). Record tests to detect non stationarity in the tails with an application to climate change. *Stochastic Environmental Research and Risk Assessment, 36*(2), 313–330. https://doi.org/10.1007/s00477-021-02122-w

6. **Castillo-Mateo, J.** (2022). Distribution-free changepoint detection tests based on the breaking of records. *Environmental and Ecological Statistics, 29*(3), 655–676. https://doi.org/10.1007/s10651-022-00539-2 [arXiv:2105.08186]

7. **Castillo-Mateo, J.**, Cebrián, A. C., & Asín, J. (2023). **RecordTest**: An R package to analyse non-stationarity in the extremes based on record-breaking events. *Journal of Statistical Software, 106*(5), 1–28. https://doi.org/10.18637/jss.v106.i05

8. **Castillo-Mateo, J.**, Cebrián, A. C., & Asín, J. (2023). Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021. *Atmospheric Research, 293*, 106934. https://doi.org/10.1016/j.atmosres.2023.106934

> *"All models are wrong, but some are wronger than others."*

> Noel A. C. Cressie, in Spatial Statistics 2023, Boulder CO, USA; rewording the common aphorism in statistics, *"All models are wrong, but some are useful."*, generally attributed to George E. P. Box

Abstract

Evidence of climate change induced from the increasing concentration of greenhouse gases in the atmosphere suggests an increment not only in mean temperature, but also more frequent warm days, heat waves, and extreme and record-breaking temperatures. This Thesis aims to develop novel stochastic models for the spatiotemporal analysis of such extreme temperature events. Additionally, the purpose of this project is to provide insights to a specific geographic region, around the Comunidad Autónoma de Aragón in Spain and peninsular Spain, about its state of the climate in terms of temperatures.

The datasets for this project contain series of over 60 years of daily maximum temperature observations at 18 and 40 meteorological stations around Aragón and Spain, respectively. Most of the proposed methods for modeling such data fall within the Bayesian hierarchical framework for point-referenced spatiotemporal data. Fitting of these models proceeds through specifically designed Markov chain Monte Carlo algorithms, and a model-based perspective offers full and exact inference with proper assessment of uncertainty. As an alternative to the asymptotic extreme value theory, the proposed spatiotemporal models include mean and—multiple and joint—quantile autoregressive in time specifications with fixed and random effects modeled as spatial Gaussian processes. Applications of the models are presented including prediction of the daily temperature series or any quantile of interest at unobserved sites. Inference to investigate climate change comparison is presented including predictive spatial probability surfaces and spatial extents for an event, including extreme heat events.

In a different direction, acknowledging the lack of exploratory or inferential tools for the analysis of record-breaking events, some of the proposed methods consist of graphical tools and hypothesis testing within the classical framework. The underlying idea of all these non-parametric tools is to use the distribution of the occurrence of records under series of independent and identically distributed continuous random variables to analyze if the observed records are compatible with that behavior. After proposing a general framework for the exploration of the occurrence of records, a first attempt is made in the modeling of such events. A Bayesian hierarchical spatial logistic

regression model is proposed for the record indicators, which incorporate suitable fixed effects and strong daily spatial random effects. Applications of the model are presented including predictions of record indicator series at unobserved sites, the spatial extent of record surface, and comparison with stationarity to assess warming rates.

Remarkable evidences of climate change found in this project include increasing global temperature trends in the summer season for Aragón, both in the mean and high quantiles, but weaker in the latter. The mean exhibits a homogeneous global warming behavior across space, while high quantiles reveal different types of trends—generally more pronounced than for the mean—across space. The occurrence of record-breaking events displays distinct characteristics in temperature's distribution, along with a significant increase in the number of records in Spain with respect to the stationary case. This increase is neither homogeneous across years—with the past two decades being the most affected—, across seasons—with summer being the most affected—, nor across space—with the Cantabrian coast being the least affected. Another notable finding is the strong effect of persistence—previous day's temperature dependence—being more pronounced in the central part of the distribution than in high quantiles, and being also very strong in the occurrence of records for consecutive days.

These findings highlight, on one hand, the importance of employing specialized statistical models that focus on characterizing the specific features of interest in the dataset, as well as models that adequately capture the spatiotemporal dependence to borrow strength across sites and provide a proper assessment of uncertainty. On the other hand, the proposed methods allow analyzing the magnitude of the potential effects of climate change along with a measure of uncertainty, and indicate the need to inform management strategies to mitigate the impact of extreme temperature events on human health, agriculture, and the economy.

Resumen

La evidencia del cambio climático inducido por el aumento de la concentración de gases de efecto invernadero en la atmósfera sugiere un incremento no solo en la temperatura media, sino también en días cálidos más frecuentes, olas de calor, y temperaturas extremas y récord. Esta Tesis tiene como objetivo desarrollar modelos estocásticos novedosos para el análisis espaciotemporal de tales eventos de temperatura extrema. Además, el propósito de este proyecto es brindar información sobre una región geográfica específica, alrededor de la Comunidad Autónoma de Aragón en España y la España peninsular, acerca de su estado del clima en cuanto a temperaturas.

Los conjuntos de datos de este proyecto contienen series de más de 60 años de observaciones diarias de temperatura máxima en 18 y 40 estaciones meteorológicas en torno a Aragón y España, respectivamente. La mayoría de los métodos propuestos para modelizar tales datos caen dentro del marco jerárquico bayesiano para datos espaciotemporales punto referenciados. El ajuste de estos modelos se realiza a través de algoritmos Monte Carlo basados en cadenas de Markov diseñados específicamente, y una perspectiva basada en modelos ofrece inferencia completa y exacta con una evaluación adecuada de la incertidumbre. Como alternativa a la teoría asintótica de valores extremos, los modelos espaciotemporales propuestos incluyen especificaciones autorregresivas en el tiempo de medias y cuantiles—múltiples y conjuntos—con efectos fijos y aleatorios modelados como procesos gaussianos espaciales. Se presentan las aplicaciones de los modelos, incluida la predicción de la serie diaria de temperaturas o cualquier cuantil de interés en sitios no observados. Se presenta la inferencia para investigar la comparación del cambio climático, incluidas superficies de probabilidad espacial predictivas y extensiones espaciales de un evento, incluidos los eventos de calor extremo.

En una dirección diferente, reconociendo la falta de herramientas exploratorias o inferenciales para el análisis de eventos récord, algunos de los métodos propuestos consisten en herramientas gráficas y contrastes de hipótesis dentro del marco clásico. La idea subyacente de todas estas herramientas no paramétricas es utilizar la distribución

de la ocurrencia de récords bajo series de variables aleatorias continuas independientes e idénticamente distribuidas para analizar si los récords observados son compatibles con ese comportamiento. Después de proponer un marco general para la exploración de la ocurrencia de récords, se hace un primer intento en la modelización de tales eventos. Se propone un modelo espacial jerárquico bayesiano de regresión logística para los indicadores de récord, que incorpora efectos fijos adecuados y fuertes efectos aleatorios espaciales diarios. Se presentan aplicaciones del modelo, incluidas predicciones de series de indicadores de récord en sitios no observados, la extensión espacial de la superficie récord, y comparación con estacionariedad para evaluar las tasas de calentamiento.

Evidencias notables de cambio climático encontradas en este proyecto incluyen tendencias globales crecientes en las temperaturas en la estación de verano para Aragón, tanto en la media como en los cuantiles altos, pero más débiles en estos últimos. La media muestra un comportamiento homogéneo de calentamiento global en todo el espacio, mientras que los cuantiles altos revelan diferentes tipos de tendencias—generalmente más pronunciadas que para la media—a lo largo del espacio. La ocurrencia de récords muestra características distintas en la distribución de temperaturas, junto con un aumento significativo en el número de récords en España con respecto al caso estacionario. Este aumento no es homogéneo ni a lo largo de los años—siendo las dos décadas pasadas las más afectadas—, ni a lo largo de las estaciones—siendo el verano el más afectado—, ni a lo largo del espacio—siendo la costa Cantábrica la menos afectada. Otro hallazgo notable es el fuerte efecto de la persistencia—la dependencia con respecto a la temperatura del día anterior—, que es más pronunciada en la parte central de la distribución que en los cuantiles altos, y siendo también muy fuerte en la ocurrencia de récords en días consecutivos.

Estos hallazgos resaltan, por una parte, la importancia de emplear modelos estadísticos especializados que se enfoquen en caracterizar las particularidades de interés del conjunto de datos, así como modelos que capturen adecuadamente la dependencia espaciotemporal para compartir evidencia entre sitios y proporcionar una estimación adecuada de la incertidumbre. Por otro lado, los métodos propuestos permiten analizar junto a una medida de incertidumbre la magnitud de los posibles efectos del cambio climático e indican la necesidad de informar estrategias de gestión para mitigar el impacto de eventos de temperatura extrema en la salud humana, la agricultura y la economía.

Contents

Acronyms

AD absolute deviation

AL asymmetric Laplace

AUC area under the receiver operating characteristic curve

BS Brier score

c.i.i.d. continuous independent and identically distributed

cdf cumulative distribution function

CI credible interval

CRAN Comprehensive R Archive Network

CRM classical record model

CRPS continuous ranked probability score

CVG coverage

EDA exploratory data analysis

EHE extreme heat event

EVT extreme value theory

GEV generalized extreme value

GP Gaussian process

i.i.d. independent and identically distributed

LDM linear drift model

LOOCV leave-one-(site)-out cross-validation

MAE mean absolute error

MCMC Markov chain Monte Carlo

OR odds ratio

pdf probability density function

QAR quantile autoregression

QR quantile regression

RI reference interval

RMSE root-mean-square error

WMAE weighted mean absolute error

Chapter 1

Introduction

The classical modeling of spatial extremes relies on asymptotic models for block maxima or peaks over high thresholds. Usual criticisms of these models include that their dependence structure is too rigidly constrained and the prohibitive computational cost to fit in high dimensions. Asymptotic arguments and parametric assumptions are needed to estimate very extreme quantities, however, many important applications where the focus is less extreme could benefit from the less restrictive and more robust quantile regression approach. For example, obtaining 1,000-year return levels required by many flood risk management commissions versus studying the effect of climate change on the 20-year return level of daily temperature. In addition, the continuously increasing interest on record-breaking events in atmospheric sciences, but an overall lack of research regarding how to model the occurrence of such events provides an attractive opportunity for research. This Thesis aims to develop novel stochastic models for the spatiotemporal analysis of extreme temperatures such as extreme heat, high quantile, and record-breaking events. This chapter will provide an introduction to the project by first discussing the background and the research problems, followed by the research objectives and their thematic unit, and finally the outline of the Thesis.

1.1 Background and methodological problems

Natural hazards in the context of temperature such us heat waves, or extreme and record-breaking temperatures arise due to physical processes that are spatial in extent and temporal in duration. In atmospheric science, spatiotemporal extreme data is a subject of great interest because extreme weather events usually yield the largest human, agricultural, and economic impacts (Coumou & Rahmstorf, 2012). Understanding the possible space-time alterations in phenomena characteristics like temperature due to climate change is vital for assessing the risk of future extreme events. The following sections will introduce key ideas about spatial statistics,

extreme value theory, quantile regression, and theory of records to provide background knowledge for this Thesis. This includes the meaning of these concepts, main pieces of theory, the current context, and gaps in the literature to be filled with this project.

1.1.1   Bayesian geostatistics

Arguably, the most common scenario in spatial data involves what is known as *geostatistics*—data modeled as values from a spatial process defined on the continuum but observed at fixed spatial locations. This section provides a brief overview of a fully model-based perspective for such data, the approach of hierarchical modeling fitted within a Bayesian framework. Standard texts from spatial, spatiotemporal, and Bayesian hierarchical perspectives are Cressie (1993), Cressie and Wikle (2011), and Banerjee et al. (2014), respectively.

The most basic geostatistical hierarchical model considers a response $Y(\mathbf{s})$ at location $\mathbf{s}$ over a spatial domain D specified as

$$Y(\mathbf{s}) = \mathbf{x}^\top(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1.1}$$

where the residual is partitioned into a spatial piece $w(\mathbf{s})$ and a non-spatial piece $\epsilon(\mathbf{s})$. The statistical problem is to make inference for the response elsewhere in D. The geostatistical story is explained by $w(\mathbf{s})$, usually modeled as a Gaussian process (GP) (Gelman et al., 2013, Chapter 21) with an isotropic stationary covariance function, $\mathrm{cov}(w(\mathbf{s}), w(\mathbf{s}')) = C(||\mathbf{s} - \mathbf{s}'||; \sigma_w^2, \phi_w)$, with *partial sill* parameter σ_w^2 and *decay* parameter ϕ_w. The pure error $\epsilon(\mathbf{s})$ is usually modeled as independent and identically distributed (i.i.d.) normal with *nugget* parameter σ_ϵ^2. The *sill* is the sum of the variance components $\sigma_w^2 + \sigma_\epsilon^2$.

Given a dataset $Y(\mathbf{s}_i)$ for $i = 1, \ldots, n$, let $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^\top$ and $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^\top$, and denote by $R(\phi_w)$ the correlation matrix arising for $\mathbf{w}$. There are two inferentially equivalent but alternative representations of the model, a marginal in $\mathbf{w}$ representation where the likelihood is

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_w^2 R(\phi_w) + \sigma_\epsilon^2 I),$$

and a hierarchical conditional in $\mathbf{w}$ representation

$$\begin{aligned}
\text{First stage} \quad & \mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \sigma_\epsilon^2 I), \\
\text{Second stage} \quad & \mathbf{w} \mid \sigma_w^2, \phi_w \sim N(\mathbf{0}, \sigma_w^2 R(\phi_w)), \\
\text{Third stage} \quad & \text{Priors on } [\boldsymbol{\theta}],
\end{aligned}$$

that emerges as a special case of the general multi-level process model

$$[data \mid process, parameters][process \mid parameters][parameters].$$

In the Bayesian framework, independent priors are usually chosen for the model parameters, yielding $[\boldsymbol{\theta}] = [\boldsymbol{\beta}][\sigma_w^2][\sigma_\epsilon^2][\phi_w]$. Customary weakly informative priors are multivariate normal for $[\boldsymbol{\beta}]$, inverse gamma for $[\sigma_w^2]$ and $[\sigma_\epsilon^2]$, and usually uniform or gamma for $[\phi_w]$ although depends upon the choice of C. Care must be taken with regard to σ_w^2, σ_ϵ^2, and ϕ_w, since an improper prior for all of them leads to an improper posterior.

Markov chain Monte Carlo. Model fitting can then proceed using Markov chain Monte Carlo (MCMC) (Gelman et al., 2013, Part III) algorithms. MCMC offers a powerful tool to approximate complex posterior distributions that arise from hierarchical models. These algorithms allow to explore the parameter space of the model, obtaining samples that provide insights into the uncertainty associated with the model parameters. Bayesian kriging obtains posterior predictive samples from the response at new locations using these posterior samples from the model parameters by means of composition sampling (Banerjee et al., 2014, Chapter 6). Full Bayesian model-based inference is straightforward using these posterior predictive samples to provide a posterior distribution for any quantity of interest. Here, *full* emphasizes obtaining the entire posterior distribution, while *model-based* underscores the use of probabilistic models that describe the data generating process.

For example, frequently, for a function g, of great interest is the block average,

$$Y(D) = \frac{1}{|D|} \int_D g(Y(\mathbf{s})) \, d\mathbf{s},$$

where $|D|$ is the area of D. Also of interest is the particular case $g(y) = \mathbf{1}(y \leq c)$ that leads to the definition of the spatial cumulative distribution function (cdf) (Banerjee et al., 2014, Chapter 15). The spatial cdf behaves like a cdf in the sense that it is non-decreasing and goes to 0 as $c \to -\infty$ and 1 as $c \to \infty$, however, since it is a function of the process realization, it is a random variable and arises as a stochastic integral.

A spatiotemporal geostatistical model. In recent years, there has been a dramatic surge in interest towards spatiotemporal modeling attributed to the rapid growth of datasets with both spatial and temporal indices. Rather than revising a huge list of works in this topic, the Schliep et al. (2021) model used to illustrate novel model-based tools in the publication of Section 3.2 will be revisited. Their spatiotemporal model is specifically developed to predict the occurrence and incidence of extreme heat events (EHEs) in space. Extreme heat is defined in terms of temperatures that are much hotter than average for a particular time and location,

and an EHE is defined as a period of persistent extreme heat. Specifically, EHEs are defined locally and are based on exceedance of a suitable local threshold. To address this, they introduce thresholding, i.e., a model which switches between two observed states, one that defines extreme heat days (those above the temperature threshold) and the other that defines non-extreme heat days (those below the temperature threshold).

Let $Y_t(\mathbf{s})$ denote the daily maximum temperature at day t and location $\mathbf{s}$. Schliep et al. (2021) proposed a two-state model where the state for a given day defines whether or not the location is experiencing an EHE. Considering known a threshold (quantile), $q(\mathbf{s})$ at location $\mathbf{s}$, the threshold state indicator $U_t(\mathbf{s})$ takes the value 1 if $Y_t(\mathbf{s}) > q(\mathbf{s})$ and 0 otherwise. The joint distribution for temperature and state is specified in a first order Markov fashion explicitly as follows. Given $Y_{t-2}(\mathbf{s})$, the joint distribution $[U_{t-1}(\mathbf{s}), Y_{t-1}(\mathbf{s}), U_t(\mathbf{s}), Y_t(\mathbf{s})]$ is written as

$$[Y_t(\mathbf{s}) \mid U_t(\mathbf{s}), Y_{t-1}(\mathbf{s})][U_t(\mathbf{s}) \mid Y_{t-1}(\mathbf{s})][Y_{t-1}(\mathbf{s}) \mid U_{t-1}(\mathbf{s}), Y_{t-2}(\mathbf{s})][U_{t-1}(\mathbf{s}) \mid Y_{t-2}(\mathbf{s})].$$

This formulation requires three model specifications:

(i) $[Y_t(\mathbf{s}) \mid U_t(\mathbf{s}) = 0, Y_{t-1}(\mathbf{s})]$,

(ii) $[Y_t(\mathbf{s}) \mid U_t(\mathbf{s}) = 1, Y_{t-1}(\mathbf{s})]$,

(iii) $[U_t(\mathbf{s}) \mid Y_{t-1}(\mathbf{s})]$.

Truncated distributions are needed for (i) and (ii), i.e., $[Y_t(\mathbf{s}) = y]\mathbf{1}(y \leq q(\mathbf{s}))$ and $[Y_t(\mathbf{s}) = y]\mathbf{1}(y > q(\mathbf{s}))$, respectively. For (i), a truncated normal distributions with autoregressive centering is adopted,

$$TN\left(\mu_t^0(\mathbf{s}) - \rho^0(Y_{t-1}(\mathbf{s}) - \mu_{t-1}^0(\mathbf{s})), \sigma^{2,0}(\mathbf{s})\right) \mathrm{I}(-\infty, q(\mathbf{s})).$$

For (ii), a truncated t distribution with autoregressive centering is adopted,

$$Tt\left(\mu_t^1(\mathbf{s}) - \rho^1(Y_{t-1}(\mathbf{s}) - \mu_{t-1}^1(\mathbf{s})), \sigma^{2,1}(\mathbf{s})\right) \mathrm{I}(q(\mathbf{s}), \infty).$$

For (iii), a probit link is employed to define

$$P(U_t(\mathbf{s}) = 1 \mid Y_{t-1}(\mathbf{s})) = \Phi(\eta_t(\mathbf{s})).$$

These $\mu_t^0(\mathbf{s})$, $\mu_t^1(\mathbf{s})$, and $\eta_t(\mathbf{s})$ are linear expressions of regressors and a spatially varying intercept modeled as a GP. Further, spatially varying variances modeled as the logarithm of a GP are introduced for the truncated distributions, and $\eta_t(\mathbf{s})$ includes expressions of $Y_{t-1}(\mathbf{s})$ centered by the threshold. Putting (i), (ii), and (iii) together, a mixture distribution for $Y_t(\mathbf{s})$ results: (i) a truncated normal distribution for the bulk

of the distribution, (ii) a truncated t distribution for the upper tail of the distribution, and (iii) mixture weights according to $P(U_t(\mathbf{s}) = 0)$ or $P(U_t(\mathbf{s}) = 1)$, respectively.

Model inference is obtained in a Bayesian framework, the model is completely specified giving customary weakly informative prior distributions for each of the model parameters, and model fitting proceeds through MCMC to obtain samples from the joint posterior distribution. These samples allow full inference based on model-based geostatistics.

Bayesian models can be either general, designed for versatility across various problems, or application-specific, tailored to address unique challenges within a specific context. The generic model in (1.1) offers flexibility but might lack precision for assessing EHEs, while the application-specific Schliep et al. (2021) model provides precision but might be limited to the analysis of threshold state. For example, the Schliep et al. (2021) model can be useful to learn about notions of the spatial extent of heat waves and EHEs, which have previously been considered informally and descriptively in the climate community. As a result, the existing research might be useful to offer context but is sometimes inadequate for a particular objective and dataset, as every problem should be data-driven to capture well the specific characteristics of interest.

1.1.2 Extreme value theory

Extreme value theory (EVT) concerns inference for rare events in data. Often very extreme values have never yet been observed, and their probabilities must therefore be estimated by extrapolation of asymptotic tail models fitted to available data. This section is included here to give a detailed background for most approaches available for the analysis of extremes, but it will not be used in the methodology proposed in this Thesis. Following the reviews by Davison et al. (2012) and Huser and Wadsworth (2022), this section gives a very brief outline of the classical theory of extremes for maxima and threshold exceedances of stationary series, and the main approaches for the analysis of spatial and spatiotemporal extremes. Standard texts from mathematical and statistical points of view are de Haan and Ferreira (2006) and Coles (2001), respectively.

Statistical modeling of extremes is usually based on limiting families of distributions for maxima that satisfy the property of max-stability. If a non-degenerate limiting distribution for the scaled maximum of a series of continuous independent and identically distributed (c.i.i.d.) variables exists, then it must be max-stable. The only non-degenerate distribution with that property is the generalized extreme value (GEV) distribution, which includes the Weibull, the Gumbel, and the Fréchet distributions.

A typical goal in applications is the estimation of a high τ-quantile of the distribution of the maximum. If the available observations are annual maxima and $\tau = 1 - 1/T$, then the τ-quantile is called the T-year return level, interpreted as the level exceeded once on average every T years. Under mild conditions on the dependence structure of stationary time series, the GEV also emerges as the only possible non-degenerate limiting distribution for scaled maxima of blocks of observations.

The GEV may also be regarded as giving an approximation for the upper tail of the distribution of an individual variable. In this case, the only limiting distribution of the excess of the individual variable conditioned to exceed a limiting high threshold is the generalized Pareto distribution, which is commonly used for modeling exceedances over high thresholds. The standard approach to such modeling uses that the times of exceedances over the high threshold are the realization of a stationary Poisson process.

Multivariate approaches to extreme values require additional max-stability properties for the joint distribution. Unlike for univariate extremes, there is no simple parametric form for the multivariate limiting distribution. While various parametric forms for the joint distribution have been proposed, those commonly employed now often lack flexibility, and non-parametric estimation has mainly been limited to the bivariate scenario due to the curse of dimensionality.

Numerous statistical techniques, such as Bayesian hierarchical models, copulas, max-stable or r-Pareto processes, subasymptotic models, and conditional spatial extremes, have been employed very recently for spatial extreme modeling (see Huser & Wadsworth, 2022, for a remarkably rich review on spatial extremes). Bayesian hierarchical models consider GEV or generalized Pareto marginals with distribution parameters modeled in terms of regressors and a GP to obtain spatially varying return levels. The copula models extend the hierarchical ones by modeling spatial dependence between observations. While hierarchical models model the response surface as everywhere discontinuous, copula models model it as continuous. The Gaussian or t copulas are simple approaches, while extremal copulas are more difficult to work with but satisfy the max-stable conditions. Max-stable and r-Pareto processes are theoretically justified, but computation in most cases is difficult and requires composite likelihood methods. They are asymptotically dependent processes. In practice, however, environmental data often tend to exhibit weakening dependence for increasing quantile levels and to support asymptotic independence. By contrast with the asymptotic max-stable and r-Pareto processes, subasymptotic models are more flexible hybrid versions that can bridge the two asymptotic dependence regimes. The last approach, conditional spatial extremes, aims at describing the spatial behavior of a random process conditional on single points being large. It allows for very flexible

forms of extremal dependence and can naturally capture both asymptotic dependence and independence.

1.1.3 Quantile regression

Classical quantile regression (QR) models the relationship between a set of regressors and a specific quantile (equivalent to a return level) of the response variable, and it offers a less restrictive and more robust approach. Simplifying interpretation, the focus here is on *linear* QR, while *nonlinear* methods are beyond the scope of this project. This section reviews the two main approaches for QR—multiple and joint—, and a quantile autoregression (QAR) model for time series. A standard text for QR is Koenker (2005), but as far as I know there is not systematic texts nor reviews that include spatial or spatiotemporal QR, so the existing approaches are summarized at the end of this section.

Multiple quantile regression. Initiated by Koenker and Bassett (1978), the scope of QR methods has broadened considerably in recent years. The so-called *multiple* QR offers a separate regression model for each of the τ-quantiles of interest, and inference proceeds by minimizing a piecewise linear check loss function,

$$\min_{\boldsymbol{\beta}} \sum \delta_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $\delta_\tau(u) = u[\tau - \mathbf{1}(u < 0)]$ and $\tau \in (0, 1)$. The simplest case employs $\tau = 0.5$ and yields the sum of absolute deviations, which is minimized by the median.

In a manner similar to how the check loss corresponds to quantiles just as the sum of squares corresponds to the mean, the normal distribution also has an equivalent counterpart for quantiles. First introduced by K. Yu and Moyeed (2001), the Bayesian approach assumes

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) + \epsilon_i(\tau)$$

where $\epsilon_i(\tau)$ are i.i.d. asymmetric Laplace (AL) (Kotz et al., 2001) errors. The AL distribution is characterized by location, scale, and asymmetry parameters; i.e., μ, $\sigma > 0$, and $\tau \in (0, 1)$, respectively. Fixing $\mu = 0$ to ensure $P(\epsilon \leq 0) = \tau$, the density of $\epsilon \sim AL(0, \sigma, \tau)$ is written as

$$f(\epsilon) = \tau(1 - \tau)\sigma \exp\{-\sigma\delta_\tau(\epsilon)\}.$$

The τ-quantile of Y_i given $\mathbf{x}_i$ is then simply $\mathbf{x}_i^\top \boldsymbol{\beta}(\tau)$. A convenient representation of ϵ in terms of standard normal and exponential variables is

$$\epsilon = \sqrt{\frac{2U}{\sigma^2\tau(1-\tau)}}Z + \frac{1 - 2\tau}{\sigma\tau(1-\tau)}U, \text{ with } Z \sim N(0, 1) \text{ and } U \sim Exp(1). \quad (1.2)$$

A primary issue with multiple QR is the so-called quantile crossing problem—increasing quantiles in τ are not guaranteed. However, multiple QR imposes *soft* (stochastic) order on the quantiles, and rich modeling for a given quantile is relatively straightforward; which motivates its implementation when interest is focused on one or a few separated quantiles.

Joint quantile regression. Recently, advances have been made in the so-called *joint* (or simultaneous) QR, supplying an appropriate joint model for all quantiles which avoids crossing of the regression across quantiles. Bondell et al. (2010) suggested to alleviate quantile crossing by solving the classical optimization problem adding monotonicity constraints on a grid of quantiles, but estimates based on this method can be sensitive to the number and location of the chosen grid.

Foundational work appears in Tokdar and Kadane (2012) presenting a simple characterization of the joint QR with one regressor through an interpolation of two monotone curves modeled via logistic transformations of a smooth GP. Their approach, embedded within a Bayesian framework, leads to a likelihood function that can be computed through a one-dimensional root-finder. The same inferential framework is used by Yang and Tokdar (2017) proposing a novel parameterization that characterizes any collection of non-crossing quantile planes over arbitrarily shaped convex regressor domains. Their parameterization uses unconstrained scalar, vector and function valued parameters with GP prior distributions. Other joint approaches were implemented directly in the spatial framework and will be summarized below.

These joint methods offer coherence from a generative model perspective, possibilities for joint inference across quantile levels, and strength borrowing across proximate quantile levels; at the cost of a perhaps more limited modeling, a bounded support on regressors, and approximate and very demanding computation. Choosing between multiple or joint methods should depend on the research problem and researchers expertise.

Quantile autoregression. Time series data has traditionally been based on Gaussian models that exclusively use first- and second-moment information. However, asymmetries and heterogeneous patterns of dependence across different levels of the distribution—invisible features for Gaussian models—, can be revealed with the aid of QR methods. Koenker and Xiao (2006) offered theoretical results for QAR(p) models of the form

$$Q_{Y_t}(\tau \mid y_{t-1}, \ldots, y_{t-p}) = \theta_0(\tau) + \sum_{j=1}^{p} \theta_j(\tau) y_{t-j}.$$

The usual i.i.d. error autoregression with errors having quantile function $\theta_0(\tau)$ is a particular case when the $\theta_j(\tau)$ do not depend upon τ for $j = 1, \ldots, p$. Provided non-crossing quantile functions, they linked them to the generative model,

$$Y_t = \theta_0(U_t) + \sum_{j=1}^{p} \theta_j(U_t) Y_{t-j}$$

with $U_t \sim U(0,1)$. For the monotonicity constraint, they required all $\theta_j(\tau)$ for $j = 0, \ldots, p$ to be strictly increasing functions (referred to as comonotonic coefficients). Comonotonic coefficients ensure non-crossing of the resulting quantiles, but under the restrictive assumption that the autoregression coefficients strictly increase in τ. They illustrate for $p = 1$ with $\theta_0(\tau) = \sigma\Phi^{-1}(\tau)$ and $\theta_1(\tau) = \min\{\gamma_0 + \gamma_1\tau, 1\}$ for $\gamma_0 \in (0,1)$ and $\gamma_1 > 0$.

A somewhat restrictive limitation of any linear specification is that there must be a bounded support for Y_t beyond which the ordering of quantiles is reversed, so linear joint QAR models must be regarded as useful local approximations over a prespecified bounded support.

Spatial quantile regression. A few Bayesian QR models have been proposed for the analysis of point-referenced spatial and spatiotemporal data. Lum and Gelfand (2012) extended the usual i.i.d. AL error to a spatial process that incorporates spatial dependence; the latent normal and exponential variables from the mixture representation are modeled as a GP and a copula transformation.

Reich et al. (2011) developed a spatial joint QR model with spatially varying regression coefficients to account for spatial dependence; the coefficients are expressed as a weighted sum of Bernstein basis polynomials where the weights are constrained spatial GPs. A somewhat similar approach was followed by Reich (2012) in a spatiotemporal context using only one regressor; he considered piecewise Gaussian basis functions rather than Bernstein polynomials, and allowed for spatiotemporal residual correlation via a dynamic spatial Gaussian copula process over the quantile levels of the observation units. The so-called constraint-free parameterization by Yang and Tokdar (2017) was extended to spatial data by Chen and Tokdar (2021); they characterized again spatial dependence via a Gaussian or t copula process on the underlying quantile levels, but obtained quantiles that do not vary spatially.

Spatiotemporal QR development is important for the analysis of data with spatial and temporal dependence. A few studies have proposed methods to manage spatial or temporal data with the QR approach. However, these studies have rarely focused on both of them simultaneously and never from a spatial QAR perspective. Expectedly,

explicit assessment of persistence—previous times' data dependence—within a proper implementation of spatial dependence, or assessment of changes in the distribution in space and time, are difficult.

1.1.4 Record-breaking events

In recent times, records have gained special importance for climate researchers in the context of global warming. The question of how a changing climate impacts the number of temperature records that are observed has intrigued both the general public and scientists. It is not surprising that the breaking of temperature records is now a hot topic, more and more frequent all over the world (Plumer & Shao, 2023). Analysis of such events is an area that has received attention primarily within the probability community, so only a few recent studies have focused on applied record-breaking events. This section introduces the main statistics of records and reviews their theory for the analysis of climate data, dominated by results for records in stationary series with some additional work addressing trends. Standard texts for the probabilistic theory of records are Arnold et al. (1998) and Nevzorov (2001).

Probabilistic properties of records. Initiated by Chandler (1952), probabilistic properties of record events have been pursued quite extensively. An observation Y_t in a time series is called a *record* if its value exceeds that of all previous observations, i.e., if $Y_t > \max\{Y_1, \ldots, Y_{t-1}\}$. By definition, Y_1 is always considered a *trivial* record. The occurrence of records is characterized by the series of record indicators defined as I_t taking the value 1 if Y_t is a record and 0 otherwise. The number of records up to time t is given by $N_t = \sum_{j=1}^{t} I_j$.

The classical record model (CRM) characterizes records arising in a series of c.i.i.d. random variables and provides the expected behavior in stationary climatic series. The main property of the variables associated with the occurrence of records under the CRM is that they do not depend on the underlying distribution of the c.i.i.d. variables. Foster and Stuart (1954) demonstrated that the record indicators are mutually independent and follow a Bernoulli distribution with probability $p_t = P(I_t = 1) = 1/t$. Then, the number of records up to time t is asymptotically normal and its expected value grows as the logarithm of the number of variables, $E(N_t) = \sum_{j=1}^{t} 1/j = \log(t) + \gamma + O(1/t)$, where γ is the Euler-Mascheroni constant. The distribution-free properties under the CRM have been used to develop statistical hypothesis tests to detect non-stationary behavior in the occurrence of records, first from a probabilistic point of view by Foster and Stuart (1954) and Diersen and Trenkler (1996), and then from a climatic applied perspective initiated by Benestad (2003, 2004).

In climatic data, stationary evolution over time is an unrealistic assumption and an alternative specification for the temperature series is the linear drift model (LDM), $Y_t = ct + \epsilon_t$, where $c > 0$ is a constant and ϵ_t are c.i.i.d. errors. In this simple scenario, the probability density function (pdf) $f_t(y)$ of Y_t is of the form $f_t(y) = f(ct + x)$ with a fixed pdf $f(y)$ and the corresponding cdf $F(y)$, which is the distribution of the errors. The probability of record at time t in the LDM depends on the distribution of the error term and is given by

$$p_t(c) = \int_{-\infty}^{\infty} f(x) \prod_{j=1}^{t-1} F(cj + x) \, dx.$$

It was shown by Ballerini and Resnick (1985, 1987) that $p_t(c)$ exists and becomes constant in the asymptotic limit for $t \to \infty$ provided the distribution of ϵ_t has a finite first moment. In other words, in the LDM the probabilities of record do not tend to zero as they do in the CRM. Gouet et al. (2020) generalized some of the LDM results for δ-records (observations higher than the previous record plus a constant δ).

Franke et al. (2010) investigated the asymptotic behavior of the probability $p_t(c)$ for a few families of distributions. Their approximation for the special case of the normal distribution has been used quite extensively to analyze records in temperature data (see, e.g., Wergen & Krug, 2010; McBride et al., 2022). Other authors such as Rahmstorf and Coumou (2011) or Fischer et al. (2021) also consider the normal distribution but with more complex trends and derive the probabilities by numerical integration.

Record-breaking investigation is vital, as records yield particularly large impacts because society often tends to react and adjust based on the most extreme event it has encountered within a lifetime, rather than considering events beyond that scope. When events of comparable magnitude recur after several years, the resulting impacts might be notably reduced, as society has had the opportunity to adapt, at least to some extent. Numerous studies have investigated probabilistic approaches to describe the occurrence of temperature records. However, these studies have been focused on the CRM, the parametric LDM, and simple extensions with more elaborate trends. Probabilistic record models are useful exploratory tools that should be readily available, but effective analysis for adaptation decisions would benefit from richer modeling strategies.

1.2 Motivation, objectives and thematic unit

Given the lack of research within the space-time domain regarding extreme heat, high quantile, and record-breaking methodology, this project will aim to develop model-based tools, spatiotemporal QAR models, and record-breaking methodology

to identify and evaluate the effects of climate change in different characteristics of daily temperature extremes around the Comunidad Autónoma de Aragón in Spain and peninsular Spain—Aragón and Spain from now on.

Study region and meteorological stations. The data that motivates most of the methods proposed in this project are the extended summer season for Aragón; while the data motivating the analysis of records are all year round in peninsular Spain. Peninsular Spain has an area of 492,175 km^2, while Aragón represents almost a 10% of it, 47,719 km^2. Spain is a diverse geographic region with several mountain ranges and a long coastline. Aragón, an inland region located in the northeast of Spain, is not an exception; with the Pyrenees to the north rising above 3,000 meters, the Ebro Valley as a central axis below 300 meters, and the Iberian system to the south exceeding 2,000 meters. Outside Aragón, Spain has the Central Plateau in the center, the Sierra Nevada near the Mediterranean Sea to the south and east, and the Atlantic Ocean to the north and west completes its long coastline. Both of them represent regions of exceptional climatic interest, not only because of their climatic variability, but also because Spain has experienced some of the most significant temperature increments in the world over the past few decades (Lionello & Scarascia, 2018).

The datasets for this project contain series of over 60 years (1956–2015 and 1960–2021) of daily maximum temperature observations at 18 and 40 meteorological stations throughout Aragón and Spain, respectively. The stations are irregularly distributed across the regions representing their diverse climatic zones. The Aragón dataset do not have any missing data, while the Spanish dataset contains a very small amount of missing data. Figure 1.1 shows an elevation map with the region of Aragón and the stations in the Iberian Peninsula. While Aragón is given by an irregular polygon, the actual study region is a rectangle similar in dimension and location that includes a few stations outside the Comunidad itself.

For both regions, one can anticipate strong persistence between two consecutive daily temperatures. The study of this type of temporal dependence in space through explicit autoregressive models will be of key interest in the methodology. Regressors could also bear important information in describing the daily maximum temperatures. Spatially varying time trends must be included in the models to assess different trends of climate change in space. To interpolate over the study regions and produce maps, the spatial regressor information must be available for the entire region and not just at the meteorological stations. Two readily available spatial covariates are elevation and distance to coast. For an area with both mountains and valleys, it is highly likely that elevation will have a crucial role on the behavior of temperatures. It is not so clear
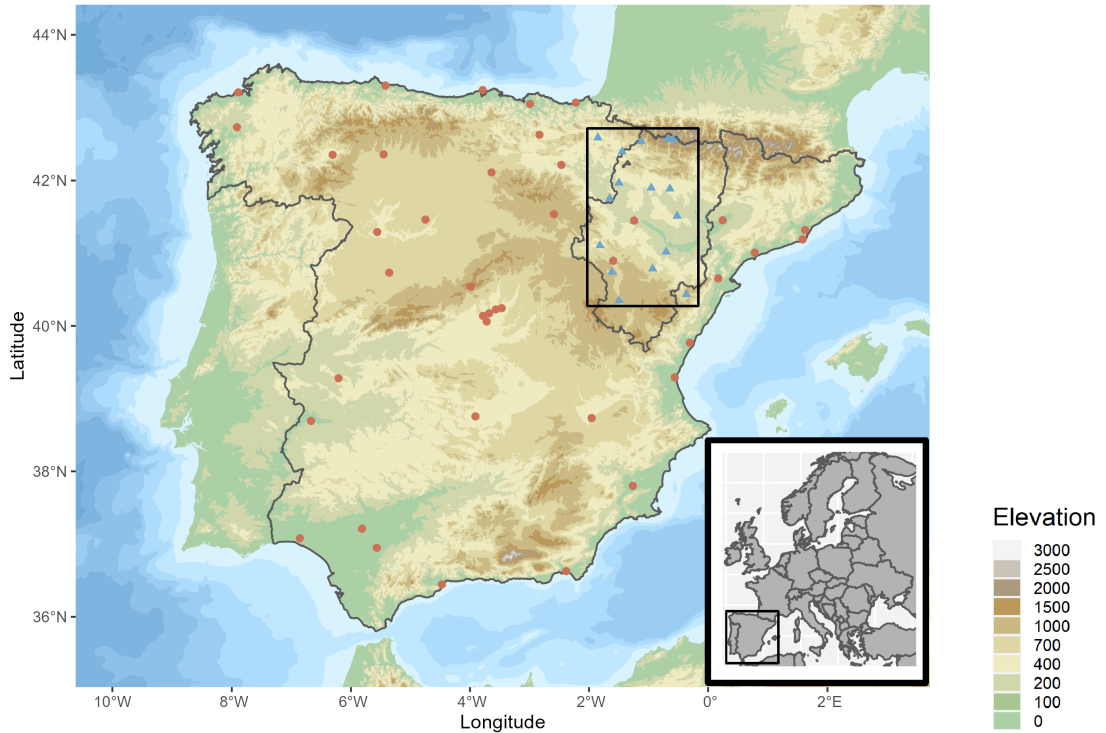
Figure 1.1: Map of the 18 Aragón stations (blue triangles) and the 40 Spanish stations (red dots) in southwestern Europe. The big gray irregular polygon is peninsular Spain, the small gray irregular polygon is the Comunidad Autónoma de Aragón, and the black rectangle is the actual study region around Aragón.

whether these regressors have a relationship with climate change.

Objectives and thematic unit. The overall aim is divided into three specific objectives, which will be developed in the subsequent sections in Chapter 3 coinciding with the publications that compose this Thesis. These objectives are: (i) to develop a spatiotemporal geostatistical mean model and propose simple model-based tools to promote such methodology within the climatic community, and to extend those tools for the analysis of EHEs; (ii) to develop novel spatiotemporal multiple and joint QAR models; and (iii) to offer an accessible framework for the analysis of records, and to develop a modeling framework for the occurrence of record-breaking events. All these objectives are addressed from an applied methodological perspective, and mainly, but not exclusively, within the Bayesian hierarchical model-based paradigm. Although the models are driven by the Aragón or Spanish datasets, all methods can be easily adapted for application in other ecological or environmental contexts. The role that each publication has had in the fulfillment of these objectives will be specified in Chapter 2. There are, however, two unpublished works submitted for publication that address the parts of model-based tools and record-breaking modeling, which could not

be included in the compendium as they have not yet been accepted for publication at the time of depositing this Thesis. Chapter 2 will also provide a comprehensive summary of these two works. Below I attempt to give greater clarity to the thematic unit of the project.

Spatiotemporal models for extremes entail two challenges. First, a spatiotemporal approach needs to incorporate dependence in both, time and space. For instance, it entails including trend effects, seasonal component, serial correlation, spatial dependence, other regressors, and random effects, among others. Second, modeling extremes calls for specific techniques tailored to their analysis, which in climate applications should be spatiotemporal in nature.

The first objective is to develop more standard spatiotemporal models for the mean of the response distribution. The flexibility of these spatiotemporal models with complex dependence structures represented through fixed and random effects is enhanced within the Bayesian hierarchical framework and the associated fitting methods like MCMC algorithms. These models are particularly valuable in climate applications as they enable the quantification of trends alongside their seasonal and spatial variability. Specifically, the model-based approach allows for the generation of practical derivatives such as probability surfaces and extents, including those of EHEs.

The second objective is to develop more general spatiotemporal models for the entire response distribution, particularly for extremes, based on quantile modeling. Insights gained from modeling means can be extended to a multiple QR model, facilitating the comparison of effects on the mean and high quantiles of the distribution. In the search for a data-generating model, working with joint QR addresses the entire response distribution simultaneously.

The third objective is the analysis of the most extreme observed values of the response, i.e., record-breaking events. Analyzing records presents certain challenges, such as scarcity of data or weak evidence of trend effects compared to the variability in the data. The spatiotemporal dependence in record series may not be as strong as in the bulk of the distribution or even high quantiles, necessitating preliminary statistical tools for the analysis of records. Having inference tools with sufficient power to determine the trend effect on their occurrence is crucial and that analysis would be facilitated by user-friendly software implementing these tools. Ultimately, the analysis of records benefits from the spatiotemporal modeling of their occurrence through Bayesian hierarchical spatiotemporal models, enabling full inference at unobserved locations.

Significance. This study will contribute to the methodological development of spatiotemporal statistics for the analysis of extremes within the Bayesian hierarchical framework, and of record-breaking events within a broader extent. It will offer complementary approaches to EVT which is rapidly and constantly growing. This will help address the current shortage of research for space-time EHEs, QAR, and especially records. This will provide real-world value to society by advancing our understanding of extreme events in Aragón and Spain, and offering a valuable methodology for application in other regions or contexts. Understanding the role of uncertainty in assessing trends in EHEs, high quantiles, and temperature records is crucial for addressing climate change, safeguarding public health, and ensuring food security and disaster management. It provides the foundation for informed decision-making and effective adaptation.

1.3   Outline

The outline of the Thesis is as follows. In Chapter 1, the bibliographic review of the study has been introduced. The research objectives and their thematic unit have been identified, and the value of such research argued. In Chapter 2, the methodological objectives and contributions will be summarized publication by publication, including key ideas and the challenges addressed. Also the two unpublished works will be summarized in this chapter. In Chapter 3, the full publications together with their supplementary materials will be inserted. These eight publications are the core contribution of this Thesis. In Chapter 4, the contributed works will be linked to the existing literature and discussed in relation to the objectives and research questions. In Chapter 5, a brief summary of the contributions and findings, limitations, improvement opportunities and future work will conclude the Thesis.

Chapter 2

Summary of the publications

This chapter summarizes and highlights the main methodological contributions offered in the eight publications and the two unpublished works that relate to the Thesis objectives but are not included in the compendium. The original notation from the publications is simplified and unified for ease of readability. The outline follows the three branches of research followed in this project within the study of extremes. Section 2.1 deals with Bayesian geostatistics, model-based tools, and EHEs; Section 2.2 deals with Bayesian multiple and joint QAR; and Section 2.3 deals with hypothesis testing and graphical tools for records, software and a real-world data application, and Bayesian logistic regression for the record indicators.

2.1 Spatiotemporal mean autoregression

The first approach was to develop an intensive exploratory data analysis (EDA) that would lead us to the construction of a first data-driven geostatistical spatiotemporal model. Subsequently, the aim was to exploit the model from a climatic point of view, which gave rise to novel model-based tools. This section summarizes the model proposed in Castillo-Mateo et al. (2022), the model-based tools proposed in the unpublished work Cebrián et al. (2023), and the EHE tools proposed in Cebrián, Asín, et al. (2022).

2.1.1 Spatial mean autoregression modeling

The Castillo-Mateo et al. (2022) model offers a Bayesian hierarchical spatiotemporal specification for daily maximum temperature which introduces several innovations in a standard framework. The model adopts two temporal scales, year and day within year. It captures temporal dependence through autoregression on days within year and on years. In addition to the fixed effects (linear trend, seasonality, and elevation), the complex spatiotemporal structure of temperature requires four spatial GPs to model

spatially varying intercept, slope, autoregression and residual variance parameters, and three pure error terms (years, sites within years, and sites for days within years).

Let $Y_{t\ell}(\mathbf{s})$ denote the daily maximum temperature for day ℓ, $\ell = 1, \ldots, L$ of year t, $t = 1, \ldots, T$ at location $\mathbf{s} \in D$, where D is the study region. In Castillo-Mateo et al. (2022), for all years from 1956 to 2015 associated with $t = 1$ and $T = 60$, $\ell = 1$ corresponds to May 1 and $L = 153$ corresponds to September 30, and D is Aragón. We model daily maximum temperature for day ℓ, $\ell = 2, \ldots, L$, year t, $t = 1, \ldots, T$, and location $\mathbf{s} \in D$ by

$$Y_{t\ell}(\mathbf{s}) = \mu_{t\ell}(\mathbf{s}) + \rho(\mathbf{s}) \left(Y_{t,\ell-1}(\mathbf{s}) - \mu_{t,\ell-1}(\mathbf{s})\right) + \epsilon_{t\ell}(\mathbf{s}).$$

We specify $\mu_{t\ell}(\mathbf{s})$, the fixed and random effects components, by

$$\mu_{t\ell}(\mathbf{s}) = \beta_0 + \alpha t + \beta_1 \sin(2\pi\ell/365) + \beta_2 \cos(2\pi\ell/365) + \beta_3 \mathrm{elev}(\mathbf{s}) + \gamma_t(\mathbf{s}), \qquad (2.1)$$

in which β_0 is a global intercept, α is a global linear trend coefficient that captures long-term climate change, the sin and cos terms are introduced to provide an annual seasonal component, and $\mathrm{elev}(\mathbf{s})$ is the elevation at $\mathbf{s}$. The random effects component, $\gamma_t(\mathbf{s})$, is specified as

$$\gamma_t(\mathbf{s}) = \beta_0(\mathbf{s}) + \alpha(\mathbf{s})t + \psi_t + \eta_t(\mathbf{s}),$$

where $\beta_0(\mathbf{s})$ is a mean-zero GP with an exponential covariance function having variance parameter $\sigma_{\beta_0}^2$ and decay parameter ϕ_{β_0}, and $\alpha(\mathbf{s})$ is a mean-zero GP with an exponential covariance function having variance parameter σ_α^2 and decay parameter ϕ_α. Thus, $\beta_0(\mathbf{s})$ provides local spatial adjustment to the intercept and $\alpha(\mathbf{s})$ provides local slope adjustment to the linear trend. They provide a flexible, locally linear baseline specification. Further, we add local space-time varying random effects, $\eta_t(\mathbf{s})$, and an autoregression in years for annual intercepts, $\psi_t = \rho_\psi \psi_{t-1} + \lambda_t$, to provide adjustment to this baseline.

As a result, we have introduced three pure error terms: $\lambda_t \sim$ i.i.d. $N(0, \sigma_\lambda^2)$ at yearly scale, $\eta_t(\mathbf{s}) \sim$ i.i.d. $N(0, \sigma_\eta^2)$ at sites within years, and $\epsilon_{t\ell}(\mathbf{s}) \sim$ ind. $N(0, \sigma^2(\mathbf{s}))$ at sites for days within years. Additionally, $\rho(\mathbf{s})$ and $\sigma^2(\mathbf{s})$ are, respectively, a spatially varying autoregressive term and a spatially varying variance at location $\mathbf{s}$, both of which are assumed constant over days and years. We model $\log\left\{(1 + \rho(\mathbf{s}))/(1 - \rho(\mathbf{s}))\right\} = Z_\rho(\mathbf{s}) \sim GP(Z_\rho, C(\cdot; \sigma_\rho^2, \phi_\rho))$, and $\log\{\sigma^2(\mathbf{s})\} = Z_\sigma(\mathbf{s}) \sim GP(Z_\sigma, C(\cdot; \sigma_\sigma^2, \phi_\sigma))$, again with exponential covariance function.

The entire specification is supplied distributionally in the form of a multi-level

hierarchical model as

$$[Y_{t\ell}(\mathbf{s}) \mid Y_{t,\ell-1}(\mathbf{s}), \boldsymbol{\theta}_f, \gamma_t(\mathbf{s}), \rho(\mathbf{s}), \sigma^2(\mathbf{s})]$$

$$[\gamma_t(\mathbf{s}) \mid \beta_0(\mathbf{s}), \alpha(\mathbf{s}), \psi_t, \sigma_\eta^2]$$

$$[\beta_0(\mathbf{s}) \mid \sigma_{\beta_0}^2, \phi_{\beta_0}][\alpha(\mathbf{s}) \mid \sigma_\alpha^2, \phi_\alpha][\psi_t \mid \psi_{t-1}, \rho_\psi, \sigma_\lambda^2]$$

$$[Z_\rho(\mathbf{s}) \mid Z_\rho, \sigma_\rho^2, \phi_\rho][Z_\sigma(\mathbf{s}) \mid Z_\sigma, \sigma_\sigma^2, \phi_\sigma]$$

$$[\boldsymbol{\theta}_f][\boldsymbol{\theta}_r][\sigma_\eta^2][Z_\rho][\sigma_\rho^2][\phi_\rho][Z_\sigma][\sigma_\sigma^2][\phi_\sigma],$$

where we denote the *fixed* effect parameters by $\boldsymbol{\theta}_f = (\beta_0, \alpha, \beta_1, \beta_2, \beta_3)$ and the random effects parameters by $\boldsymbol{\theta}_r = (\rho_\psi, \sigma_\lambda^2, \sigma_{\beta_0}^2, \phi_{\beta_0}, \sigma_\alpha^2, \phi_\alpha)$. Motivation for all the terms in the model comes from a comprehensive EDA.

Model fitting. Model inference is implemented in a Bayesian framework, requiring prior distributions for each of the model parameters. In general, weakly informative and, when available, conjugate prior distributions are chosen.

A Metropolis-within-Gibbs algorithm is developed with code in R to obtain MCMC samples from the joint posterior distribution. For example, the regression coefficients and the variance hyperparameters have a normal or an inverse gamma full conditional distribution, respectively. On the other hand, the spatial autoregressive parameters or the spatial variance of the response require Metropolis steps that are tuned manually.

Spatial and spatiotemporal prediction. For a given year, and day within year, the model enables kriging to unobserved locations. Given a new site $\mathbf{s}_0 \in D$, we obtain posterior predictive samples, $\{Y_{t\ell}^{(b)}(\mathbf{s}_0) : b = 1, \ldots, B\}$, using composition sampling, from the posterior distribution of $Y_{t\ell}(\mathbf{s}_0)$, which is a function of the parameters, process realizations, and $Y_{t,\ell-1}(\mathbf{s}_0)$. Posterior samples for the parameters are available from the model fitting, posterior samples for the GPs are available through usual Bayesian kriging, and posterior samples from $Y_{t,\ell-1}(\mathbf{s}_0)$ are obtained dynamically.

Model validation. Model validation is carried out using leave-one-(site)-out cross-validation (LOOCV) to compare the spatial predictive performance between the full model and reduced nested models. The metrics used are: (i) the root-mean-square error (RMSE), (ii) the mean absolute error (MAE), (iii) the continuous ranked probability score (CRPS), and (iv) the 90% coverage (CVG).

2.1.2 Model-based tools for space-time analysis

The model-based tools introduced in Cebrián et al. (2023) and Cebrián, Asín, et al. (2022) leverage predictive Bayesian modeling to enhance the spatial and temporal

analyses. These tools include predictive spatial probability surfaces and spatial extents, employing spatial cdfs for diverse climatic events. It is established a crucial link between generative models, like the aforementioned mean model, and the utilization of its output in conducting comprehensive space-time analyses, emphasizing their utility in addressing uncertainty. This framework proves particularly valuable for climate analyses, including mean analysis and extreme temperature assessments.

Probability surface and extent of an event

The events of interest are defined in terms of temperature data, which are represented by the output of a Bayesian model. Subsequently, statistical tools are derived from this Bayesian model output to quantify the probability of these events and their spatial extent.

Events. The events of interest $A_{t\ell}(\mathbf{s})$ are defined in terms of the daily maximum temperature $Y_{t\ell}(\mathbf{s})$ on day ℓ within year t and a local reference value $r(\mathbf{s})$ at location $\mathbf{s}$. The simplest event $\{Y_{t\ell}(\mathbf{s}) - r(\mathbf{s}) \geq w\}$ corresponds to the daily maximum temperature being higher than the value $r(\mathbf{s}) + w$. Persistence is studied with the events $\{Y_{t\ell}(\mathbf{s}) - r(\mathbf{s}), Y_{t,\ell+1}(\mathbf{s}) - r(\mathbf{s}) \geq w\}$ and $\{Y_{t,\ell-1}(\mathbf{s}) - r(\mathbf{s}), Y_{t\ell}(\mathbf{s}) - r(\mathbf{s}), Y_{t,\ell+1}(\mathbf{s}) - r(\mathbf{s}) \geq w\}$, which correspond to consecutive daily maximum temperatures being higher than the value $r(\mathbf{s}) + w$. The same type of events are defined in terms of averages of daily maximum temperatures, e.g., seasonal averages in a decade. Also events defined in terms of the difference between the seasonal averages in two different decades are analyzed.

Probability surface. Let $A_{t\ell}(\mathbf{s})$ denote an event related to $Y_{t\ell}(\mathbf{s})$ on day ℓ within year t at location $\mathbf{s}$. The posterior probability associated with $A_{t\ell}(\mathbf{s})$ is estimated by the proportion of occurrences in the collection of posterior predictive samples of the event $\{A_{t\ell}^{(b)}(\mathbf{s}) : b = 1, \ldots, B\}$. In other words, the mean across samples of the binary variable indicating the occurrence of the event,

$$\hat{P}(A_{t\ell}(\mathbf{s})) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(A_{t\ell}^{(b)}(\mathbf{s})).$$

These daily probabilities kriged over a fine spatial grid of D, G_D, can be drawn on a map to reveal the probability surface of the event.

Extent of an event. Thinking of $Y_{t\ell}(\mathbf{s})$ at location $\mathbf{s} \in D$ as a spatial process, we define the extent of area in D that is experiencing an event $A_{t\ell}(\mathbf{s})$ on day ℓ within year t by

$$\text{Ext}(A_{t\ell}(D)) = \frac{1}{|D|} \int_D \mathbf{1}(A_{t\ell}(\mathbf{s})) \, d\mathbf{s}, \tag{2.2}$$

where $|D|$ is the area of D. This is referred to as a block average of indicator functions. We also integrate over subregions of D to observe different temporal evolutions in different climatic regions.

Though in Cebrián, Asín, et al. (2022) we calculate explicit expressions for the first and second moments of the spatial extent, it is difficult to work with the integral above in practice. However, approximation via Monte Carlo integration is natural, i.e., replacing $\text{Ext}(A_{t\ell}(D))$ by

$$\widehat{\text{Ext}}(A_{t\ell}(D)) = \frac{1}{|G_D|} \sum_{\mathbf{s}_j \in G_D} \mathbf{1}(A_{t\ell}(\mathbf{s}_j)),$$

where G_D is a fine spatial grid of D, and $|G_D|$ is the number of grid cells in G_D. We obtain samples from $\widehat{\text{Ext}}(A_{t\ell}(D))$, $\{\widehat{\text{Ext}}(A_{t\ell}^{(b)}(D)) : b = 1, \ldots, B\}$, using the posterior predictive samples of $Y_{t\ell}(\mathbf{s}_j)$ and kriged values of $r(\mathbf{s}_j)$ for all $\mathbf{s}_j \in G_D$ if necessary.

Probability surfaces and extents in practice. With interest on summarizing the behavior of the probability surface or extent across years, or changes in their seasonal pattern, in practice we analyze the distribution of different averages of the daily probabilities or extent. For example, for the latter we consider

$$\text{Av}_{t \in \mathcal{T}, \ell \in \mathcal{L}} \text{Ext}(A_{t\ell}(D)) = \frac{1}{|\mathcal{T}||\mathcal{L}|} \sum_{t \in \mathcal{T}} \sum_{\ell \in \mathcal{L}} \text{Ext}(A_{t\ell}(D)),$$

where $\mathcal{T}$ and $\mathcal{L}$ are the subsets of years and days within year over which the daily extents are averaged, and $|\mathcal{T}|$ and $|\mathcal{L}|$ are the number of years and days within year in each subset, respectively.

Probability surface and extent of an event around the mean

In Cebrián et al. (2023), $r(\mathbf{s})$ is a quantity related to the mean, and consequently the events are related to the mean behavior of temperatures. The output used to compute the events are spatiotemporal predictions by the Castillo-Mateo et al. (2022) modeling work.

Extent of an extreme heat event

Extreme temperature might be of greater interest than mean behavior. In Cebrián, Asín, et al. (2022), the local reference value is denoted by $q(\mathbf{s})$ and represents the threshold surface given by a 95th percentile of local daily maximum temperatures. Special interest receives the extent of area in D that is experiencing extreme heat at least w degrees above local thresholds. This is a *formal* definition of the spatial extent in (2.2) of an EHE on day ℓ within year t, $A_{t\ell}(\mathbf{s}) = \{Y_{t\ell}(\mathbf{s}) - q(\mathbf{s}) \geq w\}$. In particular,

EHE is applicable to the extent when $w = 0$. With interest on EHEs, in Cebrián, Asín, et al. (2022) we use output from the Schliep et al. (2021) modeling work described in Section 1.1.1 to illustrate the notion.

2.2 Spatiotemporal quantile autoregression

Models focused on the bulk of the distribution like those with Gaussian errors often fail to accurately represent the extreme values observed in the data. In addition, the association between predictor variables and the response variable may vary across different quantile levels τ. In an EDA for extremes, very different marginal effects of covariates were observed across quantile levels. The first approach to address this was to extend the previous mean model to a spatial multiple QAR model. A challenge appears when one tries to do interpolation at unobserved locations without knowing the previous day's temperatures. Having successfully addressed these issues, the approach of modeling the entire conditional distribution simultaneously was appealing, leading to a joint model with a methodological focus. This added substantial complication to the quantile modeling, which lacks a conventional joint approach in the literature. This section summarizes the multiple QAR model proposed in Castillo-Mateo, Asín, Cebrián, Gelfand, et al. (2023), and the joint QAR model proposed in Castillo-Mateo et al. (*in press*).

2.2.1 Multiple quantile autoregression modeling

The Castillo-Mateo, Asín, Cebrián, Gelfand, et al. (2023) model extends the specification in Castillo-Mateo et al. (2022) summarized in Section 2.1.1 to the QR framework by substituting the Gaussian errors by AL errors. Again, the model adopts two temporal scales, and offers a flexible mixed effects autoregressive structure using four spatial GPs to capture space-time dependence. Using the AL specification, a method to extract marginal quantiles from the conditional quantiles in the autoregression is proposed.

Let $Y_{t\ell}(\mathbf{s})$ denote the daily maximum temperature for day ℓ within year t at location $\mathbf{s}$ as in Section 2.1.1. In Castillo-Mateo, Asín, Cebrián, Gelfand, et al. (2023), we propose the following spatiotemporal multiple QAR model. Given a quantile level $\tau \in (0, 1)$ and denoting by $Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$ the τ conditional quantile of $Y_{t\ell}(\mathbf{s})$ given $Y_{t,\ell-1}(\mathbf{s})$,

$$
\begin{aligned}
Y_{t\ell}(\mathbf{s}) &= Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s})) + \epsilon_{t\ell}^{\tau}(\mathbf{s}) \\
&= q_{t\ell}^{\tau}(\mathbf{s}) + \rho^{\tau}(\mathbf{s}) \left(Y_{t,\ell-1}(\mathbf{s}) - q_{t,\ell-1}^{\tau}(\mathbf{s}) \right) + \epsilon_{t\ell}^{\tau}(\mathbf{s}).
\end{aligned}
$$

In particular,

$$q_{t\ell}^\tau(\mathbf{s}) = \beta_0^\tau + \alpha^\tau t + \beta_1^\tau \sin(2\pi\ell/365) + \beta_2^\tau \cos(2\pi\ell/365) + \beta_3^\tau \operatorname{elev}(\mathbf{s}) + \gamma_t^\tau(\mathbf{s}),$$

is the analogue of $\mu_{t\ell}(\mathbf{s})$ in (2.1) with the same specification. On the other hand, the error term is $\epsilon_{t\ell}^\tau(\mathbf{s}) \sim$ ind. $AL(0, \sigma^\tau(\mathbf{s}), \tau)$. Again, the autoregressive coefficient in the response and the scale parameter of the error term are spatially varying.

Model fitting. Model inference is implemented in a Bayesian framework. The conditional AL distribution for all $Y_{t\ell}(\mathbf{s})$ can be expressed as normal when it is conditioned on $U_{t\ell}^\tau(\mathbf{s}) \sim Exp(1)$ as in (1.2). To complete the model we specify diffuse and, when available, conjugate priors such as normal and inverse gamma for most model parameters. A Metropolis-within-Gibbs algorithm is developed with code in C++ to obtain MCMC samples from the joint posterior distribution. Full conditional distributions for each of the parameters are derived, including the $n \times T \times (L - 1)$ reparameterized latent exponential variables $\xi_{t\ell}^\tau(\mathbf{s}) = U_{t\ell}^\tau(\mathbf{s})/\sigma^\tau(\mathbf{s})$, which conditionally follow an inverse Gaussian distribution, being this the main bottleneck of the algorithm.

Model adequacy. Model adequacy is carried out using LOOCV where the conditional quantiles are obtained using one-step ahead prediction. The metrics used are: (i) the probability $p(\tau)$ that an observation is less than the conditional quantile, (ii) the check loss or weighted mean absolute error (WMAE), and (iii) the $R^1(\tau)$, the analogue for the quantiles of the R^2 for the mean.

Marginal quantiles

Marginal quantiles are easy to interpret and interpolate, and in many climate applications they are of great interest to calculate thresholds. It is appealing to think of $q_{t\ell}^\tau(\mathbf{s})$ as a marginal quantile for $Y_{t\ell}(\mathbf{s})$, but $P(Y_{t\ell}(\mathbf{s}) \leq q_{t\ell}^\tau(\mathbf{s})) \neq \tau$. We use the conditional quantile model to extract a marginal quantile from the autoregression by adding a term $d_{t\ell}^\tau(\rho^\tau(\mathbf{s}), \sigma^\tau(\mathbf{s}))$ to $q_{t\ell}^\tau(\mathbf{s})$ to adjust the probability to τ.

For sake of simplicity, space, years, and the superscript τ are suppressed. Then, $Y_\ell = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1}) + \epsilon_\ell$ where $\epsilon_\ell \sim$ i.i.d. $AL(0, \sigma, \tau)$. We demonstrate in Castillo-Mateo, Asín, Cebrián, Gelfand, et al. (2023) that the τ marginal quantile of Y_ℓ is $q_\ell + d_\ell^\tau(\rho, \sigma)$, with $d_\ell^\tau(\rho, \sigma)$ satisfying $P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma) = \tau$ in d, and using the conditional normal form for the AL distribution,

$$\tilde{\epsilon}_\ell \mid \rho, \sigma, U_\ell, U_{\ell-1}, \ldots, U_1 \sim N\left(\frac{1 - 2\tau}{\sigma\tau(1 - \tau)} \sum_{j=0}^{\ell-1} \rho^j U_{\ell-j}, \frac{2}{\sigma^2\tau(1 - \tau)} \sum_{j=0}^{\ell-1} \rho^{2j} U_{\ell-j}\right)$$

and

$$P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma) = \int \int \cdots \int P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma, \{U_j : j = 1, \ldots, \ell\})[\{U_j\}] \, dU_1 \cdots dU_\ell.$$

The adjustment value is approximated via Monte Carlo integration and a one-dimensional root-finder.

As an example of the practical interest of calculating marginal quantiles in climate analysis, we observe that these marginal quantiles, $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) \equiv q_{t\ell}^\tau(\mathbf{s}) + d_{t\ell}^\tau(\rho^\tau(\mathbf{s}), \sigma^\tau(\mathbf{s}))$, kriged over a fine grid of D, G_D, can be drawn on a map to reveal the temperature quantile surface. Also the difference between the quantiles of two decades can be compared to study the temporal evolution of quantiles.

2.2.2 Joint quantile autoregression modeling

Multiple QR offers rich modeling capabilities but has the limitation of quantile crossing. The Castillo-Mateo et al. (*in press*) methodological work derives a characterization of the non-crossing QAR(1) model using two monotone curves. Subsequently, a novel spatial joint QAR(1) that captures spatial dependence through copula modeling and spatially varying quantiles through GPs is proposed. Also a pth order QAR(p) version, and a multivariate QAR(1) version are derived. All these proposed models are available in the R package **QAR** (Castillo-Mateo, 2023a). It is important to note that the models do not introduce any covariates in order to achieve simple conditions for non-crossing of autoregressive quantiles.

The support of the data. Linear QAR models should be cautiously interpreted as useful local approximations to more complex nonlinear global models. With a linear specification, the only non-crossing lines over an unbounded support are parallel lines yielding the constant autoregression model.

Let $\{y_t^* : t = 1, \ldots, T\}$ be a time-series data. The interest focuses on ensuring that the quantile curves do not cross for all values of y_{t-1}^* in a bounded interval. Although the region of interest for noncrossing must be assumed to be bounded, the variable space itself may still be unbounded. We take this interval to be $[0, 1]$ and implement this by making a transformation of the data, $y_t = (y_t^* - m)/(M - m)$, where $m < \min y_t^*$ and $M > \max y_t^*$. For a convenient "automatic" strategy for selecting m and M we use basic results from the theory of order statistics where $y_{(1)}^*$ is the minimum and $y_{(T)}^*$ is the maximum of the data. We propose $m = (T y_{(1)}^* - y_{(T)}^*)/(T-1)$ and $M = (T y_{(T)}^* - y_{(1)}^*)/(T-1)$.

Joint QAR model for time-series data

Koenker and Xiao (2006) offered an initial version of a joint QAR(p) model, it is described in Section 1.1.3. They required both θ_0 and θ_1 to be strictly increasing functions. In Castillo-Mateo et al. (*in press*), we offer a straightforward characterization of the required monotonicity of the QAR(1) lines in terms of two monotone curves, inspired from Tokdar and Kadane (2012).

Theorem 1. *An autoregressive specification, $Q_{Y_t}(\tau \mid y_{t-1}) = \theta_0(\tau) + \theta_1(\tau)y_{t-1}$ with $\theta_1(\tau) \in [-1, 1]$ for $\tau \in [0, 1]$, is monotonically increasing in τ for $y_{t-1} \in [0, 1]$ if and only if $Q_{Y_t}(\tau \mid y_{t-1}) = \eta_2(\tau) + (\eta_1(\tau) - \eta_2(\tau))y_{t-1}$ where $\eta_1, \eta_2 : [0, 1] \to [0, 1]$ are monotonically increasing.*

A convenient class of η's to work with are cdfs for continuous random variables with support in $[0, 1]$. A rich class would arise as probabilistic mixtures of such cdfs, leading to the general form $\eta(\tau) = \sum_{k=1}^{K} \lambda_k F(\tau \mid \mathbf{\Omega}_k)$, such that $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$, and $F : [0, 1] \to [0, 1]$ is strictly increasing for any parameters $\mathbf{\Omega}_k$. A convenient class of F's are the cdfs of the two parameter Kumaraswamy distribution. The pdf and cdf of the Kumaraswamy distribution are $f(x \mid a, b) = abx^{a-1}(1 - x^a)^{b-1}$ and $F(x \mid a, b) = 1 - (1 - x^a)^b$, respectively, where $x \in [0, 1]$ and $a, b > 0$. The Kumaraswamy distributions are a family with behavior similar to the beta distribution but much simpler, especially in the context of simulation since the pdf and cdf can be expressed in a closed form. Through simulation, we explored that these mixtures offer great flexibility.

Likelihood evaluation and model fitting. Extending to the autoregressive case the density expression noted by Tokdar and Kadane (2012), a valid joint specification of $Q_{Y_t}(\tau \mid y_{t-1})$ for all $\tau \in (0, 1)$ uniquely defines the conditional response density for $y_{t-1} \in [0, 1]$,

$$f_{Y_t}(y_t \mid y_{t-1}) = \left. \frac{1}{\frac{d}{d\tau}Q_{Y_t}(\tau \mid y_{t-1})} \right|_{\tau = \tau_{y_{t-1}}(y_t)},$$

where $\tau_{y_{t-1}}(y_t)$ solves $y_t = y_{t-1}\eta_1(\tau) + (1 - y_{t-1})\eta_2(\tau)$ in τ and is numerically approximated via a one-dimensional root-finder. Consequently, we write a valid log-likelihood score in terms of $u_t = \tau_{y_{t-1}}(y_t)$, all of the observed data $\mathbf{y} = (y_1, \ldots, y_T)^\top$, and the model parameters $\mathbf{\Omega}$ as

$$\ell(\mathbf{\Omega} \mid \mathbf{y}) = -\sum_{t=2}^{T} \log \left\{ y_{t-1}\dot{\eta}_1(u_t) + (1 - y_{t-1})\dot{\eta}_2(u_t) \right\}.$$

We conclude the model specification with the prior distribution of the parameters a's, b's, and λ's. We suggest to model the weights using the additive logistic normal

transformation and the parameters of the Kumaraswamy distribution with a weak Gaussian prior in the log scale. The root-finder used to evaluate the log-likelihood function is Brent's method. We implement an adaptive block-Metropolis sampler algorithm to obtain MCMC samples from the posterior distribution of the parameters and the conditional quantile function.

Model adequacy and comparison. Two dimensionless metrics which assess the global adequacy and comparative performance of the conditional quantile function are offered. The first metric $\tilde{p}_v$ is a standardized version of $p(\tau)$ integrated over $\tau \in (0, 1)$. The second metric $\bar{R}^1$ is a generalization of $R^1(\tau)$, again integrated over $\tau \in (0, 1)$.

Spatial joint QAR model

Let $Y_t(\mathbf{s})$ denote, in a spatial point-referenced time series, a variable for time $t = 1, \ldots, T$ at location $\mathbf{s} \in D$, where D is the study region. In Castillo-Mateo et al. (*in press*), D is Aragón, and the period spans from May 1 to September 30 in 2015. We specify the joint spatial QAR model,

$$Y_t(\mathbf{s}) = \theta_0(U_t(\mathbf{s}); \mathbf{s}) + \theta_1(U_t(\mathbf{s}); \mathbf{s})Y_{t-1}(\mathbf{s}),$$

where the θ functions are quantile and spatially varying, and the vectors $(U_t(\mathbf{s}_1), \ldots, U_t(\mathbf{s}_n))^\top$ are assumed to follow a spatial copula process.

Modeling spatial dependence. Spatial dependence is captured through spatially varying quantiles—analogue to spatially varying coefficients—and dependent quantile levels—analogue to dependent errors—as follows.
Spatially varying quantiles. For the spatially varying coefficients, we consider one cdf for each $\eta(\tau; \mathbf{s})$. In fact, at location $\mathbf{s}$, let assume $\eta_j(\tau; \mathbf{s}) = 1 - (1 - \tau^{a_j(\mathbf{s})})^{b_j(\mathbf{s})}$ with parameters $a_j(\mathbf{s})$ and $b_j(\mathbf{s})$ ($j = 1, 2$). We introduce four independent GPs for the a's and b's on the log scale. In particular, we model $\log a_j(\mathbf{s}) \sim GP(a_j, C(\cdot; \sigma_{a_j}^2, \phi_{a_j}))$ and $\log b_j(\mathbf{s}) \sim GP(b_j, C(\cdot; \sigma_{b_j}^2, \phi_{b_j}))$ where $C(\cdot; \sigma^2, \phi)$ is an exponential covariance function with variance parameter σ^2 and decay parameter ϕ.
The spatial copula process. We take the processes $U_t(\mathbf{s})$'s to follow a Gaussian copula for each t, induced by a stationary spatial GP. In the spirit of Chen and Tokdar (2021), we define
$$U_t(\mathbf{s}) = \Phi(Z_t(\mathbf{s})), \quad Z_t(\mathbf{s}) = W_t(\mathbf{s}) + \epsilon_t(\mathbf{s}),$$
$$W_t(\mathbf{s}) \sim GP(0, C(\cdot; \gamma, \phi)), \quad \epsilon_t(\mathbf{s}) \sim \text{i.i.d. } N(0, 1 - \gamma).$$
The process $W_t(\mathbf{s})$ captures spatial dependence while $\epsilon_t(\mathbf{s})$ is independent pure error. The parameter $\gamma \in [0, 1]$ determines the proportion of spatial and independent

variation. With this approach, the Gaussian copula density has correlation matrix $\gamma R(\phi) + (1 - \gamma)I$ where $R(\phi)$ is the $n \times n$ correlation matrix induced by $C(\cdot; \gamma, \phi)$.

Likelihood evaluation and spatial interpolation. It is convenient to first obtain the joint distribution for all data, **y**. By Sklar's theorem, the joint conditional density of **y** can be partitioned into a marginal part and a copula part. Subsequently, we find the expression of the log-likelihood function for the spatial QAR, and after giving weakly informative priors, inference proceeds in a similar way as in the univariate case.

With the proposed model we can interpolate the autoregressive coefficient, or conditional quantiles to any desired location in the study region given any proposed or reference value for the previous time's observation at that location.

2.3 Spatiotemporal record-breaking events

In parallel with QR, the focus was centered on the exploration of record-breaking events. As far as I know, there were only a few specific inference tools based on the occurrence of records and no complete modeling approach. The need to objectively evaluate the impact of climate change on the occurrence of records at local and regional levels is vital in many climate-related issues. Both approaches offer valuable complementary insights. A spatiotemporal model provides extensive opportunities for full inference, although its proper implementation might require expertise in multi-level stochastic modeling. On the other hand, the inference tools are straightforward, relying on conventional hypothesis testing and easily interpretable graphical tools. They are now readily accessible through a freely available software package that is particularly suitable for researchers less experienced in statistics or for preliminary EDAs preceding more complex modeling.

This section summarizes the hypothesis tests and graphical tools proposed in Cebrián, Castillo-Mateo, et al. (2022) and Castillo-Mateo (2022); the R package **RecordTest** (Castillo-Mateo, 2023b) described in Castillo-Mateo, Cebrián, et al. (2023a) that implements these and other inference tools; and their extensive application in Castillo-Mateo, Cebrián, et al. (2023b) to the Spanish dataset. This summary concludes with the model proposed in the unpublished work Castillo-Mateo, Gelfand, et al. (2023), which offers a very novel Bayesian hierarchical specification for the record indicators.

2.3.1 Record tests

In general, records are considered for individual calendar days across years, yielding 365 time series for each location, so when necessary, $(I_{1\ell}(\mathbf{s}), \ldots, I_{T\ell}(\mathbf{s}))^\top$ denote the sequence of record indicators across years within day ℓ at location $\mathbf{s}$.

Non-stationarity detection

The null hypothesis H_0 of the tests is that the probability of record at each time t in a series of length T is the probability of record under the stationary condition characterized by c.i.i.d. series. Under climate change, higher probabilities of upper record than in c.i.i.d. series are expected. The null and one-sided alternative hypotheses are, respectively,

$$
\begin{aligned}
H_0 &: p_t = 1/t, & t = 2, \ldots, T, \\
H_1 &: p_t > 1/t, & \text{for at least one } t = 2, \ldots, T.
\end{aligned}
\tag{2.3}
$$

In Cebrián, Castillo-Mateo, et al. (2022), we propose a family of distribution-free tests to detect deviations from c.i.i.d. series using the likelihood function of the record indicators under the null hypothesis. If the analyzed series is formed by continuous independent observations with no seasonal behavior, deviations from the c.i.i.d. hypothesis suggest a change in the mean, a change in the shape of the pdf, or a combination of both. The tests assume L mutually independent series of length T available. These series can be a subset of uncorrelated series extracted from series measured at different spatial points or series obtained from splitting the original data.

Tests based on the number of records. The most basic statistic is the total number of records up to time T in the L series,

$$
N = \sum_{\ell=1}^{L} N_{T\ell} = \sum_{\ell=1}^{L} \sum_{t=1}^{T} I_{tm}.
$$

Under the null hypothesis, N is asymptotically normal when T and/or M tend to ∞. Diersen and Trenkler (1996) suggested to include weights in the record indicators according to their position in the series to improve the power of the tests, i.e.,

$$
\mathcal{N} = \sum_{\ell=1}^{L} \sum_{t=1}^{T} w_t I_{t\ell}.
$$

They suggested the use of linear weights $w_t = t - 1$ because records become less likely for increasing time, and an occurrence at high t gives more evidence against the null hypothesis. Although different in nature, using the score—the gradient of

the log-likelihood function—we found in Cebrián, Castillo-Mateo, et al. (2022) an equivalent statistic with weights $w_t = t^2/(t-1)$ ($w_1 = 0$), which give the locally most powerful unbiased score test. Under the null hypothesis, $\mathcal{N}$ is still asymptotically normal in L. The p-value of the test is $P(Z \geq (\mathcal{N}_0 - 0.5 - \mu)/\sigma)$, where Z is a standard normal variable, μ and σ are the mean and standard deviation of the statistic under the null hypothesis, $\mathcal{N}_0$ is the observed statistic, and 0.5 is a continuity correction.

Further, we use the asymptotic normal distribution of the statistics N and $\mathcal{N}$ to compute reference intervals (RIs) under the null hypothesis of the number of records and of the weighted number of records up to time t. These RIs together with the observed values are drawn across t to observe deviations from the null hypothesis.

Joining information from different types of records. Foster and Stuart (1954) already observed that there are four types of records available in a series. The upper and lower records in the forward series, $(Y_1, \ldots, Y_T)^\top$, and in the backward series, $(Y_T, \ldots, Y_1)^\top$. We use the superscripts L and B to denote lower records and records in backward series, respectively.

The null hypothesis is the same for all types of record. Joining information from different types of records allows to study the tails of the distribution, not only the observed record-breaking events in the classical sense. The main statistics we consider are: (i) $\mathcal{N}_{upp} = \mathcal{N} - \mathcal{N}^B$ based on upper records, for the analysis of an increasing trend in the upper tail; (ii) $\mathcal{N}_{low} = \mathcal{N}^{BL} - \mathcal{N}^L$ based on lower records, for the analysis of an increasing trend in the lower tail; (iii) $\mathcal{N}_{both} = \mathcal{N} - \mathcal{N}^L - \mathcal{N}^B + \mathcal{N}^{BL}$ based on four types of records, for the analysis of an increasing trend in both tails. The sign of the statistic for each type of record is positive or negative according to whether a higher or lower number of records is expected under the alternative hypothesis for that type of record. Under the null hypothesis, all these statistics have an asymptotic normal distribution in L, and/or in T only if no weights are included as above. They have zero mean, and using notions from order statistics we derive in Cebrián, Castillo-Mateo, et al. (2022) the covariances between all types of record indicators, necessary to obtain the variance of the statistics. The p-value is obtained as usual.

Again, we use the asymptotic normal distribution of these statistics to compute RIs under the null hypothesis for the statistic up to each time t. These RIs together with the observed values are drawn across t to observe deviations from the null hypothesis.

Additional record tests, a powerful approach for joining dependent p-values coming from individual tests for different types of records, and graphical tools based on probabilities and times of record are given in Cebrián, Castillo-Mateo, et al. (2022).

Monte Carlo analysis of size and power. The size and power of these tests are assessed and compared through Monte Carlo simulations. The estimated power is obtained under several scenarios of the alternative hypothesis; e.g., models with linear (the LDM), concave, and convex trends, and normal, generalized Pareto, and generalized extreme value error terms. The record tests result even more powerful than the widely used Mann-Kendall test—used to detect trends in the mean—in some scenarios where an error term with a light or bounded tail is considered. The power of the tests based on different types of records is higher than those using only one type.

Change-point detection

In Castillo-Mateo (2022), I propose three distribution-free statistics to detect a change-point if the record occurrence stops being stationary. The statistics test the null hypothesis H_0 in (2.3) against the two-sided alternative hypothesis

$$H_1 : p_t = 1/t, \quad t = 1, \ldots, t_0, \qquad \text{and} \qquad p_t \neq 1/t, \quad t = t_0 + 1, \ldots, T,$$

where t_0 is the change-point. The main test statistic for change-point detection in the record occurrence is $\mathcal{K} = \max_{1 \leq t \leq T} |K_T(t)|$, where

$$K_T(t) = \frac{N_t - E(N_t)}{\sqrt{Var(N_T)}} - \frac{Var(N_t)}{Var(N_T)} \frac{N_T - E(N_T)}{\sqrt{Var(N_T)}}.$$

The change-point estimate is defined as $\hat{t}_0 = \arg\max_{1 \leq t \leq T} |K_T(t)|$. Under the null hypothesis, the distribution of $\mathcal{K}$ is Kolmogorov in the limit as $T \to \infty$. The p-value is computed in the usual way, and a significant change-point occurs at time $\hat{t}_0$ if the null hypothesis is rejected.

The change-point statistic can also consider L series and weights for the record indicators with $w_t = \sqrt{t^2/(t-1)}$ ($w_1 = 0$). I proved that the statistic does no longer follow the Kolmogorov distribution in the limit, but the p-value can be estimated using Monte Carlo simulations under the null hypothesis.

Monte Carlo analysis of size, power, and change-point estimate. The size, power, and ability of these statistics to detect the actual change-point are assessed through Monte Carlo simulations. The estimator is right-sided biased, which can be usefully interpreted as the time when the underlying process that drives the temperature distribution truly affects the records in the observed data. In other words, the change-point is determined by the occurrence of a record; if no records are observed, there is no change-point. This bias decreases significantly when the number of series L is increased or when the effect of the change becomes greater under the alternative hypothesis.

2.3.2 The R package RecordTest and real-world application

The R package **RecordTest** (Castillo-Mateo, 2023b) provides EDA and inference tools based on theory of records to describe the record occurrence and detect trends and change-points in time series. In particular, **RecordTest** consists of graphical tools (Benestad, 2003, 2004; Cebrián, Castillo-Mateo, et al., 2022), distribution-free tests for trend in location, variation or non-stationarity in the tails (Foster & Stuart, 1954; Diersen & Trenkler, 1996; Cebrián, Castillo-Mateo, et al., 2022), and change-point detection tests (Castillo-Mateo, 2022), all of them based on the record occurrence.

We describe **RecordTest** in Castillo-Mateo, Cebrián, et al. (2023a). We detail how to prepare the data: first splitting the series with `series_split()`, and then extracting a subset of uncorrelated series ready to be used as argument in the functions with `series_uncor()`. The main tests are comprised in the functions: `N.test()` for those based on the number of records, `foster.test()` for those based on joining information from different types of records, and `change.point()` for the change-point tests. And graphical tools to show: the evolution of the number of records with `N.plot()`, the statistics joining different types of records with `foster.plot()`, and the probabilities of record with `p.plot()`. Other tests, graphical tools, and additional functions are also available in the package. As far as I know, **RecordTest** is the only statistical software package currently available for the analysis of record-breaking events; and it is available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=RecordTest.

Real-world data application. We extend and use **RecordTest** in Castillo-Mateo, Cebrián, et al. (2023b) to incorporate permutation tests and analyze records over peninsular Spain in the period 1960–2021. Permutation tests allow to apply the tests to dependent series in time or space. This property allows to evaluate the null hypothesis globally pooling data in a region. Permutation tests only rely on the assumption of exchangeability under the null hypothesis. A sample is exchangeable if any permutation of it has the same joint probability distribution. In the record tests, there is a sample of $t = 1, \ldots, T$ observations of a vector of L variables $(Y_{t1}, \ldots, Y_{tL})$. Under the null hypothesis, the T observations of the vector $(Y_{t1}, \ldots, Y_{tL})$ are independent with the same multivariate distribution, so that permutations in t are exchangeable.

In Castillo-Mateo, Cebrián, et al. (2023b), we compare the behavior of the occurrences of records in different Spanish regions, in different periods of the year, and in different signals such as daily or annual maximum temperature. Significant evidences of the effect of an increasing trend in the occurrence of upper extremes are found in

most of Spain. The effects are heterogeneous within the year, being autumn the season where the effects are weaker and summer where they are stronger. Concerning the spatial variability, the Mediterranean and the Atlantic region are the areas where the effects are more and less clear, respectively.

2.3.3   Spatial modeling of record-breaking events

Modeling the occurrence of records in Spain is challenging due to diverse climate and topography, leading to regional variations that make it difficult to identify specific drivers of temperature records and to distinguish between climate change and natural variability. In this section, and in more detail in Castillo-Mateo, Gelfand, et al. (2023), we address these challenges through the use of a Bayesian hierarchical model that accounts for large- and small-scale variation—mean behavior as well as spatial and temporal stochastic dependence—in order to obtain high-resolution posterior predictive realizations that can be used for needed inference. Although the probabilistic properties of records have been widely applied, they model the *marginal* probabilities of record. The contribution of this section is to propose space-time *conditional* models for the record indicators. This pooling of data with joint modeling is especially important when studying records because these events, though rare, are highly dependent.

Specifically, we model the probability of record $p_{t\ell}(\mathbf{s})$ at day ℓ within year $t = 2, \ldots, T$ at location $\mathbf{s}$ using a version of a logistic regression model. With regard to the Bernoulli distributions of the indicators, we introduce suitable fixed effects and daily spatial random effects, specified within a fully Bayesian hierarchical structure on the logit scale. For long-term trends, we consider forms like $p_t = 1/t^\alpha$ which, on the logit scale yields $-\log(t^\alpha - 1)$. To simplify to an asymptotically equivalent linear expression, we adopt $-\alpha \log(t-1)$ and $\alpha = 1$ gives the probability of record in the CRM. Further, the model also incorporates seasonality, persistence, geographical covariates, and useful interactions with the long-term trends, i.e., the α coefficient could be a linear expression of covariates expressing the rate of deviation from stationarity.

Data and exploratory analysis

Data precision and tied records.   The Spanish dataset is measured to the nearest 1/10th of a °C. This rounding/discretization results in some ties when records are identified. To deal with ties, an observation that is at least as large as any previous observation is called a *weak* record. Here, we define a *tied* record in terms *equal* rather than *higher or equal*, i.e., an r-tied record ($r \geq 2$) arises when an observation shares the same value with $r-1$ preceding weak records. To accommodate the tied records within the Bayesian framework, we assume that each of the true daily temperatures roundings

are i.i.d., following a distribution on the interval $(\text{observed} - 0.05, \text{observed} + 0.05)$. Therefore, for an r-tied record in the rounded data series, the probability of it being a record in the *true* daily temperature series is $1/r$. The MCMC model fitting algorithm addresses this characteristic of temperature data by sampling from the $\text{Bernoulli}(1/r)$ distribution of the indicators corresponding to r-tied records at the beginning of each iteration.

Exploring the occurrence of records. An extensive EDA was conducted to evaluate departures from stationarity in the occurrence of records, identify related regressors, and enable a data-driven variable selection for the model. The covariates studied as fixed effects aim to capture spatio-temporal variability; i.e., annual trend, persistence, seasonal behavior, and geographic features.

Two types of exploratory tools were used in the analysis: (i) graphical tools, and (ii) exploratory global or local logit models for all or individual sites, respectively. Frequentist models with different regressors were compared using the AIC as a goodness of fit measure. Below I show a couple of examples developed in the EDA.

The model in the logit scale must at least include a long-term trend, $\log(t-1)$, which in the stationary case would have a coefficient roughly equal to -1. To explore whether that term is enough to capture temporal evolution across years, the left plot in Figure 2.1 shows $t \times \hat{p}_t$ against t, whose expected value under stationarity is 1. An empirical estimate of p_t was obtained by averaging across space and days within year, $\hat{p}_t = \sum_i \sum_\ell I_{t\ell}(\mathbf{s}_i)/(n \times 365)$. The estimates of p_t are getting higher than expected in a stationary climate. To allow a flexible modeling of the deviation from stationarity, the inclusion of a polynomial function of $\log(t-1)$ was considered. The AIC concluded that a third order polynomial is unnecessary and the second is preferred.

To study the dependence between the occurrence of records in two consecutive days, we considered the joint distribution $[I_{t\ell}(\mathbf{s}), I_{t,\ell-1}(\mathbf{s})]$ expressed in terms of 2×2 tables. The evolution of this dependence across years was studied with two-way tables obtained by summing across space and days within year. The empirical log odds ratios (ORs) provide a useful tool for learning about persistence,

$$LOR_t = \log \frac{(n_{t,11} + 0.5)(n_{t,00} + 0.5)}{(n_{t,01} + 0.5)(n_{t,10} + 0.5)},$$

where $n_{t,jk} = \sum_i \sum_\ell \mathbf{1}(I_{t\ell}(\mathbf{s}_i) = j, I_{t,\ell-1}(\mathbf{s}_i) = k)$ for $j, k \in \{0,1\}$ denotes the frequencies in each cell of the table, and 0.5 is a customary continuity correction. For notation convenience $I_{t0}(\mathbf{s}) \equiv I_{t-1,365}(\mathbf{s})$. The LOR_t compares the probabilities of record given a record or a non-record the previous day. Values close to 0 express independence while positive values capture persistence. The right plot in Figure 2.1
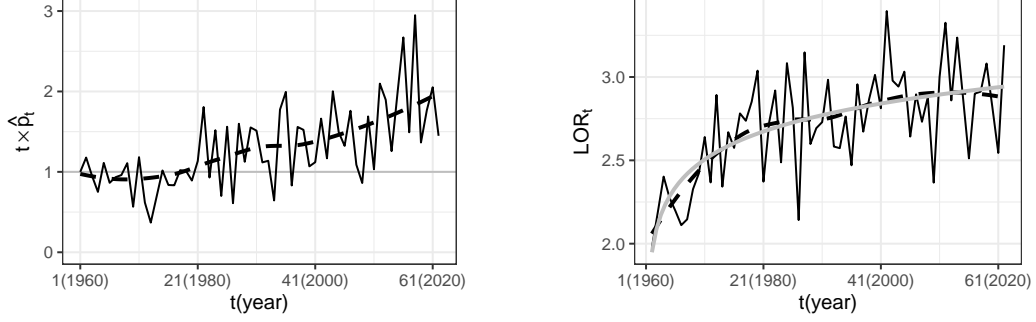
33

Figure 2.1: Left: Evolution of $t \times \hat{p}_t$ against t with reference value 1. Right: LOR_t comparing the probability of record given a record the previous day and the probability of record without a record the previous day, together with a linear model fitted to $\log(t-1)$ in a solid gray line. LOESS curves in dashed black lines.

that shows the LOR_t against t indicates a strong persistence increasing across years. The linear relationship observed between LOR_t and $\log(t-1)$ suggests the inclusion of the autoregressive term $I_{t,\ell-1}(\mathbf{s})$ and its interaction with the long-term trend in the model. Given the strong persistence of temperature, the introduction of second-order autoregressive terms was also considered with analogous results.

Additional analyses found a significant seasonal behavior in the long-term trends and spatial variability primarily explained by distance to the coast. Elevation did not appear to improve the AIC of the models once distance to the coast was introduced.

Model specifics

A rich spatial logistic regression model across days, for annual temperature records is proposed. Let $I_{t\ell}(\mathbf{s})$ denote the record indicator of the daily maximum temperature for day ℓ, $\ell = 1,\ldots,365$, within year t, $t = 1,\ldots,T$, at location $\mathbf{s} \in D$, where D is the study region. In particular, D is peninsular Spain, and the series go from January 1, 1960 to December 31, 2021; so, $t = 1$ corresponds to 1960 and $T = 62$ to 2021. We model record indicators beginning with day $\ell = 3,\ldots,365$, year $t = 2,\ldots,T$, and location $\mathbf{s}$ according to

$$I_{t\ell}(\mathbf{s}) \mid I_{t,\ell-1}(\mathbf{s}), I_{t,\ell-2}(\mathbf{s}) \sim \text{Bernoulli}(g^{-1}(\eta_{t\ell}(\mathbf{s}))) \quad \text{with} \quad \eta_{t\ell}(\mathbf{s}) = \mathbf{x}_{t\ell}(\mathbf{s})\boldsymbol{\beta} + w_{t\ell}(\mathbf{s}),$$

where $g(p) = \log\{p/(1-p)\}$ is the logit link function. Here, $p_{t\ell}(\mathbf{s}) = g^{-1}(\eta_{t\ell}(\mathbf{s}))$ is the probability of a record for day ℓ, year t, and location $\mathbf{s}$, with fixed effects $\mathbf{x}_{t\ell}(\mathbf{s})\boldsymbol{\beta}$ and random effects $w_{t\ell}(\mathbf{s})$. The $\mathbf{x}_{t\ell}(\mathbf{s}) = (1, x_{t\ell 1}(\mathbf{s}),\ldots,x_{t\ell k}(\mathbf{s}))$ are $k+1$ covariates measured on day ℓ, year t, and location $\mathbf{s}$ with $\boldsymbol{\beta}$ a column vector of length $k+1$ of regression coefficients. The $w_{t\ell}(\mathbf{s})$ are space-time correlated errors.

We first supply the fixed effects term. Apart from the intercept, the entries are: (i) the first and second degree polynomials of $\log(t-1)$; (ii) persistence terms to capture

34

first and second autoregressive dependence, $I_{t,\ell-1}(\mathbf{s})$, $I_{t,\ell-2}(\mathbf{s})$, and $I_{t,\ell-1}(\mathbf{s}) \times I_{t,\ell-2}(\mathbf{s})$, and their interactions with $\log(t-1)$; (iii) seasonal terms including one harmonic, $\sin_\ell = \sin(2\pi\ell/365)$ and $\cos_\ell = \cos(2\pi\ell/365)$, and their interactions with the second-order long-term trend; and (iv) spatial terms to model the effect of the logarithm of the distance to the coast, $\log(\mathrm{dist}(\mathbf{s}))$, including interaction with the second-order long-term trend and persistence. In summary, we have the intercept and the following 20 predictors,

$$
\begin{aligned}
\mathbf{x}_{t\ell}(\mathbf{s}) = &(1, \log(t-1), [\log(t-1)]^2, I_{t,\ell-1}(\mathbf{s}), I_{t,\ell-2}(\mathbf{s}), I_{t,\ell-1}(\mathbf{s}) \times I_{t,\ell-2}(\mathbf{s}), \\
&\log(t-1) \times I_{t,\ell-1}(\mathbf{s}), \log(t-1) \times I_{t,\ell-2}(\mathbf{s}), \log(t-1) \times I_{t,\ell-1}(\mathbf{s}) \times I_{t,\ell-2}(\mathbf{s}), \\
&\sin_\ell, \cos_\ell, \sin_\ell \times \log(t-1), \cos_\ell \times \log(t-1), \sin_\ell \times [\log(t-1)]^2, \cos_\ell \times [\log(t-1)]^2 \\
&\log(\mathrm{dist}(\mathbf{s})), \log(\mathrm{dist}(\mathbf{s})) \times \log(t-1), \log(\mathrm{dist}(\mathbf{s})) \times [\log(t-1)]^2, \\
&\log(\mathrm{dist}(\mathbf{s})) \times I_{t,\ell-1}(\mathbf{s}), \log(\mathrm{dist}(\mathbf{s})) \times I_{t,\ell-2}(\mathbf{s}), \log(\mathrm{dist}(\mathbf{s})) \times I_{t,\ell-1}(\mathbf{s}) \times I_{t,\ell-2}(\mathbf{s})).
\end{aligned}
$$

While 20 fixed effects terms may seem excessive, in explaining the roughly $900,000$ observations we do find all of them significant. Motivation for all the terms in the model comes from the novel EDA and model comparison.

We introduce explicit spatial and temporal dependence through random effects. We considered different structures for modeling these random effects, with the full model considering $w_{t\ell}(\mathbf{s})$ space-time correlated errors following GPs with mean $w_{t\ell} \sim$ i.i.d. $N(0, \sigma_1^2)$ and a common exponential covariance function having variance and decay parameters σ_0^2 and ϕ_0, respectively.

For prediction, the autoregressive model requires an initial condition for $I_{t1}(\mathbf{s})$ and $I_{t2}(\mathbf{s})$, the first and second values in year t. We model them as $\eta_{t1}(\mathbf{s}) = \mathbf{x}_{t1}(\mathbf{s})\boldsymbol{\beta}_1 + w_{t1}(\mathbf{s})$ and $\eta_{t2}(\mathbf{s}) = \mathbf{x}_{t2}(\mathbf{s})\boldsymbol{\beta}_2 + w_{t2}(\mathbf{s})$. The covariate vectors are reduced to $\mathbf{x}_{t1}(\mathbf{s}) = (1, \log(t-1), I_{t-1,365}(\mathbf{s}))$ and $\mathbf{x}_{t2}(\mathbf{s}) = (1, \log(t-1), I_{t1}(\mathbf{s}))$. The $w_{t\ell}(\mathbf{s})$ for $\ell = 1, 2$ are modeled as above, each a GP with mean $w_{t\ell} \sim N(0, \sigma_{1,\ell}^2)$ and exponential covariance function having variance $\sigma_{0,\ell}^2$ and the same decay parameter ϕ_0.

Prior specification and model fitting. Model inference is implemented in a Bayesian framework. Adopting the data augmentation strategy by Held and Holmes (2006), each binary data can be seen as the indicator of a latent standard logistic variable exceeding a particular threshold. A standard logistic variable can be represented as a scale mixture of normal form, allowing the same conjugate priors as in a standard spatial linear model. Consequently, we specified weak normal priors for the regression coefficients, weak inverse gamma priors for the variances, and a gamma prior for the decay parameter. A Metropolis-within-Gibbs algorithm is developed with code in C++ to obtain MCMC samples from the joint posterior distribution. Each

iteration of the MCMC requires the simulation of a truncated normal variable and a variable following the asymptotic distribution of the Kolmogorov-Smirnov statistic for every data, which is the main bottleneck of the algorithm.

Spatial prediction and inference. For a given year, and day within year, the model enables kriging—again, through Bayesian kriging and composition sampling—for the record indicators $I_{t\ell}(\mathbf{s}_0)$ or their probabilities $p_{t\ell}(\mathbf{s}_0)$ to unobserved locations $\mathbf{s}_0 \in D$.

Employing G_D, a fine spatial grid for D, posterior samples from $I_{t\ell}(\mathbf{s})$ can be realized at each location in G_D for every day ℓ, $\ell = 1, \ldots, 365$, within year t, $t = 2, \ldots, T$. So, we can make inference about any feature of interest related to the occurrence of records. Let $1 \leq \ell_1 \leq \ell_2 \leq 365$ and $1 \leq t_1 \leq t_2 \leq T$. Then, a general feature of primary interest is the average cumulative number of records across days from ℓ_1 to ℓ_2 and across years from t_1 to t_2. It is defined as

$$\bar{N}_{t_1:t_2,\ell_1:\ell_2}(\mathbf{s}) = \frac{1}{\ell_2 - \ell_1 + 1} \sum_{t=t_1}^{t_2} \sum_{\ell=\ell_1}^{\ell_2} I_{t\ell}(\mathbf{s}).$$

The average total number of records $\bar{N}_T(\mathbf{s})$ arises with $\ell_1 = 1$, $\ell_2 = 365$, $t_1 = 1$ and $t_2 = T$. Comparison between the average number of records predicted by the model and the expected number of records under the stationary case, $E_0[\bar{N}_{t_1:t_2,\ell_1:\ell_2}(\mathbf{s})] = \sum_{t=t_1}^{t_2} 1/t$ may be of interest. The ratio expression

$$R_{t_1:t_2,\ell_1:\ell_2}(\mathbf{s}) = \frac{\bar{N}_{t_1:t_2,\ell_1:\ell_2}(\mathbf{s})}{E_0[\bar{N}_{t_1:t_2,\ell_1:\ell_2}(\mathbf{s})]}, \tag{2.4}$$

captures records expected by the model compared to a scenario without climate change. To compare seasons, the statistics $R_{t_1:t_2,\text{DJF}}(\mathbf{s})$, $R_{t_1:t_2,\text{MAM}}(\mathbf{s})$, $R_{t_1:t_2,\text{JJA}}(\mathbf{s})$, and $R_{t_1:t_2,\text{SON}}(\mathbf{s})$ were considered, where DJF is winter (December, January, and February), MAM is spring, JJA is summer, and SON is autumn, with obvious notation.

Computing the above quantities for all $\mathbf{s} \in G_D$ we can draw maps of the posterior mean or borders of the 90% credible intervals (CIs) of the quantities of interest. This enables a useful picture of the spatio-temporal characteristics of the occurrence of records and assessment of regions and time periods with higher risk of exceeding temperatures above all of the previous measurements. Probability surfaces and the extent of record surface that extend the ideas from Section 2.1.2 to record-breaking occurrences are also proposed.

Model comparison. The spatial predictive performance of several models was compared using 10-fold cross-validation. The metrics used are: (i) the proper scoring rule for binary events Brier score (BS), (ii) the area under the receiver

operating characteristic curve (AUC), and (iii) an absolute deviation (AD) statistic, $|N_{t\ell}(\mathbf{s}) - N_{t\ell}^{(b)}(\mathbf{s})|$ comparing observed values and posterior predictive samples.

Two periods for record breaking were considered. One employs the first 30 years of the series when records are more frequent, the second employs the last 31 years when records are more rare. The simplest model considered for illustration was the stationary model $\eta_{t\ell}(\mathbf{s}) = -\log(t-1)$. For 1961–1990, its $100 \times BS$ was 7.88 and the AUC was 0.733; for 1991–2021, its $100 \times BS$ was 3.23 and the AUC was 0.514. The full model presented above obtained the best results. For 1961–1990, its $100 \times BS$ was 4.32 and the AUC was 0.944; for 1991–2021, its $100 \times BS$ was 2.13 and the AUC was a remarkable 0.924.

Results

In Castillo-Mateo, Gelfand, et al. (2023), we included an extensive list of results, e.g., the posterior distribution of the model parameters, maps based on the posterior mean and CIs for the number of records, or the time series of the extent of record surface. In this section I show two illustrative results.

Of special interest is the number of records in the past decade, as it provides a current picture of the climate and is not influenced by the high probability of occurrence in the early stages. Analyzing records by season is also important because each season has distinct spatial and temporal patterns. Figure 2.2 shows the posterior mean of the ratio $R_{53:62,\ell_1:\ell_2}(\mathbf{s})$ in (2.4) for years in 2012–2021 by season. The model estimates that the global warming trends have increased the number of records expected in the past decade almost two-fold—1.93 $(1.89, 1.98)$—, which suggests an estimated probability of around 50% that the records of the past decade over peninsular Spain would not have occurred in the absence of climate warming. By season the values are 1.88 $(1.80, 1.97)$ in winter, 1.81 $(1.72, 1.89)$ in spring, 2.13 $(2.04, 2.22)$ in summer, and 1.92 $(1.83, 2.01)$ in autumn. The number of records in the past decade is higher than in the stationary case everywhere, and this difference is significant for any point in the region. The percentage of area that has a significantly higher number of records is 100% for winter, 90.1% for spring, 98.4% for summer, and 99.9% for autumn. During this period, summer presents greater warming on average but also greater spatial variability compared to, e.g., winter.

The model is also useful to obtain maps for the probability of record across days during a particular heatwave. On August 14 ($\ell = 225$), 2021, the province of Cordoba in Spain set the highest temperature ever recorded in the country, while other stations also broke their own records (WMO, 2022). To demonstrate the spatial and temporal extent of this absolute record, Figure 2.3 shows the posterior mean of the probabilities
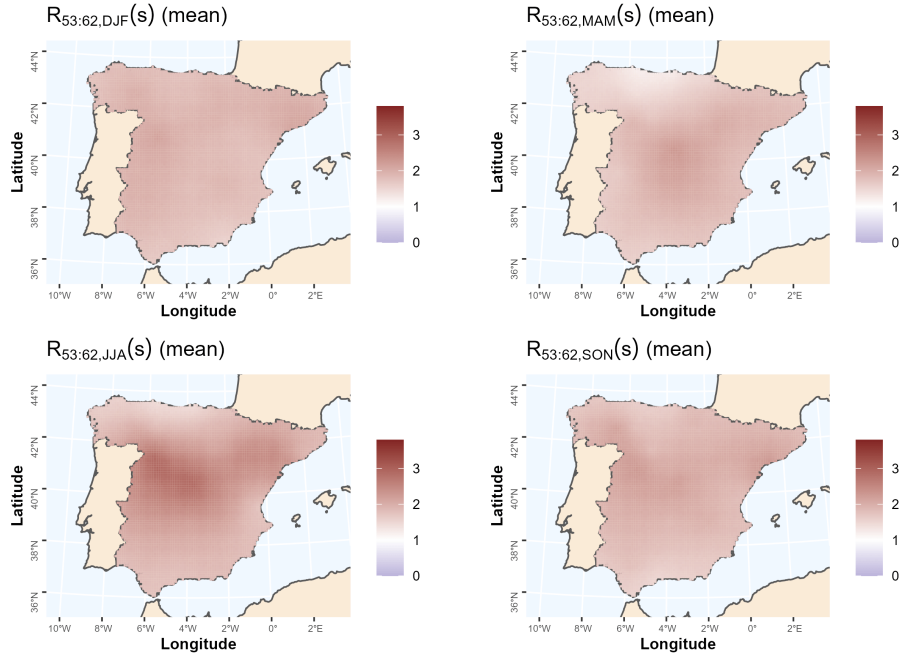
Figure 2.2: Map of the posterior mean of $R_{53:62,DJF}(\mathbf{s})$ (winter, top–left), $R_{53:62,MAM}(\mathbf{s})$ (spring, top–right), $R_{53:62,JJA}(\mathbf{s})$ (summer, bottom–left), and $R_{53:62,SON}(\mathbf{s})$ (autumn, bottom–right).

of record $p_{62\ell}(\mathbf{s})$ on days $\ell = 222, \ldots, 229$ within year 2021. This specific episode demonstrates the dynamics of persistence as well as the spatial structure of dependence. The beginning of the effects of the heatwave on the occurrence of records was observed on day 224, with posterior mean probabilities surpassing 0.5 in the northeast. Over the following three days, the heatwave continued to evolve, with high probabilities observed across most regions, except for the coast, and with the area of high probabilities gradually diminished towards the southwest.
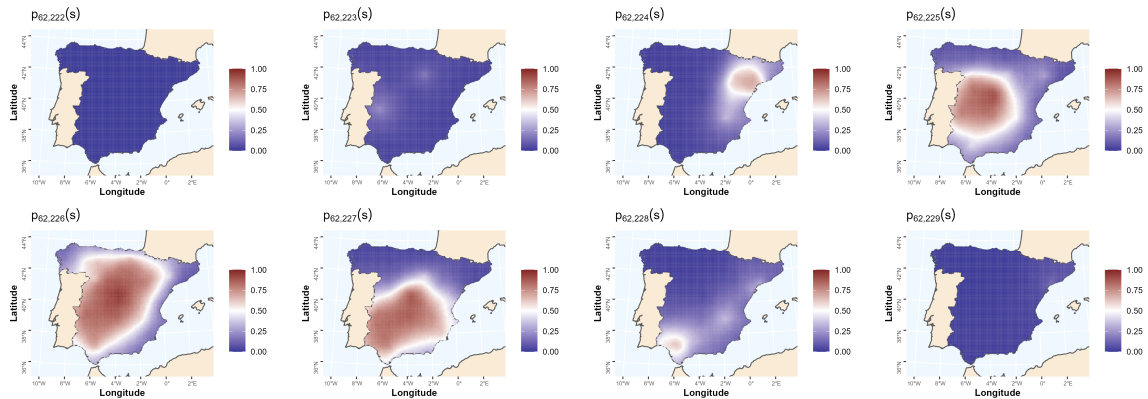

Figure 2.3: Maps of the posterior mean of $p_{62\ell}(\mathbf{s})$ on days $\ell = 222, \ldots, 229$ within year 2021.

Chapter 3

Publications

3.1 Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain

This manuscript was published in:

Castillo-Mateo, J., Lafuente, M., Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2022). Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain. *Journal of Agricultural, Biological and Environmental Statistics, 27*(3), 487–505. https://doi.org/10.1007/s13253-022-00493-3 [arXiv:2201.01687]

And it was disseminated (speaker emphasized) in:

- ***Castillo-Mateo, J.***, Gelfand, A. E., Cebrián, A. C., Asín, J., & Abaurrea, J. (2022, June 7–10). *Spatio-temporal modeling and analysis of daily maximum temperatures* [Contributed talk]. XXXIX Congreso Nacional de Estadística e Investigación Operativa, Granada, Spain.

- ***Castillo-Mateo, J.***, Lafuente, M., & Gelfand, A. E. (2021, July 12–16). *Hierarchical spatio-temporal modeling of daily maximum temperatures: A case study in the Ebro river basin* [Poster session]. Valencia International Bayesian Analysis Summer School 4, Valencia, Spain.

He came into my office one day with that classic spatial data set on scallop catches in the Atlantic Ocean, and asked, "What can I do with this stuff, and what the heck is a variogram?" I said, "I have no clue."

Alan E. Gelfand, in Carlin and Herring (2015)

Check for updates

Spatial Modeling of Day-Within-Year Temperature Time Series: An Examination of Daily Maximum Temperatures in Aragón, Spain

Jorge CASTILLO-MATEO, Miguel LAFUENTE, Jesús ASÍN, Ana C. CEBRIÁN, Alan E. GELFAND, and Jesús ABAURREA

Acknowledging a considerable literature on modeling daily temperature data, we propose a multi-level spatiotemporal model which introduces several innovations in order to explain the daily maximum temperature in the summer period over 60 years in a region containing Aragón, Spain. The model operates over continuous space but adopts two discrete temporal scales, year and day within year. It captures temporal dependence through autoregression on days within year and also on years. Spatial dependence is captured through spatial process modeling of intercepts, slope coefficients, variances, and autocorrelations. The model is expressed in a form which separates fixed effects from random effects and also separates space, years, and days for each type of effect. Motivated by exploratory data analysis, fixed effects to capture the influence of elevation, seasonality, and a linear trend are employed. Pure errors are introduced for years, for locations within years, and for locations at days within years. The performance of the model is checked using a leave-one-out cross-validation. Applications of the model are presented including prediction of the daily temperature series at unobserved or partially observed sites and inference to investigate climate change comparison.

Supplementary materials accompanying this paper appear online.

Key Words: Autoregression; Gaussian process; Hierarchical model; Long-term trend; Markov chain Monte Carlo; Spatially varying coefficients.

1. INTRODUCTION

Evidence of global warming in the climate system is strong and many of the observed changes since the 1950s are unprecedented, with an estimated anthropogenic increase of 0.2°C per decade due to past and ongoing emissions (IPCC 2013, 2018). Climate change

raises significant concerns as it may result in health problems and death, degradation of flora and fauna biodiversity, reductions in crop production, increase in pests, etc. In this framework, the analysis of daily maximum temperatures and their long-term trends over time is particularly important due to the strong potential impact on public health (Roldán et al. 2016; Rossati 2017; Watts et al. 2015), agriculture (Hatfield et al. 2011; Schlenker and Roberts 2009), and economy (Diffenbaugh and Burke 2019).

We propose a new multi-level spatiotemporal model to explain the daily maximum temperature in the summer period, in an area containing the Comunidad Autónoma de Aragón in the northeast of Spain. The region includes part of the Ebro Valley in the center, with mountainous areas in the south (Iberian System) and north (Pyrenees). The valley is an extensively irrigated production area with garden crops, fruits, and vegetables, as well as rainfed agriculture with cereals, almonds, wine, and oil. In the mountainous areas, there are some protected natural spaces with extensive forests and a high diversity of landscapes. It is an area of great biodiversity with important water resources for the region. Despite its relatively small size, spatiotemporal modeling of the temperatures in this region is a challenge due to the heterogeneous orography and the climatic variability.

The spatiotemporal model seeks to characterize spatial patterns and detect trends over time in the daily maximum temperature during the summer period. It is specified over continuous space but adopts two discrete units of time, years and days within years. This allows us to model the time evolution of daily maximum temperatures during the summer, omitting the cooler months that are not of interest here. The model introduces temporal dependence using autoregression terms for days within years and also for years. The model separates fixed and random effects in the mean. Fixed effects capture the global mean, the seasonal component across days, the average long-term trend across years, and the influence of elevation. Random effects are employed for the spatial dependence in the intercepts, the slope coefficients, the autoregression coefficients, and the variances of the responses. The two temporal scales allow us to separate space, years, and days within years for each type of effect. Three pure error processes are adopted, one for locations at days within years, one for locations within years, and one for years. The full specification is motivated by exploratory analyses. Altogether, the model provides a better understanding of the temporal evolution of temperatures for the entirety of the region along with the spatial uncertainty linked to those features.

The model is specified in a hierarchical Bayesian framework and estimated using a Markov chain Monte Carlo (MCMC) algorithm. In this framework, posterior predictive distributions for the features of daily maximum temperatures (trends, persistence, mean, variance, etc.) can be readily obtained. In particular, we can obtain posterior predictive samples of the spatial processes and the daily maximum temperature series at unobserved sites. Prediction at unobserved sites is particularly important in Aragón since this region is sparsely monitored due to rural depopulation; there is a lack of observed series in many areas of interest. The model can also be used to impute periods of missing observations in a series.

Space–time modeling of environmental series has received substantial attention in the literature. Sahu et al. (2006) proposed a random effects model for fine particulate matter concentrations in the midwestern USA. Sahu et al. (2007) proposed a space–time hierarchi-

cal model for daily 8-hour maximum ozone levels in the state of Ohio. This model includes an autoregressive part for the residuals of the fixed effects, a global annual intercept, and a spatially correlated error term. Lemos et al. (2007) modeled monthly water temperature data in a Central California Estuary. They used a Bayesian approach to separate the seasonal cycle, short-term fluctuations, and long-term trends by means of local mixtures of two patterns. With regard to temperature models, Craigmile and Guttorp (2011) built space–time hierarchical Bayesian models using daily mean temperatures in Central Sweden that emphasize modeling trend through a wavelet specification, as well as seasonality, and error that may exhibit space–time long-range dependence. Verdin et al. (2015) modeled maximum and minimum temperature to develop a weather generator using spatial Gaussian processes (GPs), where both temperature models are autoregressive with spatially varying model coefficients and spatial correlation. Li et al. (2020) proposed a three-step space–time regression-kriging model for monthly average temperature data. With such data, they first remove seasonality, then they regress the revised data on environmental predictors, and finally they take the resulting residuals and administer spatiotemporal variogram modeling. By contrast, models for daily temperatures take a different approach, seeking to explicitly express short-term persistence of temperature. They employ autoregressive terms, e.g., the one-point model by Mohammadi et al. (2021). A modeling approach very different from our mean specification considers extremes in the daily temperature series and leads to extreme value modeling under the block maxima framework or peaks-over-threshold framework (see, e.g., Reich et al. 2014; Bopp and Shaby 2017).

The outline of the paper is as follows. An exploratory analysis to motivate the complexity of the model is given in Sect. 2. Section 3 describes the modeling details, and Sect. 4 presents a leave-one-out cross-validation (LOOCV) analysis for model comparison as well as some results and applications for the selected model. Section 5 ends the paper with some conclusions and future work. Supplementary Materials accompanying this paper appear online.

2. DATA AND EXPLORATORY ANALYSIS

The point-referenced dataset we use contains 18 daily maximum temperature observational series from AEMET (the Spanish Meteorological Office) around the Comunidad Autónoma de Aragón (see Fig. 1). The time series include the daily observations from May to September (MJJAS), corresponding to the extended summer period, and span the period from 1956 to 2015. The region of interest is located in the central portion of the Ebro Basin in the northeastern part of Spain and has an area of 53,279 km^2, wherein the areas above 500 m and 1,000 m are 32,924 km^2 and 15,195 km^2, respectively. The maximum elevation is roughly 3,400 m in the Pyrenees, 2,600 m in the Iberian System, and between 200 and 400 m in the Central Valley. Most of the area is characterized by a Mediterranean-Continental dry climate with irregular rainfall and a large temperature range. However, climate differences can be distinguished by elevation and the influence from the Mediterranean Sea in the east as well as the continental conditions of the Iberian Central Plateau in the southwest (AEMET 2011).
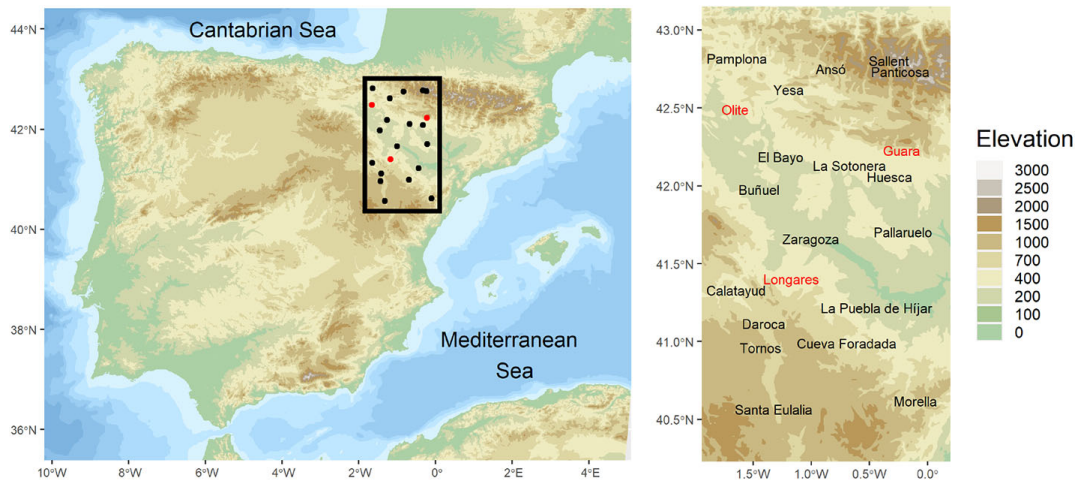
Figure 1.   Map locating within the Iberian Peninsula the 18 sites (black) used to fit the model and the 3 unobserved sites (red) where prediction is carried out (Color figure online).

We summarize an extensive exploratory data analysis of the daily maximum temperature series that helps us establish the covariates and spatiotemporal structures that are candidates for inclusion in the model. The top plots in Fig. 2 show the variability in temperature characteristics and the influence of elevation on them. The two plots on the left show the mean and the standard deviation of temperature at each site against elevation. The mean temperature shows an approximately linear decreasing relation with elevation, varying from almost 30 to 18°C. However, there exist other influential factors, e.g., Sallent in the north and Tornos in the south have both an elevation around 1,000 m, but a quite lower mean temperature is observed for the latter (see Table S1 in Supplementary Materials).

The bottom plots in Fig. 2 summarize the mean and standard deviation from data corresponding to a month in MJJAS for the 18 sites in the periods 1956–1985 and 1986–2015; the summary measures are calculated in 30-year periods following the recommendation of the WMO (2017). The seasonal pattern for all of the series is quite similar, i.e., the maximum mean temperature is observed in July and the minimum in May, with a difference of around 7°C between them. The range of the mean temperatures among sites is around 10°C, so the spatial variability of the mean is a bit higher than the variability at each site within the summer. The mean of the set of standard deviations is slightly higher than 4°C. However, relevant spatial differences are observed with a range of values around 1.5°C. Temporal variability is lower within the summer.

To explore the effect of global warming in the region, the changes between 1956–1985 and 1986–2015 periods, expressed as differences for the means and quotients for the standard deviations, are also shown on the bottom-right plot in Fig. 2 and Table S1. The mean temperature in 1986–2015 has increased from 1956–1985 by roughly 1°C, with a slightly smaller increase in the northeastern sites. The increase in the mean temperature is observed in May, June, August and, except for three sites, in July. No relevant change in the seasonal pattern is observed. The spatial variability in the two periods is similar. As for the standard deviations, no evidence of temporal change is observed, with all of the quotients between the two periods being approximately one.
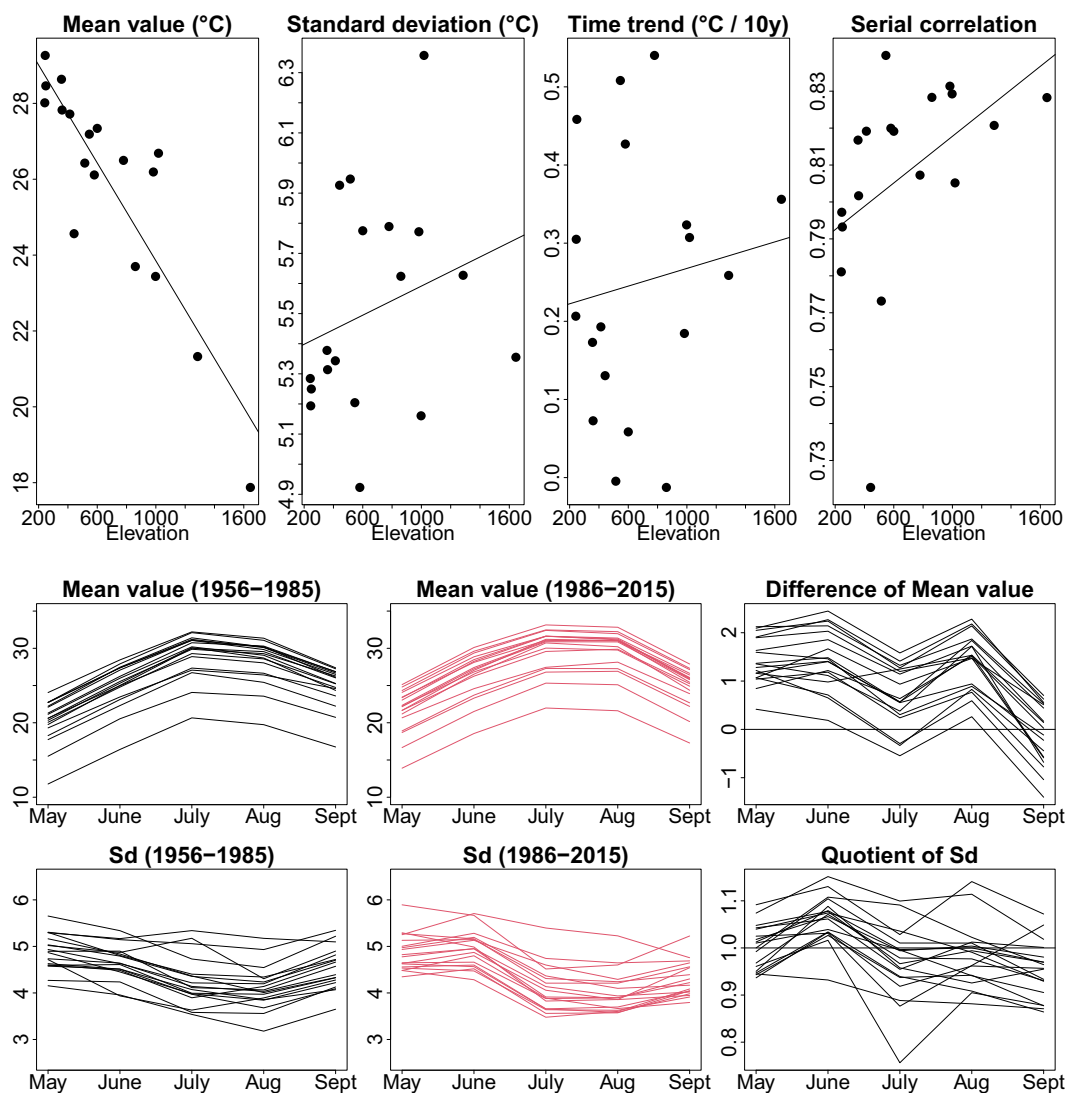
Figure 2. Top: Mean value, standard deviation, annual time trend, and serial correlation against elevation for the daily maximum temperature series at the 18 sites. Bottom: Mean value and standard deviation of the series in both 30-year periods, 1956–1985 and 1986–2015, and the change between them, expressed as differences for the mean and quotients for the standard deviation.

The two plots in the top right in Fig. 2 summarize an exploratory analysis of the behavior of the time series over time. The first shows the slope regressed against year (expressed in °C per decade), fitted by ordinary least squares to the daily maximum temperature series in each site. Clear differences are observed in the 18 fitted trends, suggesting the need to include a spatial random effect to reflect this feature. The variability in the trends does not seem to be related to the elevation. The last plot shows the serial correlation in the temperature series. A strong correlation, higher than 0.72, is observed for all the sites but with spatial differences. The strong autocorrelation is probably caused by a persistent anticyclonic situation that tends to affect the Iberian Peninsula in the summer. Sites with a higher elevation seem to show a slightly higher persistence.

As an additional exploratory analysis, 18 hierarchical temporal models were fitted, one for each of the available sites. These local models, which are summarized in Section S1.1

of the Supplementary Materials, are useful to identify the time structures required for the temperature series and to evaluate the spatial variability of the fitted terms. The results motivate the introduction of spatially varying intercepts, trends, autoregression coefficients, and variances for the spatial variability in the model.

3. THE MODEL

We propose a multi-level (i.e., hierarchical) full mean model for daily maximum temperatures that operates over continuous space and two discrete temporal scales. It captures temporal dependence through autoregression on days within year and on years. It captures spatial dependence through spatial process modeling of intercepts, slope coefficients, variances, and autocorrelations. We detail this model below and then discuss model fitting, prediction under the model, and model comparison.

3.1. MODEL CONSTRUCTION

Let $Y_{t\ell}(\mathbf{s})$ denote the daily maximum temperature for day ℓ, $\ell = 2, \ldots, L$ of year t, $t = 1, \ldots, T$ at location $\mathbf{s} \in D$, where D is our study region. Here, for all years, $\ell = 1$ corresponds to May 1 and $L = 153$ corresponds to September 30. It is convenient to express the full model in a form which separates fixed effects from random effects and also carefully separates space, years, and days for each type of effect. Specifically, we model daily maximum temperature for day ℓ, year t, and location $\mathbf{s}$ by

$$Y_{t\ell}(\mathbf{s}) = \mu_{t\ell}(\mathbf{s}; \boldsymbol{\theta}_f) + \gamma_t(\mathbf{s}) + \rho_Y(\mathbf{s}) \left(Y_{t,\ell-1}(\mathbf{s}) - (\mu_{t,\ell-1}(\mathbf{s}; \boldsymbol{\theta}_f) + \gamma_t(\mathbf{s})) \right) + \epsilon_{t\ell}^{(Y)}(\mathbf{s}). \quad (1)$$

Here, $\mu_{t\ell}(\mathbf{s}; \boldsymbol{\theta}_f)$ denotes the fixed effects component and $\gamma_t(\mathbf{s})$ the random effects component. We specify

$$\mu_{t\ell}(\mathbf{s}; \boldsymbol{\theta}_f) = \beta_0 + \alpha t + \beta_1 \sin(2\pi \ell/365) + \beta_2 \cos(2\pi \ell/365) + \beta_3 \text{elev}(\mathbf{s}) \quad (2)$$

in which β_0 is a global intercept, α is a global linear trend coefficient, the sin and cos terms are introduced to provide an annual seasonal component, and $\text{elev}(\mathbf{s})$ is the elevation at $\mathbf{s}$. We denote these *fixed* effect parameters by $\boldsymbol{\theta}_f = (\beta_0, \alpha, \beta_1, \beta_2, \beta_3)$.

We specify

$$\gamma_t(\mathbf{s}) = \beta_0(\mathbf{s}) + \alpha(\mathbf{s})t + \psi_t + \eta_t(\mathbf{s}). \quad (3)$$

In (3), ψ_t follows an AR(1) specification, i.e., $\psi_t = \rho_\psi \psi_{t-1} + \lambda_t$, providing an autoregression in years for annual intercepts. This autoregression could help to capture factors yielding correlation across years, such as the influence of variation in solar activity on the earth's surface temperature or the El Niño–Southern Oscillation. However, in Sect. 3.2, we discover that ρ_ψ is not significantly different from 0. We still need ψ's in the model to address the fact that some years are warmer or colder than others, but we do not need to specify them autoregressively. We denote the variance for this component by σ_λ^2.

Continuing, $\beta_0(\mathbf{s})$ is a mean-zero GP with an exponential covariance function having variance parameter $\sigma_{\beta_0}^2$ and decay parameter ϕ_{β_0}, and $\alpha(\mathbf{s})$ is a mean-zero GP with an

exponential covariance function having variance parameter σ_α^2 and decay parameter ϕ_α. Thus, $\beta_0(\mathbf{s})$ provides local spatial adjustment to the intercept and $\alpha(\mathbf{s})$ provides local slope adjustment to the linear trend. Due to the simplicity of linear time trends they are often used in climate studies (IPCC 2013). Here, they provide an extremely flexible, locally linear baseline specification. Further, we add local space–time varying random effects, $\eta_t(\mathbf{s})$, to provide adjustment to this baseline. We collect the random effects parameters into $\boldsymbol{\theta}_r = (\rho_\psi, \sigma_\lambda^2, \sigma_{\beta_0}^2, \phi_{\beta_0}, \sigma_\alpha^2, \phi_\alpha)$.

The entire specification is supplied distributionally in the form of a multi-level hierarchical model as

$$[Y_{t\ell}(\mathbf{s}) \mid Y_{t,\ell-1}(\mathbf{s}), \vec{\theta}_f, \gamma_t(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon^2(\mathbf{s})]$$

$$[\gamma_t(\mathbf{s}) \mid \beta_0(\mathbf{s}), \alpha(\mathbf{s}), \psi_t, \sigma_\eta^2]$$

$$[\beta_0(\mathbf{s})|\sigma_{\beta_0}^2, \phi_{\beta_0}][\alpha(\mathbf{s})|\sigma_\alpha^2, \phi_\alpha][\psi_t|\psi_{t-1}, \rho_\psi, \sigma_\lambda^2] \qquad (4)$$

$$[Z_{\rho_Y}(\mathbf{s})|Z_{\rho_Y}, \sigma_{\rho_Y}^2, \phi_{\rho_Y}][Z_{\sigma_\epsilon^2}(\mathbf{s})|Z_{\sigma_\epsilon^2}, \sigma_{\sigma_\epsilon^2}^2, \phi_{\sigma_\epsilon^2}]$$

$$[\vec{\theta}_f][\vec{\theta}_r][\sigma_\eta^2][Z_{\rho_Y}][\sigma_{\rho_Y}^2][\phi_{\rho_Y}][Z_{\sigma_\epsilon^2}][\sigma_{\sigma_\epsilon^2}^2][\phi_{\sigma_\epsilon^2}].$$

As a result, we have introduced three pure error terms: $\lambda_t \overset{iid}{\sim} N(0, \sigma_\lambda^2)$ at yearly scale, $\eta_t(\mathbf{s}) \overset{iid}{\sim} N(0, \sigma_\eta^2)$ at sites within years, and $\epsilon_{t\ell}^{(Y)}(\mathbf{s}) \overset{ind.}{\sim} N(0, \sigma_\epsilon^2(\mathbf{s}))$ at sites for days within years. Additionally, $\rho_Y(\mathbf{s})$ and $\sigma_\epsilon^2(\mathbf{s})$ are, respectively, a spatially varying autoregressive term and a spatially varying variance at location $\mathbf{s}$, both of which are assumed constant over days and years. We model $\log\{(1 + \rho_Y(\mathbf{s}))/(1 - \rho_Y(\mathbf{s}))\} = Z_{\rho_Y}(\mathbf{s}) \sim GP(Z_{\rho_Y}, C(\cdot; \sigma_{\rho_Y}^2, \phi_{\rho_Y}))$, and $\log\{\sigma_\epsilon^2(\mathbf{s})\} = Z_{\sigma_\epsilon^2}(\mathbf{s}) \sim GP(Z_{\sigma_\epsilon^2}, C(\cdot; \sigma_{\sigma_\epsilon^2}^2, \phi_{\sigma_\epsilon^2}))$, again with exponential covariance functions. Motivation for adopting spatially varying specifications for these terms arises from exploratory data analysis at the level of the individual sites. That is, suppose we fit the model above but ignore spatial structure and treat the sites as conditionally independent. We show in Section S1.1 of the Supplementary Materials that the assumptions of constant autoregression coefficients and constant variances over the region do not seem justified.

All of the components considered in the full model and their relationships are depicted in the graphical model in Fig. 3. This diagram, perhaps, reveals the complexity of the full model more readily than through Equations (1) to (4).

The reader might wonder if the GPs above are independent. We investigated dependence between the intercept and slope GPs using the following coregionalization (Banerjee et al. 2014, Chapter9). Suppose $v_1(\mathbf{s})$ and $v_2(\mathbf{s})$ are independent GPs with zero mean and unit variance whose exponential covariance functions have decay parameters ϕ_1 and ϕ_2, respectively. In the full model, we insert $\beta_0(\mathbf{s}) = a_{11}v_1(\mathbf{s})$ and $\alpha(\mathbf{s}) = a_{21}v_1(\mathbf{s}) + a_{22}v_2(\mathbf{s})$. Here, we let a_{11} and a_{22} each have a half (or folded) Gaussian prior, while a_{21} has a regular Gaussian prior. The parameter a_{21} captures the dependence between the two processes. That is, the induced covariance between $\beta_0(\mathbf{s})$ and $\alpha(\mathbf{s})$ is $a_{21}a_{11}$. We care whether a_{21} is significantly different from zero with little interest in exactly what the correlation is. Under the model above, the posterior distribution of a_{21} was centered at zero with wide credible intervals. So, this dependence was not included in the final model for which we present the inference.
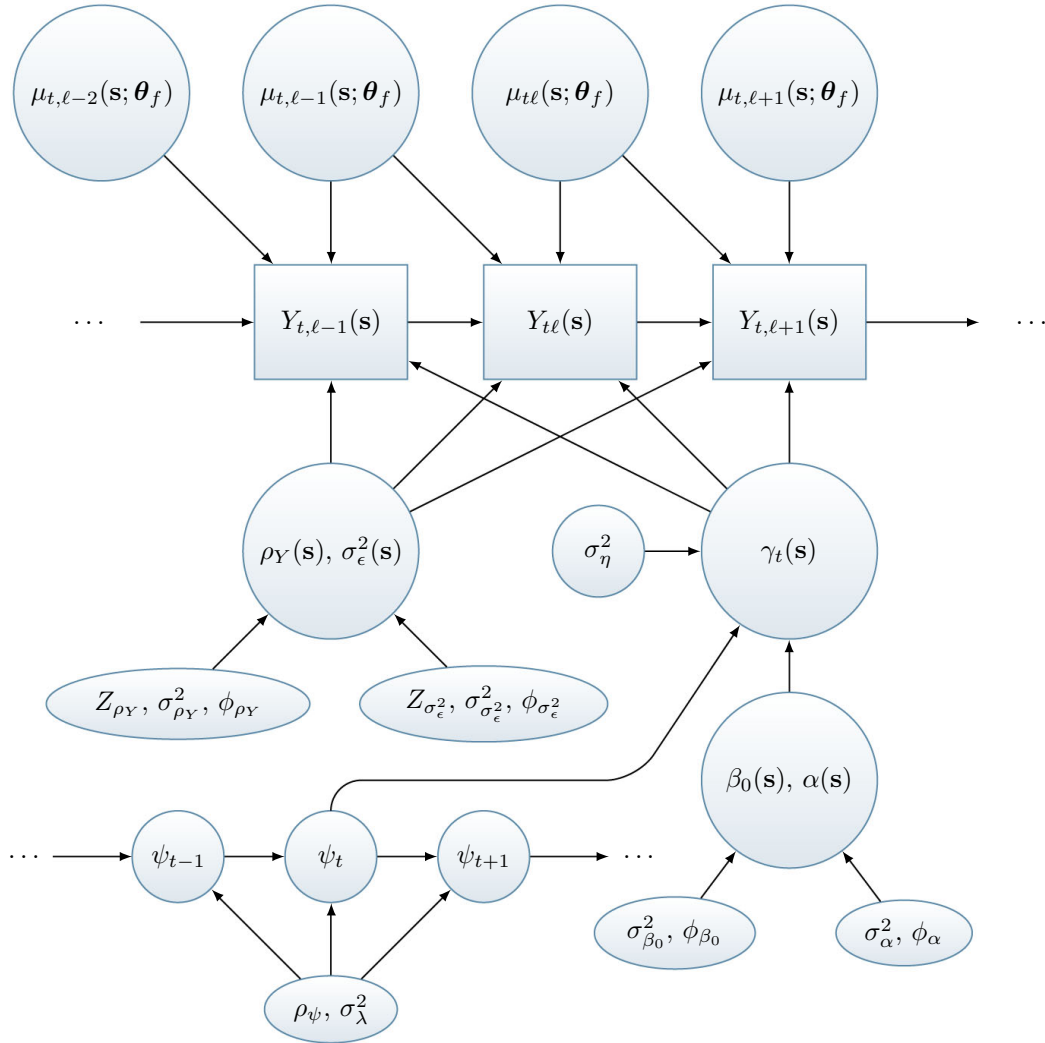
Figure 3.   Graphical model for specification in Equations (1) to (4). Rectangular nodes are observed, circular nodes are unobserved.

Returning to the full model, notice that we have separated the fixed effects according to subscripts t, ℓ, and $\mathbf{s}$. As for $\gamma_t(\mathbf{s})$, we can see that it has a spatially varying intercept, a spatially varying coefficient for drift, and an AR(1) model for years. Also, $\gamma_t(\mathbf{s})$ has both space and time dependence and, in fact, we can readily calculate $\mathrm{cov}(\gamma_t(\mathbf{s}), \gamma_{t+h}(\mathbf{s}'))$. Under independence of the intercept and slope processes, the *equilibrium* covariance becomes

$$\mathrm{cov}(\gamma_t(\mathbf{s}), \gamma_{t+h}(\mathbf{s}')) = C(||\mathbf{s} - \mathbf{s}'||; \sigma_{\beta_0}^2, \phi_{\beta_0}) + t(t+h)C(||\mathbf{s} - \mathbf{s}'||; \sigma_\alpha^2, \phi_\alpha) + \frac{\sigma_\lambda^2}{1 - \rho_\psi^2}\rho_\psi^{|h|}.$$

(5)

Finally, special cases of interest include: $\beta_0(\mathbf{s}) = 0$ implies a constant intercept over space, $\alpha(\mathbf{s}) = 0$ implies a constant linear drift over space, and $\rho_\psi = 0$ implies no yearly autoregression. These assumptions merely revise the form of $\gamma_t(\mathbf{s})$. We might consider conditioning on a longer history of maximum temperatures. We experimented with introducing additional lags in the modeling, but we found no gain in predictive performance. We could

also consider additional fixed effects, e.g., longitude, latitude or distance to coast, or even adding interactions, e.g., $t \times \mathrm{elev}(\mathbf{s})$. However, the exploratory analysis did not reveal a relationship between daily temperatures and these fixed effects, so they were not introduced in the full model.

3.2. MODEL FITTING

Model inference is implemented in a Bayesian framework, requiring prior distributions for each of the model parameters. In general, diffuse and, when available, conjugate prior distributions are chosen. Recall that the model adopts a conditional Gaussian distribution for all $Y_{t\ell}(\mathbf{s})$'s. Thus, it is appropriate to assign each of the coefficient parameters β_0, α, β_1, β_2, and β_3, independent and diffuse Gaussian prior distributions with mean 0 and standard deviation 100. The variance parameters, σ_λ^2 and σ_η^2, are assigned independent Inverse-Gamma(2, 1) prior distributions. In preliminary analyses, the autoregresive term between years, ρ_ψ, was assigned a non-informative Uniform$(-1, 1)$ prior distribution. As its posterior distribution was centered at zero with wide credible intervals, we set the parameter at $\rho_\psi = 0$. For identifiability, the random effect for the first year, ψ_1, is fixed to zero.

Hyperpriors are assigned to the mean of both $Z_{\rho_Y}(\mathbf{s})$ and $Z_{\sigma_\epsilon^2}(\mathbf{s})$. That is, Z_{ρ_Y} and $Z_{\sigma_\epsilon^2}$ are given a Gaussian prior distribution with mean 0 and standard deviation 100 and 1, respectively. The variance parameter for each of the four spatial covariance functions, $\sigma_{\beta_0}^2$, σ_α^2, $\sigma_{\rho_Y}^2$, and $\sigma_{\sigma^2}^2$, is assigned an independent Inverse-Gamma(2, 1) prior distribution. Preliminary analyses with a discrete uniform prior distribution for each of the spatial decay parameters indicated that these parameters almost always placed most mass on the smallest decay value. Due to the fact that, with an exponential covariance function, the variance and the decay parameter cannot be individually identified (Zhang 2004), and the decay parameter is 3/range, we set $\phi \equiv \phi_{\beta_0} = \phi_\alpha = \phi_{\rho_Y} = \phi_{\sigma^2} = 3/d_{max}$, where d_{max} is the maximum distance between any pair of spatial locations.

MCMC is used to obtain samples from the joint posterior distribution. The sampling algorithm is a Metropolis-within-Gibbs version. Since we only have 18 sites, we fit the model without marginalization over the spatial random effects. Also, we introduce $\tilde{\beta}_0(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$ and $\tilde{\alpha}(\mathbf{s}) = \alpha + \alpha(\mathbf{s})$ within $\gamma_t(\mathbf{s})$ for the fitting to enable the benefits of hierarchical centering in the model fitting (Gelfand et al. 1995). Details of the MCMC used for the model fitting are provided in Section S2.1 of the Supplementary Materials. All the covariates have been centered and scaled to have zero mean and standard deviation one to improve the mixing behavior of the algorithm.

3.3. SPATIAL AND SPATIOTEMPORAL PREDICTION

Under the full model, prediction at location $\mathbf{s}_0$, day ℓ', and year t' is based on the posterior predictive distribution of $Y_{t'\ell'}(\mathbf{s}_0)$ arising from the full model. Here, $\mathbf{s}_0$ may correspond to a fully observed location (held out for validation), a partially observed location (for completion of a record), or a new location in D. Our goal is not forecasting, so we restrict ourselves to the observed time period $\ell' = 2, \ldots, L$ and $t' = 1, \ldots, T$. Within the Bayesian framework, the posterior predictive distribution for $Y_{t'\ell'}(\mathbf{s}_0)$ is obtained by integrating over

the parameters with respect to the joint posterior distribution. The formal expression for the posterior predictive distribution for $[Y_{t'\ell'}(\mathbf{s}_0) \mid \mathbf{Y}]$, where $\mathbf{Y}$ is the observed data, is given in Section S2.2 of the Supplementary Materials. Customarily, the distribution is obtained empirically through posterior samples. That is, with MCMC algorithms, samples of the posterior parameters are used to obtain posterior predictions of observations, the so-called composition sampling (see Banerjee et al. 2014, Chapter 6; and Section S2.2 for the details).

3.4. MODEL EVALUATION

For model assessment, a LOOCV is carried out to compare the spatial predictive performance of the models. The full model considered includes four spatial GPs. To validate that model as well as the importance of the considered GPs, reduced models incorporating 0, 1, 2, or 3 GPs are fitted. Models are presented explicitly in Sect. 4.1 where we further clarify that removing particular terms allows explicit interpretation of the resulting reduced models.

Results from Sect. 4.1 favor the full model, and so results for this model are presented subsequently. However, several of the reduced models yield essentially equivalent global performance, though the fit at some sites is poorer. We attempt to clarify why this might be expected but also show that each set of random effects reveals differences across sites, further encouraging us to retain them in the inference presentation.

For each location in the holdout set, the entire time series of daily maximum temperatures is withheld during model fitting. Then, for location $\mathbf{s}_i$, we conduct our model comparison through the following metrics: (i) root-mean-square error (RMSE), (ii) mean absolute error (MAE), (iii) continuous ranked probability score (CRPS; Gneiting and Raftery 2007), and (iv) coverage (CVG). By definition,

$$\text{RMSE}_i = \sqrt{\frac{1}{T(L-1)} \sum_{t=1}^{T} \sum_{\ell=2}^{L} \left(\hat{Y}_{t\ell}(\mathbf{s}_i) - Y_{t\ell}(\mathbf{s}_i) \right)^2},$$

$$\text{MAE}_i = \frac{1}{T(L-1)} \sum_{t=1}^{T} \sum_{\ell=2}^{L} \left| \hat{Y}_{t\ell}(\mathbf{s}_i) - Y_{t\ell}(\mathbf{s}_i) \right|,$$

$$\text{CRPS}_i = \frac{1}{T(L-1)} \sum_{t=1}^{T} \sum_{\ell=2}^{L}$$
$$\left(\frac{1}{B} \sum_{b=1}^{B} \left| Y_{t\ell}^{(b)}(\mathbf{s}_i) - Y_{t\ell}(\mathbf{s}_i) \right| - \frac{1}{2B^2} \sum_{b_1=1}^{B} \sum_{b_2=1}^{B} \left| Y_{t\ell}^{(b_1)}(\mathbf{s}_i) - Y_{t\ell}^{(b_2)}(\mathbf{s}_i) \right| \right),$$

$$\text{CVG}_i = \frac{1}{T(L-1)} \sum_{t=1}^{T} \sum_{\ell=2}^{L} I(L_{t\ell}(\mathbf{s}_i) \leq Y_{t\ell}(\mathbf{s}_i) \leq U_{t\ell}(\mathbf{s}_i)),$$

where $\hat{Y}_{t\ell}(\mathbf{s}_i) = \sum_{b=1}^{B} Y_{t\ell}^{(b)}(\mathbf{s}_i)/B$ with $Y_{t\ell}^{(b)}(\mathbf{s}_i)$ the bth posterior predictive replicate of $Y_{t\ell}(\mathbf{s}_i)$, from the left-out location $\mathbf{s}_i$. Also, $(L_{t\ell}(\mathbf{s}_i), U_{t\ell}(\mathbf{s}_i))$ is the 90% predictive interval for $Y_{t\ell}(\mathbf{s}_i)$, i.e., the 5th and 95th percentiles of the MCMC samples $Y_{t\ell}^{(b)}(\mathbf{s}_i)$ $(b = 1, \ldots, B)$,

Table 1.   Mean value across the 18 sites of the performance metrics for models with different spatial GPs

	RMSE	MAE	CRPS	CVG
M_0	4.49	3.64	2.57	0.894
$M_1(\beta_0(\mathbf{s}))$	4.36	3.53	2.49	0.901
$M_1(\alpha(\mathbf{s}))$	4.49	3.64	2.57	0.894
$M_1(\rho_Y(\mathbf{s}))$	4.49	3.64	2.57	0.895
$M_1(\sigma_\epsilon(\mathbf{s}))$	4.49	3.63	2.56	0.893
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.36	3.53	2.48	0.901
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.36	3.53	2.49	0.899
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.49	3.63	2.56	0.894
M_4	4.36	3.53	2.48	0.900

and $I(\cdot)$ is the indicator function. The smaller the RMSE, MAE, and CRPS values, the better the model performance. However, the target for CVG is proximity to 0.90.

4. RESULTS

We summarize, using LOOCV, the comparison of models with differing inclusion of the foregoing spatial GPs. Each model was fitted to the daily maximum temperature series in months MJJAS for the 60 years from 1956 to 2015. Then, we present the results for the fitting of the full model over the study region.

In the MCMC fitting, we ran 10 chains, with 200,000 iterations for each chain, to obtain samples from the joint posterior distribution. The first 100,000 samples were discarded as burn-in, and the remaining 100,000 samples were thinned to retain 100 samples from each chain for posterior inference. MCMC diagnostics for the full model are shown in Section S2.3 of the Supplementary Materials.

4.1. Validation and Model Comparison

The full model considered includes four spatial GPs. To compare models and assess the importance of the proposed GPs, simpler models incorporating 0, 1, 2, or 3 GPs are fitted. M_p with $p = 0, 1, \ldots, 4$ denotes a model including p spatial processes that are specified in parentheses. For example, $M_1(\beta_0(\mathbf{s}))$ is the model with a single spatial process for the intercept; for simplicity, the full model is denoted M_4.

Using the criteria in Sect. 3.4 with LOOCV for each of the 18 available locations, Table 1 summarizes the averages across sites for the four metrics. The strongest improvement in predictive performance is obtained by adding a spatially varying intercept process, i.e., $M_1(\beta_0(\mathbf{s}))$. The inclusion of the other GPs does not yield a clear improvement in performance. This is not surprising, since the GP for intercepts explicitly rewards predicting the mean and random realizations well in order to agree with the held-out values. However, the usefulness of the other GPs with regard to effectively capturing autocorrelations and variances at the observed sites will be seen in Sect. 4.2.

Table 2.   Posterior mean and 90% credible intervals for the parameters of M_4

	Mean	Credible interval
β_0 (intercept)	25.70	(24.30, 27.16)
α (trend)	0.0207	$(-0.0074, 0.0490)$
β_1 (sine)	13.18	(13.00, 13.37)
β_2 (cosine)	0.633	(0.558, 0.709)
β_3 (elevation)	-0.0069	$(-0.0084, -0.0054)$
ρ_Y	0.691	(0.606, 0.762)
σ_ϵ	2.963	(2.433, 3.515)
σ_η	0.230	(0.201, 0.264)
σ_λ	0.936	(0.799, 1.088)
σ_{β_0}	1.492	(1.154, 1.939)
σ_α	0.0283	(0.0211, 0.0376)
σ_{ρ_Y}	0.339	(0.263, 0.435)
$\sigma_{\sigma_\epsilon^2}$	0.404	(0.312, 0.522)

Table S4 in the Supplementary Materials provides details, by site, for the metrics in Table 1. The locations with poorest fit for all of the models are Pamplona and Tornos, the only ones with CRPS greater than 3. They also show large RMSE and MAE as well as poor CVG. For the other locations, the CVG of all the models is closer to the nominal value 0.90. In particular, M_4 not only has the best CVG on average, but the variability of the CVG$_i$'s with respect to the nominal 0.90 is the lowest of all the models.

4.2. RESULTS FOR THE FULL MODEL

Here, we show fitted and prediction results for the full model, M_4, and demonstrate the need to include the four GPs. The parameters α, β_1, β_2, β_3, $\alpha(\mathbf{s})$, and σ_α have been rescaled to interpret them in terms of the original measure of the covariates. Table 2 summarizes the posterior mean and credible intervals of the model parameters, including standard deviation of random effects.

The harmonic coefficients β_1 and β_2 indicate the strong seasonality in the temperature series. The coefficient β_3 supplies the gradient of temperature corresponding to elevation, approximately -7°C per 1,000 m. This value agrees with the exploratory analysis in Sect. 2, and the average environmental lapse rate (Navarro-Serrano et al. 2018). The linear trend coefficient, α, indicates that the average increase in temperature is 0.21°C per decade. Peña-Angulo et al. (2021) found a similar trend (0.27°C per decade) in the summer maximum temperature in Spain (1956–2015). The posterior mean of the autoregresive spatial process, ρ_Y, confirms the strong serial correlation of daily temperatures.

The other parameters are standard deviations linked to the spatiotemporal effects of the model. The posterior mean of σ_ϵ, the mean of the spatially varying standard deviations of the pure error process $\epsilon_{t\ell}^{(Y)}(\mathbf{s})$, is close to 3°C. This value doubles the posterior mean of σ_{β_0} which represents the spatial variability of the mean level $\beta_0(\mathbf{s})$ and triples the posterior mean of σ_λ, linked to the variability of the yearly random effects ψ_t. The magnitude of the remaining standard deviation parameters is smaller.
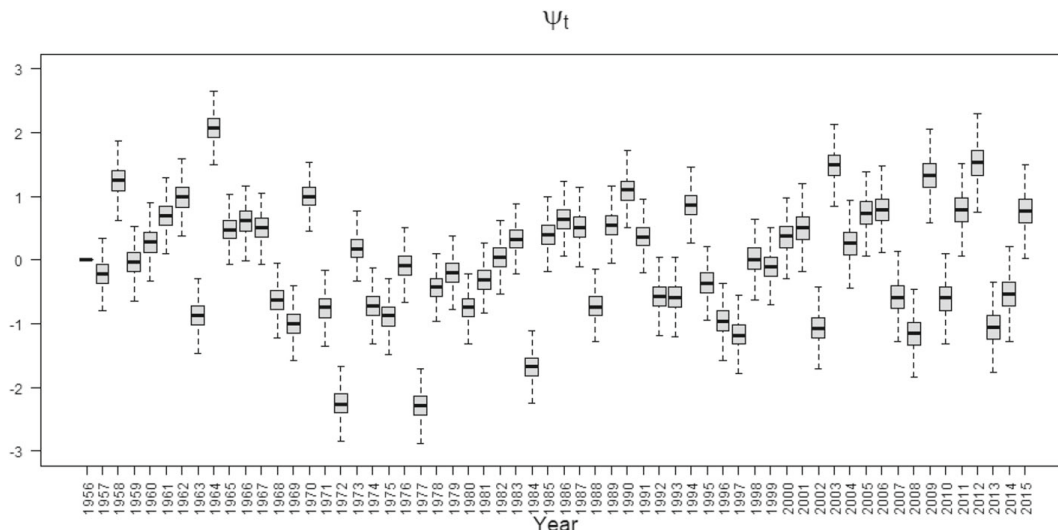
ψ_t

Figure 4.   Box plots of the posterior distributions of the annual random effects ψ_t in M_4.

With $\rho_\psi = 0$, the yearly random effects, ψ_t, are, a priori, distributed as $N(0, \sigma_\lambda^2)$. The posteriors are summarized using box plots in Fig. 4. It is observed that the effects may add or subtract in a given year up to roughly 2.5°C, with a standard deviation close to 1°C. These yearly random effects are able to capture historical events like the extremely cold summer of 1977 in Spain or the European heat wave in 2003 (Peña-Angulo et al. 2021).

The posterior distributions at the observed locations of the four spatial processes in M_4, $\tilde{\beta}_0(\mathbf{s}), \tilde{\alpha}(\mathbf{s}), \rho_Y(\mathbf{s})$, and $\sigma_\epsilon(\mathbf{s})$, are summarized in Fig. 5 using box plots. The box plots of the locations are sorted from the lowest to the highest elevation in the horizontal axis. They confirm the need to consider the four GPs to represent the great climatic variability of the region under study. To show the spatial behavior of the spatial processes over the entire region, maps of their posterior means, obtained by a model-based Bayesian kriging, are presented in Fig. 6. In Section S3.2 of the Supplementary Materials, the parameters of M_4 are compared with the parameters of the local models described in Section S1.1, and both show good agreement.

The top-left plots in Figs. 5 and 6 correspond to $\tilde{\beta}_0(\mathbf{s})$. The posterior distributions for most of the locations show remarkable differences. In particular, $\tilde{\beta}_0(\mathbf{s})$ has a clear climatic interpretation. The spatial adjustments provided by this GP help to improve the fit for the two areas with a similar elevation around 1,000 m but different climates. These areas are the southwest and the north of the region. The former has a warmer climate than the latter, whose climate is influenced by the proximity of the Atlantic Ocean.

With regard to the spatially varying yearly linear trend, $\tilde{\alpha}(\mathbf{s})$, the top-right plots in Figs. 5 and 6 reveal clear spatial differences in the warming trend. The posterior distributions for higher locations and for the Central Valley are shifted with respect to others. Most of the area shows warming trends, except some areas in the northwest, e.g., Yesa or Ansó, whose posterior distributions are centered at zero.

The spatial process for the autoregressive term, $\rho_Y(\mathbf{s})$, is clearly necessary in the model. The bottom-left plot in Fig. 5 shows that the posterior distributions for the 18 locations differ substantially. The posterior means of the $\rho_Y(\mathbf{s})$ are positive in all locations, and their
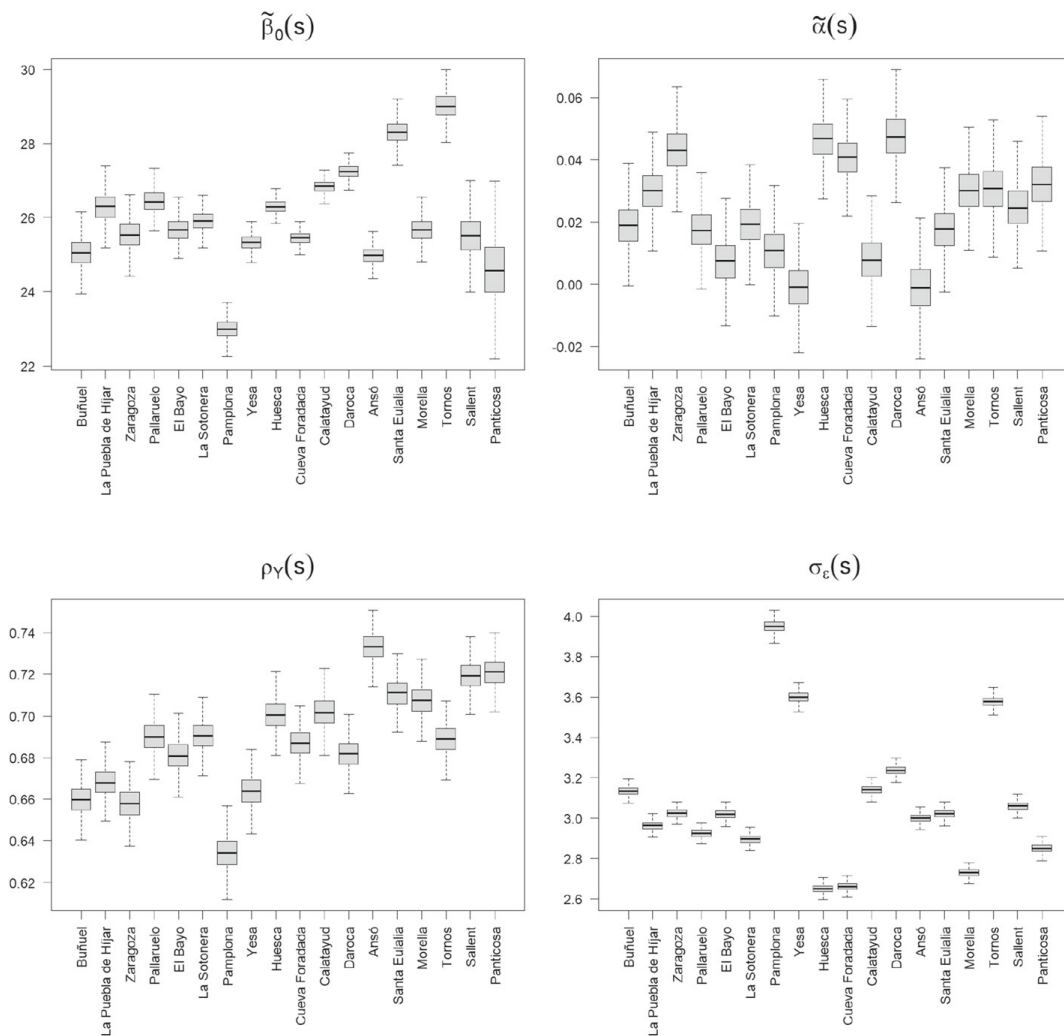
Figure 5.   Box plots of the posterior distributions of the spatial random effects, $\tilde{\beta}_0(\mathbf{s})$, $\tilde{\alpha}(\mathbf{s})$, $\rho_Y(\mathbf{s})$, $\sigma_\epsilon(\mathbf{s})$, in M_4. Locations are sorted by elevation, from lowest to highest.

values seem to have an increasing relation with the elevation. According to the bottom-left plot in Fig. 6, the posterior mean is also related to *cierzo*, a severe northwesterly cold wind that gives rise to a renewal of the atmospheric condition with less warm air masses. This wind reduces the persistence of the temperature and therefore the dependence with respect to the previous day. In the areas affected by *cierzo*, the mean is around 0.65, lower than the posterior mean of the mean of the process $\rho_Y(\mathbf{s})$, close to 0.7.

The need for the $\sigma_\epsilon(\mathbf{s})$ process is also clear. The bottom-right plot in Fig. 5 reveals strong differences among the posterior distributions of the standard deviations across locations. The high variability of Pamplona, Yesa, and Tornos stands out. The bottom-right plot in Fig. 6 confirms the spatial variability of the standard deviation and shows that higher standard deviations are observed in the western part of the region.

4.2.1. Prediction at Unobserved Locations

Now, we illustrate the use of the full model for prediction at three unobserved sites in the region: Longares (530 m), Olite (390 m), and Guara (800 m). The new sites are marked in
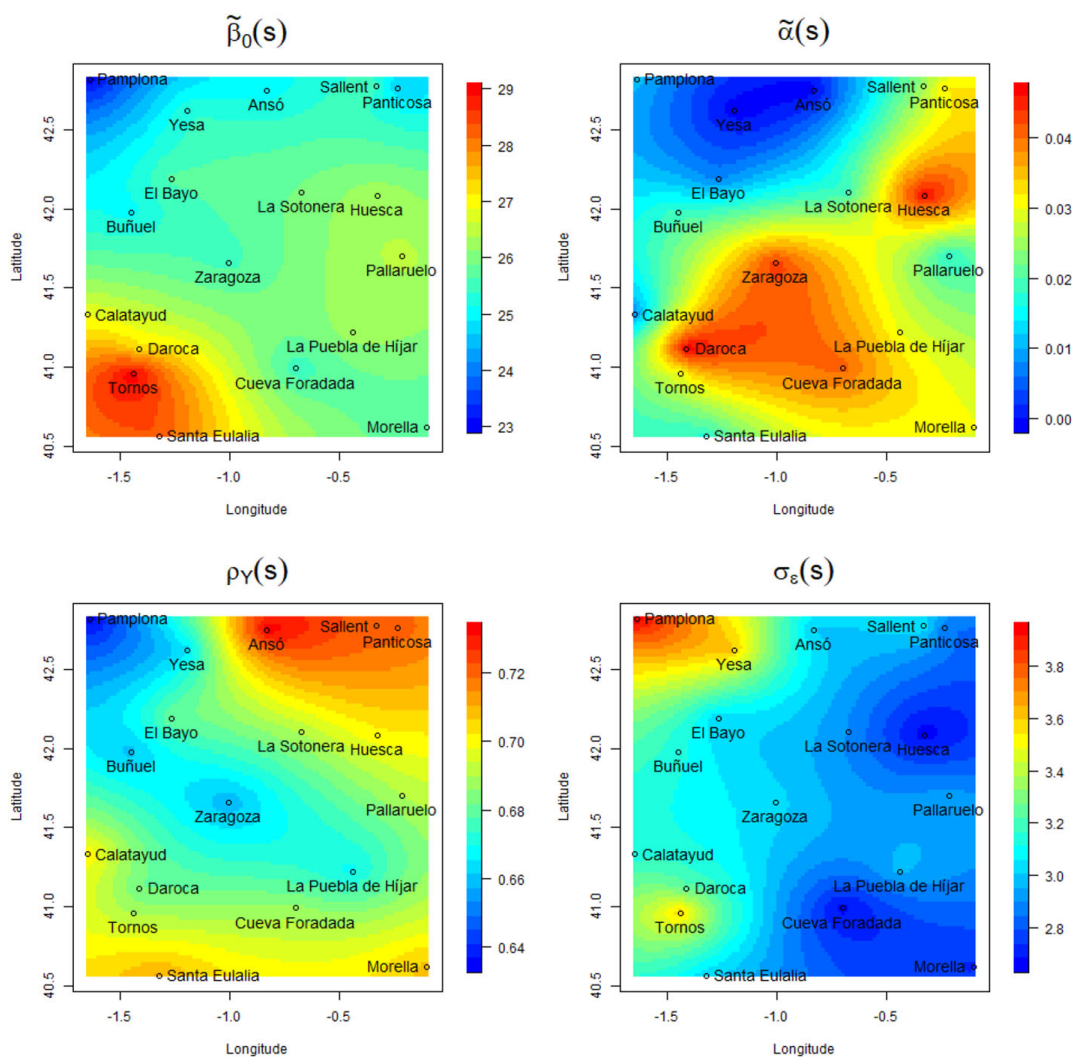
Figure 6.   Maps of the posterior means of the four spatial processes included in M_4, obtained by a model-based Bayesian kriging, with resolution 100×100 (Color figure online).

red in Fig. 1 and represent areas with different environmental and climatic characteristics. Longares is located in the southern half of the region in a rainfed agricultural area dedicated to the production of wine. Vines are seriously affected by global warming since high temperatures lead to both a decrease in production and a premature ripening of the grapes. Olite is located in a rural area in the northwest where smaller increases in the temperature have been observed; an incomplete series of observed values is available at this site. Guara is an uninhabited area in the Natural Park Sierra and Cañones de Guara. The prediction of the temperature evolution in this area is essential to better understand the changes that have been observed in the ecosystem of the Natural Park.

We use the model to impute missing values in an observed series using the posterior predictive distribution. Daily temperatures in Olite are available in the AEMET database from 1968 to 2007, although with many missing observations. As an example, Fig. 7 shows the plot of the observed series and the posterior predictive means with 90% credible intervals for MJJAS days in 1968 and, as a summary, the plot of the observed and the posterior yearly averages with 90% credible intervals. The 90% CVG in the observed data is 92.0%. The
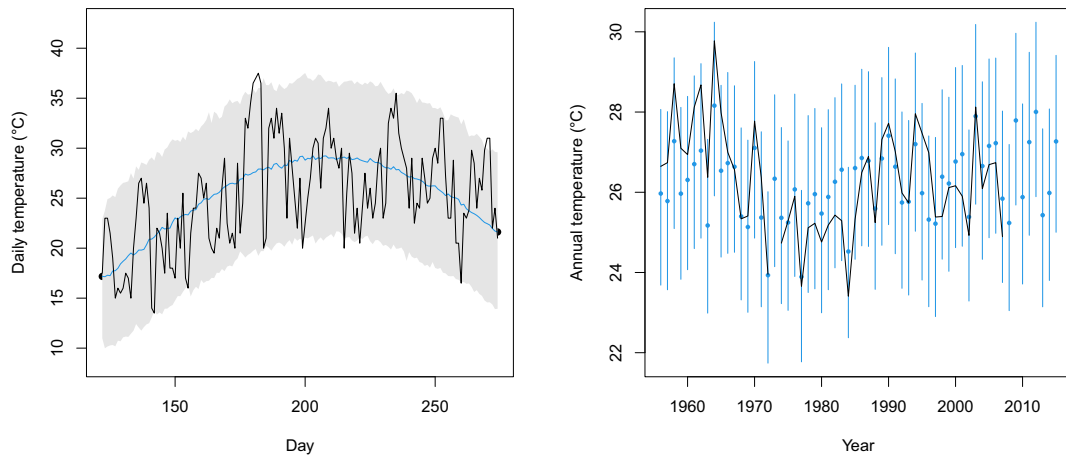
Figure 7.   Left: Observed (black rough curve) and posterior predictive means (blue smooth curve) with associated 90% credible intervals of daily maximum temperatures in Olite (1968). Right: Observed yearly averages (black curve) and associated posterior mean and 90% credible intervals (Color figure online).

agreement between the observed and the predicted data confirms that M_4 can be used effectively to impute missing values in Olite.

The posterior distribution of the four spatial processes $\tilde{\beta}_0(\mathbf{s})$, $\tilde{\alpha}(\mathbf{s})$, $\rho_Y(\mathbf{s})$, and $\sigma_\epsilon(\mathbf{s})$ for the three predicted locations are shown in Figure S5 of the Supplementary Materials. The posterior distributions for $\tilde{\beta}_0(\mathbf{s})$ in Longares and Guara are similar despite having different elevations. The posterior distributions of $\tilde{\alpha}(\mathbf{s})$ in Longares and Guara are very similar, while the distribution of Olite is shifted with a posterior mean almost 0.3°C per decade lower. The fitted $\rho_Y(\mathbf{s})$'s show the differences in the autocorrelation of temperature in the three locations with posterior means varying from 0.65 to 0.72. The largest differences in the posterior distributions appear in the $\sigma_\epsilon(\mathbf{s})$.

M_4 is also used to evaluate the change over time of the temperature in the three predicted sites, using the posterior predictive distribution of the difference between the average in the 30-year periods 1956–1985 and 1986–2015 (see Figure S6 in Supplementary Materials). Despite the difference in elevation, the posterior mean of the increment is similar in Longares and Guara, around 1.4°C, while in Olite it is smaller, 0.5°C and its 90% credible interval $(-0.010, 1.028)$ contains zero. The posterior probability that the mean in 1986–2015 is higher than in 1956–1985 is 0.94 in Olite and essentially 1 in Longares and Guara.

5. SUMMARY AND FUTURE WORK

We have proposed a very rich space–time mean model for daily maximum temperatures, fitted over a 60-year period for a region in Spain. Our specification is continuous in space and autoregressive in time. In time, autoregression was examined annually and also daily for the summer season within each year. We find novel spatial structure including spatially varying intercepts and trend coefficients as well as spatially varying autoregression coefficients and variances.

The proposed modeling can be adapted to other regions, perhaps considering other geographical covariates such as latitude, longitude, or distance to the sea. Also, the modeling

can omit spatial processes that are not necessary, e.g., avoiding $\rho_Y(\mathbf{s})$ in a more homogeneous region with a lower variation in elevation. The modeling might also be adapted to other response variables in spatiotemporal problems, such as daily minimum temperature and other environmental variables including daily evapotranspiration or hourly temperature in the sea. The flexible autoregression terms can express behavior in series where serial correlation is an important source of variation.

A limitation of the present analysis is that we have only 18 monitoring stations so that learning about the spatial surfaces in our modeling is less than we would want. Despite this small number of sites, the model has been able to capture the climate variability of the region under study. The spatial random effects identify areas with a different mean temperature level, but also areas where the observed warming over time shows a different trend, areas where temperature is more persistent (i.e., with a stronger daily serial correlation) or with different variability. The capacity of the fitted model to impute temperature over the entire region allows us to obtain reliable predictions and credible intervals for daily temperature series at unobserved sites. This can be valuable for economical, agricultural, or environmental reasons.

Future work will consider different regions providing more available spatial locations n. However, the $\mathcal{O}(n^3)$ computational complexity of inverting a $n \times n$ covariance matrix can be prohibitive for implementing the above model for data with large n. Reduced rank approximations to GPs may be used to address this computation bottleneck, e.g., Gaussian predictive process (Banerjee et al. 2008) or nearest-neighbor GP (Datta et al. 2016). As a different challenge, one may wonder whether the low trend values (blue region) in the top-right plot in Fig. 6 are actually meaningful. Future work could implement a version of a spatially dependent multiple testing analysis (Risser et al. 2019) given the posterior draws of $\tilde{\alpha}(\mathbf{s})$. A different future direction will move away from mean modeling to quantile modeling in order to investigate extremes of temperature, both hot and cold. This will lead to novel development for spatiotemporal quantile regression.

ACKNOWLEDGEMENTS

REFERENCES

AEMET (2011) Atlas climático ibérico – Iberian climate atlas. Ministerio de Medio Ambiente, y Medio Rural y Marino; Agencia Estatal de Meteorología; and Instituto de Meteorologia de Portugal, https://doi.org/10.31978/784-11-002-5

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc: Ser B (Stat Methodol) 70(4):825–848. https://doi.org/10.1111/j.1467-9868.2008.00663.x

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical Modeling and Analysis for Spatial Data, 2nd edn. Chapman and Hall/CRC, New York, NY, USA, https://doi.org/10.1201/b17115

Bopp GP, Shaby BA (2017) An exponential-gamma mixture model for extreme Santa Ana winds. Environmetrics 28(8):e2476. https://doi.org/10.1002/env.2476

Craigmile PF, Guttorp P (2011) Space-time modelling of trends in temperature series. J Time Ser Anal 32(4):378–395. https://doi.org/10.1111/j.1467-9892.2011.00733.x

Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J Am Stat Assoc 111(514):800–812. https://doi.org/10.1080/01621459.2015.1044091

Diffenbaugh NS, Burke M (2019) Global warming has increased global economic inequality. Proc Natl Acad Sci 116(20):9808–9813. https://doi.org/10.1073/pnas.1816020116

Gelfand AE, Sahu SK, Carlin BP (1995) Efficient parametrisations for normal linear mixed models. Biometrika 82(3):479–488. https://doi.org/10.1093/biomet/82.3.479

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Hatfield JL, Boote KJ, Kimball BA, Ziska LH, Izaurralde RC, Ort D, Thomson AM, Wolfe D (2011) Climate impacts on agriculture: implications for crop production. Agron J 103(2):351–370. https://doi.org/10.2134/agronj2010.0303

IPCC (2013) Summary for policymakers. In: Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA

IPCC (2018) Summary for policymakers. In: Masson-Delmotte V, Zhai P, Pörtner HO, Roberts D, Skea J, Shukla PR, Pirani A, Moufouma-Okia W, Péan C, Pidcock R, Connors S, Matthews JBR, Chen Y, Zhou X, Gomis MI, Lonnoy E, Maycock T, Tignor M, Waterfield T (eds) Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty, World Meteorological Organization, Geneva, Switzerland

Lemos RT, Sansó B, Los Huertos M (2007) Spatially varying temperature trends in a Central California Estuary. J Agric Biol Environ Stat 12(3):379. https://doi.org/10.1198/108571107X227603

Li S, Griffith DA, Shu H (2020) Temperature prediction based on a space-time regression-kriging model. J Appl Stat 47(7):1168–1190. https://doi.org/10.1080/02664763.2019.1671962

Mohammadi B, Mehdizadeh S, Ahmadi F, Lien NTT, Linh NTT, Pham QB (2021) Developing hybrid time series and artificial intelligence models for estimating air temperatures. Stoch Env Res Risk Assess 35(6):1189–1204. https://doi.org/10.1007/s00477-020-01898-7

Navarro-Serrano F, López-Moreno JI, Azorin-Molina C, Alonso-González E, Tomás-Burguera M, Sanmiguel-Vallelado A, Revuelto J, Vicente-Serrano SM (2018) Estimation of near-surface air temperature lapse rates over continental Spain and its mountain areas. Int J Climatol 38(8):3233–3249. https://doi.org/10.1002/joc.5497

Peña-Angulo D, Gonzalez-Hidalgo JC, Sandonís L, Beguería S, Tomas-Burguera M, López-Bustins JA, Lemus-Canovas M, Martin-Vide J (2021) Seasonal temperature trends on the Spanish mainland: a secular study (1916–2015). Int J Climatol 41(5):3071–3084. https://doi.org/10.1002/joc.7006

Reich BJ, Shaby BA, Cooley D (2014) A hierarchical model for serially-dependent extremes: a study of heat waves in the western US. J Agric Biol Environ Stat 19(1):119–135. https://doi.org/10.1007/s13253-013-0161-y

Risser MD, Paciorek CJ, Stone DA (2019) Spatially dependent multiple testing under model misspecification, with application to detection of anthropogenic influence on extreme climate events. J Am Stat Assoc 114(525):61–78. https://doi.org/10.1080/01621459.2018.1451335

Roldán E, Gómez M, Pino MR, Pórtoles J, Linares C, Díaz J (2016) The effect of climate-change-related heat waves on mortality in Spain: uncertainties in health on a local scale. Stoch Env Res Risk Assess 30(3):831–839. https://doi.org/10.1007/s00477-015-1068-7

Rossati A (2017) Global warming and its health impact. Int J Occup Environ Med 8(1):7–20. https://doi.org/10.15171/ijoem.2017.963

Sahu SK, Gelfand AE, Holland DM (2006) Spatio-temporal modeling of fine particulate matter. J Agric Biol Environ Stat 11(1):61–86. https://doi.org/10.1198/108571106X95746

Sahu SK, Gelfand AE, Holland DM (2007) High-resolution space-time ozone modeling for assessing trends. J Am Stat Assoc 102(480):1221–1234. https://doi.org/10.1198/016214507000000031

Schlenker W, Roberts MJ (2009) Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proc Natl Acad Sci 106(37):15594–15598. https://doi.org/10.1073/pnas.0906865106

Verdin A, Rajagopalan B, Kleiber W, Katz RW (2015) Coupled stochastic weather generation using spatial and generalized linear models. Stoch Env Res Risk Assess 29(2):347–356. https://doi.org/10.1007/s00477-014-0911-6

Watts N, Adger WN, Agnolucci P, Blackstock J, Byass P, Cai W, Chaytor S, Colbourn T, Collins M, Cooper A, Cox PM, Depledge J, Drummond P, Ekins P, Galaz V, Grace D, Graham H, Grubb M, Haines A, Hamilton I, Hunter A, Jiang X, Li M, Kelman I, Liang L, Lott M, Lowe R, Luo Y, Mace G, Maslin M, Nilsson M, Oreszczyn T, Pye S, Quinn T, Svensdotter M, Venevsky S, Warner K, Xu B, Yang J, Yin Y, Yu C, Zhang Q, Gong P, Montgomery H, Costello A (2015) Health and climate change: policy responses to protect public health. The Lancet 386(10006):1861–1914. https://doi.org/10.1016/S0140-6736(15)60854-6

WMO (2017) WMO Guidelines on the Calculation of Climate Normals (WMO-No. 1203). Geneva, Switzerland, https://library.wmo.int/doc_num.php?explnum_id=4166

Zhang H (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J Am Stat Assoc 99(465):250–261. https://doi.org/10.1198/016214504000000241

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Materials for "Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain"

S1 Data and exploratory analysis

Table S1 shows the elevation in meters for each site. It also summarizes the differences in mean value and standard deviation of the daily maximum temperature in degrees Celsius for each site between both 30-year periods, 1956–1985 and 1986–2015.

S1.1 The local model

In order to motivate our spatial modeling decisions in Section 3 of the Main Manuscript, in this section we fit independent local models for each location following the steps in that section. However, here we do not center or scale the covariates. The local model for day ℓ, year t at any location simplifies the full model as

$$Y_{t\ell} = \mu_{t\ell} + \psi_t + \rho_Y(Y_{t,\ell-1} - (\mu_{t,\ell-1} + \psi_t)) + \epsilon_{t\ell}^{(Y)},$$

where the fixed effects are

$$\mu_{t\ell} = \beta_0 + \alpha t + \beta_1 \sin(2\pi\ell/365) + \beta_2 \cos(2\pi\ell/365).$$

We consider $\psi_t \overset{iid}{\sim} N(0, \sigma_\lambda^2)$ and $\epsilon_{t\ell}^{(Y)} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. The interpretation of the model terms is equivalent to that given for the full model in the Main Manuscript, so we do not repeat it here.

The parameters are shown in Figure S1 and the seasonal pattern is summarized in S2. The first figure shows the posterior mean and 90% credible interval for $\beta_0, \alpha, \rho_Y, \sigma_\epsilon$, for the independently fitted models at each location. Significant differences between locations are observed for the four parameters. The parameter σ_λ did not show a remarkable difference between most locations (not shown).

It is clear that β_0 is related with elevation, i.e., the temperature is inversely proportional to the elevation of each location. However, this relationship shows considerable noise due to the specific climatic conditions in the southern region, which has stations with an elevation close to 1,000 m with a significantly warmer temperature than stations in the north region

Table S1: Elevation, and mean value and standard deviation of the daily maximum temperature for each site in the months MJJAS for the 30-year periods, 1956–1985 and 1986–2015. Difference between means (Δ mean) and quotient of standard deviations (Q sd) of each period.

Location	Elevation	1956–1985		1986–2015		Δ mean	Q sd
		Mean	Sd	Mean	Sd		
Pamplona	442	24.2	5.9	25.0	5.9	0.8	1.0
Buñuel	242	27.5	5.1	28.5	5.4	1.0	1.1
El Bayo	360	27.6	5.4	28.0	5.3	0.4	1.0
Morella	998	22.9	5.2	24.0	5.1	1.1	1.0
Huesca	546	26.3	5.1	28.0	5.2	1.7	1.0
Tornos	1,018	26.1	6.3	27.3	6.3	1.2	1.0
Santa Eulalia	983	25.8	5.7	26.6	5.8	0.9	1.0
Calatayud	600	27.1	5.8	27.6	5.7	0.5	1.0
Panticosa	1,645	17.1	5.4	18.7	5.2	1.6	1.0
La Puebla de Híjar	245	28.7	5.2	29.8	5.1	1.1	1.0
Ansó	860	23.8	5.6	23.6	5.6	-0.2	1.0
Daroca	779	25.6	5.7	27.4	5.7	1.8	1.0
Zaragoza	249	27.6	5.2	29.3	5.2	1.7	1.0
La Sotonera	413	27.3	5.4	28.1	5.2	0.8	1.0
Pallaruelo	356	28.2	5.6	29.0	5.1	0.8	0.9
Cueva Foradada	580	25.5	4.7	26.8	5.0	1.3	1.1
Sallent	1,285	20.9	5.7	21.8	5.5	0.9	1.0
Yesa	515	26.3	6.1	26.5	5.7	0.2	0.9

with a similar elevation. The spatial variability of α suggests that there is a different warming between locations, that does not seem to be related to elevation. The ρ_Y parameter and the elevation are clearly related, but the variability around the linear relationship cannot be captured by a fixed effect. Finally, the posterior mean of σ_ϵ is quite different between locations with narrow credible intervals, showing a spatial variability of this parameter throughout the region without any relation to elevation.

The left plot in Figure S2 shows the empirical seasonal pattern at each location. In particular, the mean value over the years is drawn for each day of MJJAS and each location. The right plot shows the posterior mean of $\beta_1 \sin(2\pi\ell/365) + \beta_2 \cos(2\pi\ell/365)$ for $\ell = 1, \ldots, L$. In both plots the patterns have been centered to compare them between locations. In conclusion, the climate of Aragón shows a common and unimodal seasonal pattern in MJJAS. In particular, the seasonal component can be characterized by a single harmonic for the entire region.

In summary, this local modeling is useful to find spatial differences and similarities between the full model parameters for different points in space. Furthermore, the inclusion of spatial random effects in the model associated with the intercept, the linear trend, the autoregression coefficient and the variance is justified by these results.
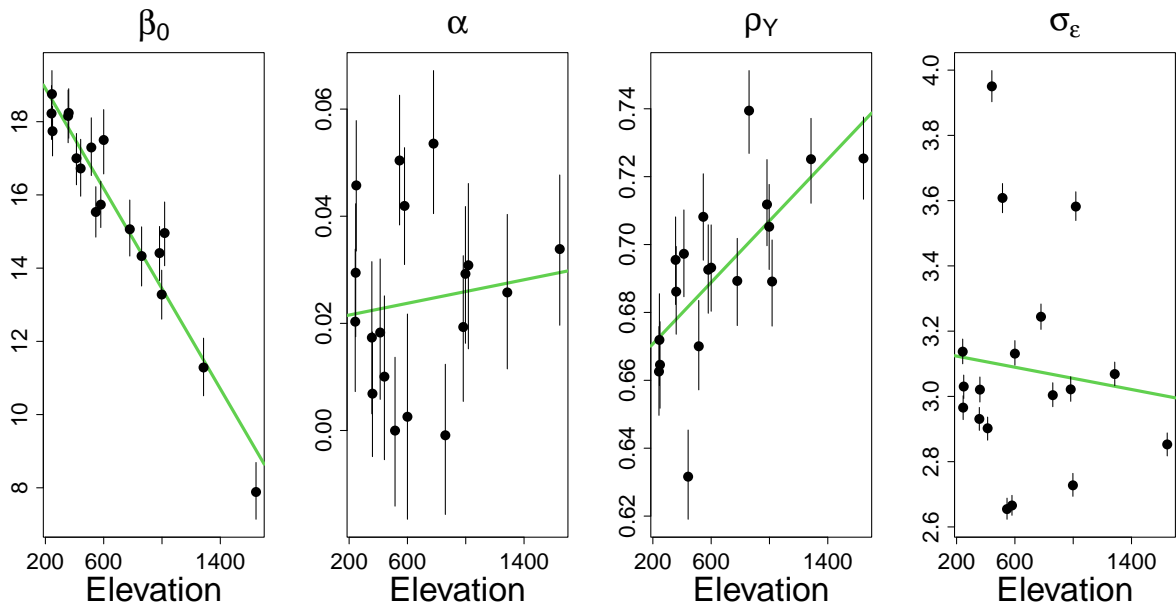
Figure S1: Posterior mean and 90% credible interval of the parameters $\beta_0, \alpha, \rho_Y, \sigma_\epsilon$ in the local models. Summaries are sorted by the elevation of the locations.
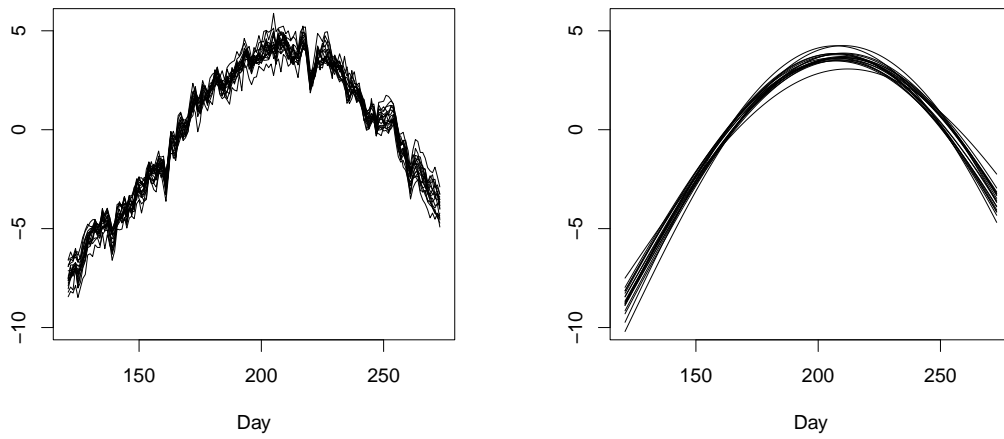

Figure S2: Left: Centered empirical mean value of each series across each day of MJJAS. Rigth: Posterior mean of the centered seasonal pattern in the local models.

S2 Sampling methods

S2.1 Gibbs sampling algorithm

The Bayesian spatio-temporal model can be represented in a hierarchical structure, following Gelfand (2012), we specify distributions for data, process and parameters in three stages,

$$
\begin{aligned}
&\text{First stage:} \quad \big[data \mid process, parameters\big] \\
&\text{Second stage:} \quad \big[process \mid parameters\big] \\
&\text{Third stage:} \quad \big[(hyper)parameters\big].
\end{aligned}
$$

Although the hierarchical model can be flattened by suitable marginalization, the advantage of the hierarchical structure lies in convenience of specification, ease of interpretation and facilitation of model fitting. In particular, since the model is Gaussian and linear, the Gibbs sampler is expected to be well behaved and convergence to be fairly quick before 50,000 iterations. The hierarchical form leads to the following joint distribution for data, processes and parameters,

$$
\prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{\ell=2}^{L}\big[Y_{t\ell}(\mathbf{s}_i) \mid Y_{t,\ell-1}(\mathbf{s}_i), \beta_1, \beta_2, \beta_3, \gamma_t(\mathbf{s}_i), \rho_Y(\mathbf{s}_i), \sigma_\epsilon^2(\mathbf{s}_i)\big]
$$

$$
\prod_{i=1}^{n}\prod_{t=1}^{T}\big[\gamma_t(\mathbf{s}_i) \mid \tilde{\beta}_0(\mathbf{s}_i), \tilde{\alpha}(\mathbf{s}_i), \psi_t, \sigma_\eta^2\big]\prod_{t=2}^{T}[\psi_t \mid \psi_{t-1}, \rho_\psi, \sigma_\lambda^2]
$$

$$
\big[\{\tilde{\beta}_0(\mathbf{s}_i)\} \mid \beta_0, \sigma_{\beta_0}^2, \phi_{\beta_0}\big]\big[\{\tilde{\alpha}(\mathbf{s}_i)\} \mid \alpha, \sigma_\alpha^2, \phi_\alpha\big]\big[\{Z_{\rho_Y}(\mathbf{s}_i)\} \mid Z_{\rho_Y}, \sigma_{\rho_Y}^2, \phi_{\rho_Y}\big]\big[\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\} \mid Z_{\sigma_\epsilon^2}, \sigma_{\sigma_\epsilon^2}^2, \phi_{\sigma_\epsilon^2}\big]
$$

$$
\big[\beta_0\big]\big[\alpha\big]\big[\beta_1\big]\big[\beta_2\big]\big[\beta_3\big]\big[Z_{\rho_Y}\big]\big[Z_{\sigma_\epsilon^2}\big]\big[\rho_\psi\big]\big[\sigma_\lambda^2\big]\big[\sigma_\eta^2\big]\big[\sigma_{\beta_0}^2\big]\big[\sigma_\alpha^2\big]\big[\sigma_{\rho_Y}^2\big]\big[\sigma_{\sigma_\epsilon^2}^2\big]\big[\phi_{\beta_0}\big]\big[\phi_\alpha\big]\big[\phi_{\rho_Y}\big]\big[\phi_{\sigma_\epsilon^2}\big],
\tag{S1}
$$

provided one starts any year t with the observed $Y_{t1}(\mathbf{s})$.

Defining notation that will be used to shorten the expressions, we denote the elements of the correlation matrices by $\left(r_{jk}^{(\cdot)}\right)^{-1} = R(\phi_\cdot) = (\exp\{-\phi_\cdot||\mathbf{s}_j - \mathbf{s}_k||\})$ where $\cdot$ is any of $\beta_0, \alpha, \rho_Y, \sigma_\epsilon^2$. We denote $X_{t\ell i} = Y_{t\ell}(\mathbf{s}_i) - (\mu_{t\ell}(\mathbf{s}_i; \boldsymbol{\theta}_f) + \gamma_t(\mathbf{s}_i))$ and $X_{t\ell i}^{(-\cdot)} = Y_{t\ell}(\mathbf{s}_i) - (\mu_{t\ell}(\mathbf{s}_i; \boldsymbol{\theta}_f) + \gamma_t(\mathbf{s}_i))^{(-\cdot)}$, where here $\cdot$ is any of $\beta_1, \beta_2, \beta_3, \gamma_t(\mathbf{s}_i)$ and represents that $\mu_{t\ell}(\mathbf{s}_i; \boldsymbol{\theta}_f) + \gamma_t(\mathbf{s}_i)$ does not have that component. Finally, we shorten $\sin_\ell = \sin(2\pi\ell/365)$ and $\cos_\ell = \cos(2\pi\ell/365)$. And a and b denote chosen hyperpriors in each case.

The Gibbs sampler algorithm for Equation S1 is initialized giving initial values to all the parameters. Then, updating from iteration b to $b+1$ consists of drawing a sample from the following full conditional distributions:

- The full conditional distributions of $\beta_0, \alpha, \beta_1, \beta_2, \beta_3, Z_{\rho_Y}, Z_{\sigma_\epsilon^2}$ are all Gaussian, in particular

$$
[\beta_0 \mid \cdots] \propto N\left(\beta_0 \mid \frac{\sum_{j,k} r_{jk}^{(\beta_0)} \tilde{\beta}_0(\mathbf{s}_k)}{\sum_{j,k} r_{jk}^{(\beta_0)}}, \frac{\sigma_{\beta_0}^2}{\sum_{j,k} r_{jk}^{(\beta_0)}}\right) \times N(\beta_0 \mid a_{\beta_0}, b_{\beta_0}^2)
$$

$$
[\alpha \mid \cdots] \propto N\left(\alpha \mid \frac{\sum_{j,k} r_{jk}^{(\alpha)} \tilde{\alpha}(\mathbf{s}_k)}{\sum_{j,k} r_{jk}^{(\alpha)}}, \frac{\sigma_\alpha^2}{\sum_{j,k} r_{jk}^{(\alpha)}}\right) \times N(\alpha \mid a_\alpha, b_\alpha^2)
$$

$$[\beta_1 \mid \cdots] \propto \prod_{i=1}^{n} N \left(\beta_1 \mid \frac{\sum_{t=1}^{T} \sum_{\ell=2}^{L} (\sin_\ell - \rho_Y(\mathbf{s}_i) \sin_{\ell-1})(X_{t\ell i}^{(-\beta_1)} - \rho_Y(\mathbf{s}_i) X_{t,\ell-1,i}^{(-\beta_1)})}{T \sum_{\ell=2}^{L} (\sin_\ell - \rho_Y(\mathbf{s}_i) \sin_{\ell-1})^2}, \right.$$

$$\left. \frac{\sigma_\epsilon^2(\mathbf{s}_i)}{T \sum_{\ell=2}^{L} (\sin_\ell - \rho_Y(\mathbf{s}_i) \sin_{\ell-1})^2} \right) \times N(\beta_1 \mid a_{\beta_1}, b_{\beta_1}^2)$$

$$[\beta_2 \mid \cdots] \propto \prod_{i=1}^{n} N \left(\beta_2 \mid \frac{\sum_{t=1}^{T} \sum_{\ell=2}^{L} (\cos_\ell - \rho_Y(\mathbf{s}_i) \cos_{\ell-1})(X_{t\ell i}^{(-\beta_2)} - \rho_Y(\mathbf{s}_i) X_{t,\ell-1,i}^{(-\beta_2)})}{T \sum_{\ell=2}^{L} (\cos_\ell - \rho_Y(\mathbf{s}_i) \cos_{\ell-1})^2}, \right.$$

$$\left. \frac{\sigma_\epsilon^2(\mathbf{s}_i)}{T \sum_{\ell=2}^{L} (\cos_\ell - \rho_Y(\mathbf{s}_i) \cos_{\ell-1})^2} \right) \times N(\beta_2 \mid a_{\beta_2}, b_{\beta_2}^2)$$

$$[\beta_3 \mid \cdots] \propto \prod_{i=1}^{n} N \left(\beta_3 \mid \frac{\sum_{t=1}^{T} \sum_{\ell=2}^{L} (X_{t\ell i}^{(-\beta_3)} - \rho_Y(\mathbf{s}_i) X_{t,\ell-1,i}^{(-\beta_3)})}{T(L-1)(1 - \rho_Y(\mathbf{s}_i))\text{elev}(\mathbf{s}_i)}, \right.$$

$$\left. \frac{\sigma_\epsilon^2(\mathbf{s}_i)}{T(L-1)(1 - \rho_Y(\mathbf{s}_i))^2 \text{elev}(\mathbf{s}_i)^2} \right) \times N(\beta_3 \mid a_{\beta_3}, b_{\beta_3}^2)$$

$$[Z_{\rho_Y} \mid \cdots] \propto N \left(Z_{\rho_Y} \mid \frac{\sum_{j,k} r_{jk}^{(\rho_Y)} Z_{\rho_Y}(\mathbf{s}_k)}{\sum_{j,k} r_{jk}^{(\rho_Y)}}, \frac{\sigma_{\rho_Y}^2}{\sum_{j,k} r_{jk}^{(\rho_Y)}} \right) \times N(Z_{\rho_Y} \mid a_{\rho_Y}, b_{\rho_Y}^2)$$

$$[Z_{\sigma_\epsilon^2} \mid \cdots] \propto N \left(Z_{\sigma_\epsilon^2} \mid \frac{\sum_{j,k} r_{jk}^{(\sigma_\epsilon^2)} Z_{\sigma_\epsilon^2}(\mathbf{s}_k)}{\sum_{j,k} r_{jk}^{(\sigma_\epsilon^2)}}, \frac{\sigma_{\sigma_\epsilon^2}^2}{\sum_{j,k} r_{jk}^{(\sigma_\epsilon^2)}} \right) \times N(Z_{\sigma_\epsilon^2} \mid a_{\sigma_\epsilon^2}, b_{\sigma_\epsilon^2}^2)$$

- The full conditional distribution of ρ_ψ is a truncated Gaussian distribution within the interval (a, b).

$$[\rho_\psi \mid \cdots] \sim TN \left(\frac{\sum_{t=2}^{T} \psi_t \psi_{t-1}}{\sum_{t=2}^{T} \psi_{t-1}^2}, \frac{\sigma_\lambda^2}{\sum_{t=2}^{T} \psi_{t-1}^2}, (a_{\rho_\psi}, b_{\rho_\psi}) \right)$$

- The full conditional distributions for $\sigma_\lambda^2, \sigma_\eta^2, \sigma_{\beta_0}^2, \sigma_\alpha^2, \sigma_{\rho_Y}^2, \sigma_{\sigma_\epsilon^2}^2$ are all inverse gamma as

follows,

$$[1/\sigma_\lambda^2 \mid \cdots] \sim G\left(\frac{T-1}{2} + a_{\sigma_\lambda}, \frac{1}{2}\sum_{t=2}^{T}(\psi_t - \rho_\psi \psi_{t-1})^2 + b_{\sigma_\lambda}\right)$$

$$[1/\sigma_\eta^2 \mid \cdots] \sim G\left(\frac{nT}{2} + a_{\sigma_\eta}, \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\gamma_t(\mathbf{s}_i) - (\tilde{\beta}_0(\mathbf{s}_i) + \tilde{\alpha}(\mathbf{s}_i)t + \psi_t)\right)^2 + b_{\sigma_\eta}\right)$$

$$[1/\sigma_{\beta_0}^2 \mid \cdots] \sim G\left(\frac{n}{2} + a_{\sigma_{\beta_0}}, \frac{1}{2}(\{\tilde{\beta}_0(\mathbf{s}_i)\} - \beta_0 \mathbf{1})^\top R(\phi_{\beta_0})^{-1}(\{\tilde{\beta}_0(\mathbf{s}_i)\} - \beta_0 \mathbf{1}) + b_{\sigma_{\beta_0}}\right)$$

$$[1/\sigma_\alpha^2 \mid \cdots] \sim G\left(\frac{n}{2} + a_{\sigma_\alpha}, \frac{1}{2}(\{\tilde{\alpha}(\mathbf{s}_i)\} - \alpha \mathbf{1})^\top R(\phi_\alpha)^{-1}(\{\tilde{\alpha}(\mathbf{s}_i)\} - \alpha \mathbf{1}) + b_{\sigma_\alpha}\right)$$

$$[1/\sigma_{\rho_Y}^2 \mid \cdots] \sim G\left(\frac{n}{2} + a_{\rho_Y}, \frac{1}{2}(\{Z_{\rho_Y}(\mathbf{s}_i)\} - Z_{\rho_Y}\mathbf{1})^\top R(\phi_{\rho_Y})^{-1}(\{Z_{\rho_Y}(\mathbf{s}_i)\} - Z_{\rho_Y}\mathbf{1}) + b_{\rho_Y}\right)$$

$$[1/\sigma_{\sigma_\epsilon^2}^2 \mid \cdots] \sim G\left(\frac{n}{2} + a_{\sigma_\epsilon^2}, \frac{1}{2}(\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\} - Z_{\sigma_\epsilon^2}\mathbf{1})^\top R(\phi_{\sigma_\epsilon^2})^{-1}(\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\} - Z_{\sigma_\epsilon^2}\mathbf{1}) + b_{\sigma_\epsilon^2}\right)$$

- For $\phi_{\beta_0}, \phi_\alpha, \phi_{\rho_Y}, \phi_{\sigma_\epsilon^2}$ the simplest solution is to fix the parameter at some reasonable value. An alternative is to discretize the support to say m between 10 to 20 values, obtain and store the collection of $n \times n$ matrices, i.e., inverses and determinants, and then make discrete updates from the following full conditionals.

$$[\phi_{\beta_0} \mid \cdots] \propto |R(\phi_{\beta_0})|^{-1/2}\exp\left\{\frac{-1}{2\sigma_{\beta_0}^2}(\{\tilde{\beta}_0(\mathbf{s}_i)\} - \beta_0 \mathbf{1})^\top R(\phi_{\beta_0})^{-1}(\{\tilde{\beta}_0(\mathbf{s}_i)\} - \beta_0 \mathbf{1})\right\}$$
$$\times U\left\{a_{\phi_{\beta_0}}^{(1)}, \ldots, a_{\phi_{\beta_0}}^{(m)}\right\}$$

$$[\phi_\alpha \mid \cdots] \propto |R(\phi_\alpha)|^{-1/2}\exp\left\{\frac{-1}{2\sigma_\alpha^2}(\{\tilde{\alpha}(\mathbf{s}_i)\} - \alpha \mathbf{1})^\top R(\phi_\alpha)^{-1}(\{\tilde{\alpha}(\mathbf{s}_i)\} - \alpha \mathbf{1})\right\}$$
$$\times U\left\{a_{\phi_\alpha}^{(1)}, \ldots, a_{\phi_\alpha}^{(m)}\right\}$$

$$[\phi_{\rho_Y} \mid \cdots] \propto |R(\phi_{\rho_Y})|^{-1/2}\exp\left\{\frac{-1}{2\sigma_{\rho_Y}^2}(\{Z_{\rho_Y}(\mathbf{s}_i)\} - Z_{\rho_Y}\mathbf{1})^\top R(\phi_{\rho_Y})^{-1}(\{Z_{\rho_Y}(\mathbf{s}_i)\} - Z_{\rho_Y}\mathbf{1})\right\}$$
$$\times U\left\{a_{\phi_{\rho_Y}}^{(1)}, \ldots, a_{\phi_{\rho_Y}}^{(m)}\right\}$$

$$[\phi_{\sigma_\epsilon^2} \mid \cdots] \propto |R(\phi_{\sigma_\epsilon^2})|^{-1/2}\exp\left\{\frac{-1}{2\sigma_{\sigma_\epsilon^2}^2}(\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\} - Z_{\sigma_\epsilon^2}\mathbf{1})^\top R(\phi_{\sigma_\epsilon^2})^{-1}(\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\} - Z_{\sigma_\epsilon^2}\mathbf{1})\right\}$$
$$\times U\left\{a_{\phi_{\sigma_\epsilon^2}}^{(1)}, \ldots, a_{\phi_{\sigma_\epsilon^2}}^{(m)}\right\}$$

- The full conditionals for the $\tilde{\beta}_0(\mathbf{s}_i)$'s and $\tilde{\alpha}(\mathbf{s}_i)$'s are Gaussian, coming from the joint multivariate Gaussian distributions of $\tilde{\beta}_0(\mathbf{s})$ and $\tilde{\alpha}(\mathbf{s})$ respectively, and the part of the random effects. Note that we consider the hierarchical centering of these random effects

to improve convergence behavior. For $i = 1, \ldots, n$, the full conditionals are

$$[\tilde{\beta}_0(\mathbf{s}_i) \mid \cdots] \propto N\left(\tilde{\beta}_0(\mathbf{s}_i) \mid \frac{1}{T}\sum_{t=1}^{T}(\gamma_t(\mathbf{s}_i) - \tilde{\alpha}(\mathbf{s}_i)t - \psi_t), \frac{\sigma_\eta^2}{T}\right)$$

$$\times N\left(\tilde{\beta}_0(\mathbf{s}_i) \mid \beta_0 + \frac{\sum_{k\neq i} r_{ik}^{(\beta_0)}(\beta_0 - \tilde{\beta}_0(\mathbf{s}_k))}{r_{ii}^{(\beta_0)}}, \frac{\sigma_{\beta_0}^2}{r_{ii}^{(\beta_0)}}\right)$$

$$[\tilde{\alpha}(\mathbf{s}_i) \mid \cdots] \propto N\left(\tilde{\alpha}(\mathbf{s}_i) \mid \frac{\sum_{t=1}^{T} t(\gamma_t(\mathbf{s}_i) - \tilde{\beta}_0(\mathbf{s}_i) - \psi_t)}{\sum_{t=1}^{T} t^2}, \frac{\sigma_\eta^2}{\sum_{t=1}^{T} t^2}\right)$$

$$\times N\left(\tilde{\alpha}(\mathbf{s}_i) \mid \alpha + \frac{\sum_{k\neq i} r_{ik}^{(\alpha)}(\alpha - \tilde{\alpha}(\mathbf{s}_k))}{r_{ii}^{(\alpha)}}, \frac{\sigma_\alpha^2}{r_{ii}^{(\alpha)}}\right)$$

- The full conditional distributions of the $Z_{\rho_Y}(\mathbf{s}_i)$'s and $Z_{\sigma_\epsilon^2}(\mathbf{s}_i)$'s are non-standard. To draw samples from them, we suggest a random walk Metropolis-Hastings algorithm with Gaussian distribution proposals with the mean at the current parameter value. According to Gelman et al. (1996), the variance of the proposals should be tuned until the acceptance rate is between 15% and 40%. For $i = 1, \ldots, n$, the full conditionals are proportional to

$$[Z_{\rho_Y}(\mathbf{s}_i) \mid \cdots] \propto \exp\left\{\frac{-1}{2\sigma_\epsilon^2(\mathbf{s}_i)}\sum_{t=1}^{T}\sum_{\ell=2}^{L}\left(X_{t\ell i} - \frac{e^{Z_{\rho_Y}(\mathbf{s}_i)} - 1}{e^{Z_{\rho_Y}(\mathbf{s}_i)} + 1}X_{t,\ell-1,i}\right)^2\right\}$$

$$\times N\left(Z_{\rho_Y}(\mathbf{s}_i) \mid Z_{\rho_Y} + \frac{\sum_{k\neq i} r_{ik}^{(\rho_Y)}(Z_{\rho_Y} - Z_{\rho_Y}(\mathbf{s}_k))}{r_{ii}^{(\rho_Y)}}, \frac{\sigma_{\rho_Y}^2}{r_{ii}^{(\rho_Y)}}\right)$$

$$[Z_{\sigma_\epsilon^2}(\mathbf{s}_i) \mid \cdots] \propto \exp\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\}^{-T(L-1)/2}\exp\left\{\frac{-1}{2\exp\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\}}\sum_{t=1}^{T}\sum_{\ell=2}^{L}(X_{t\ell i} - \rho_Y(\mathbf{s}_i)X_{t,\ell-1,i})^2\right\}$$

$$\times N\left(Z_{\sigma_\epsilon^2}(\mathbf{s}_i) \mid Z_{\sigma_\epsilon^2} + \frac{\sum_{k\neq i} r_{ik}^{(\sigma_\epsilon^2)}(Z_{\sigma_\epsilon^2} - Z_{\sigma_\epsilon^2}(\mathbf{s}_k))}{r_{ii}^{(\sigma_\epsilon^2)}}, \frac{\sigma_{\sigma_\epsilon^2}^2}{r_{ii}^{(\sigma_\epsilon^2)}}\right)$$

- We obtain the Gaussian full conditionals for the ψ's as follows. For identifiability, ψ_1 is fixed to zero. Then, two cases are considered: (i) when $t = 2, \ldots, T-1$, and (ii) when $t = T$. Then, respectively

$$[\psi_t \mid \psi_{t-1}, \psi_{t+1}, \cdots] \propto N\left(\psi_t \mid \frac{1}{n}\sum_{i=1}^{n}(\gamma_t(\mathbf{s}_i) - \tilde{\beta}_0(\mathbf{s}_i) - \tilde{\alpha}(\mathbf{s}_i)t), \frac{\sigma_\eta^2}{n}\right)$$

$$\times N\left(\psi_t \mid \frac{\rho_\psi(\psi_{t-1} + \psi_{t+1})}{1 + \rho_\psi^2}, \frac{\sigma_\lambda^2}{1 + \rho_\psi^2}\right)$$

$$[\psi_T \mid \psi_{T-1}, \cdots] \propto N\left(\psi_T \mid \frac{1}{n}\sum_{i=1}^{n}(\gamma_T(\mathbf{s}_i) - \tilde{\beta}_0(\mathbf{s}_i) - \tilde{\alpha}(\mathbf{s}_i)T), \frac{\sigma_\eta^2}{n}\right) \times N\left(\psi_T \mid \rho_\psi\psi_{T-1}, \sigma_\lambda^2\right)$$

7

- Finally, the full conditionals for the $\gamma_t(\mathbf{s}_i)$'s are all Gaussian. For $i = 1, \ldots, n$ and $t = 1, \ldots, T$,

$$[\gamma_t(\mathbf{s}_i) \mid \cdots] \propto N\left(\gamma_t(\mathbf{s}_i) \mid \frac{\sum_{\ell=2}^{L}(X_{t\ell i}^{(-\gamma_t(\mathbf{s}_i))} - \rho_Y(\mathbf{s}_i)X_{t,\ell-1,i}^{(-\gamma_t(\mathbf{s}_i))})}{(L-1)(1-\rho_Y(\mathbf{s}_i))}, \frac{\sigma_\epsilon^2(\mathbf{s}_i)}{(L-1)(1-\rho_Y(\mathbf{s}_i))^2}\right)$$
$$\times N\left(\gamma_t(\mathbf{s}_i) \mid \tilde{\beta}_0(\mathbf{s}_i) + \tilde{\alpha}(\mathbf{s}_i)t + \psi_t, \sigma_\eta^2\right)$$

Note in the expressions above that the product of Gaussian densities is proportional to a Gaussian density with parameters as follows

$$\prod_{i=1}^{n} N\left(x \mid \mu_i, \sigma_i^2\right) \propto N\left(x \mid \sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2}/\sum_{i=1}^{n}\frac{1}{\sigma_i^2}, 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\right).$$

S2.2 Composition sampling algorithm and Bayesian kriging

Following the notation of Section 3.3 of the Main Manuscript. Once samples of the join posterior distribution have been obtained using the Gibbs sampler algorithm in Section S2.1, one may want to make spatial or space-time predictions. Formally, the posterior predictive distribution for $Y_{t'\ell'}(\mathbf{s}_0)$ is

$$\begin{aligned}
&[Y_{t'\ell'}(\mathbf{s}_0) \mid \mathbf{Y}] \\
&= \int \prod_{\ell=2}^{\ell'} [Y_{t'\ell}(\mathbf{s}_0) \mid Y_{t',\ell-1}(\mathbf{s}_0), \boldsymbol{\theta}_f, \gamma_{t'}(\mathbf{s}_0), \rho_Y(\mathbf{s}_0), \sigma_\epsilon^2(\mathbf{s}_0)] \\
&\quad \times [\gamma_{t'}(\mathbf{s}_0) \mid \tilde{\beta}_0(\mathbf{s}_0), \tilde{\alpha}(\mathbf{s}_0), \psi_{t'}, \sigma_\eta^2] \\
&\quad \times [\tilde{\beta}_0(\mathbf{s}_0) \mid \beta_0, \sigma_{\beta_0}^2, \phi_{\beta_0}, \{\tilde{\beta}_0(\mathbf{s}_i)\}] \times [\tilde{\alpha}(\mathbf{s}_0) \mid \alpha, \sigma_\alpha^2, \phi_\alpha, \{\tilde{\alpha}(\mathbf{s}_i)\}] \\
&\quad \times [Z_{\rho_Y}(\mathbf{s}_0) \mid Z_{\rho_Y}, \sigma_{\rho_Y}^2, \phi_{\rho_Y}, \{Z_{\rho_Y}(\mathbf{s}_i)\}] \times [Z_{\sigma_\epsilon^2}(\mathbf{s}_0) \mid Z_{\sigma_\epsilon^2}, \sigma_{\sigma_\epsilon^2}^2, \phi_{\sigma_\epsilon^2}, \{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\}] \\
&\quad \times [\boldsymbol{\theta}, \{\tilde{\beta}_0(\mathbf{s}_i)\}, \{\tilde{\alpha}(\mathbf{s}_i)\}, \{Z_{\rho_Y}(\mathbf{s}_i)\}, \{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\}, \psi_{t'} \mid \mathbf{Y}] \\
&\quad \times \prod_{\ell=2}^{\ell'-1} dY_{t'\ell}(\mathbf{s}_0)d\gamma_{t'}(\mathbf{s}_0)d\tilde{\beta}_0(\mathbf{s}_0)d\tilde{\alpha}(\mathbf{s}_0)dZ_{\rho_Y}(\mathbf{s}_0)dZ_{\sigma_\epsilon^2}(\mathbf{s}_0)d\boldsymbol{\theta} \\
&\quad\quad d\{\tilde{\beta}_0(\mathbf{s}_i)\}d\{\tilde{\alpha}(\mathbf{s}_i)\}d\{Z_{\rho_Y}(\mathbf{s}_i)\}d\{Z_{\sigma_\epsilon^2}(\mathbf{s}_i)\}d\psi_{t'},
\end{aligned}$$

where $\mathbf{Y}$ denotes the observed data and $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \boldsymbol{\theta}_r, Z_{\rho_Y}, Z_{\sigma_\epsilon^2}, \sigma_\eta^2, \sigma_{\rho_Y}^2, \sigma_{\sigma_\epsilon^2}^2, \phi_{\rho_Y}, \phi_{\sigma_\epsilon^2})$ all the parameters. Note that prediction at a new location would require additional modeling for $Y_{t'1}(\mathbf{s}_0)$, although no additional complications arise. The simplest solution is to obtain these values by ordinary kriging.

The samples from the Gibbs sampler are used to obtain samples from the posterior predictive distribution. First, a random sample is drawn from the posterior distribution using the details in the Gibbs sampler algorithm described in Section S2.1. Then, from each GP, say $W(\mathbf{s})$, Bayesian kriging is used to draw a sample from the conditional distribution of $W(\mathbf{s}_0)$ given $\{W(\mathbf{s}_i)\}$, details are given in the paragraph below. A sample of $\gamma_{t'}(\mathbf{s}_0)$ is drawn from

$$[\gamma_{t'}(\mathbf{s}_0) \mid \cdots] \sim N\left(\tilde{\beta}_0(\mathbf{s}_0) + \tilde{\alpha}(\mathbf{s}_0)t' + \psi_{t'}, \sigma_\eta^2\right).$$

Finally, a sample $Y_{t'\ell}(\mathbf{s}_0)$, $\ell = 2, \ldots, \ell'$, is drawn sequentially from the top level model

$$\left[Y_{t'\ell}(\mathbf{s}_0) \mid \cdots\right] \sim N\left(\mu_{t'\ell}(\mathbf{s}_0; \boldsymbol{\theta}_f) + \gamma_{t'}(\mathbf{s}_0) + \rho_Y(\mathbf{s}_0)(Y_{t',\ell-1}(\mathbf{s}_0) - (\mu_{t',\ell-1}(\mathbf{s}_0; \boldsymbol{\theta}_f) + \gamma_{t'}(\mathbf{s}_0))), \sigma_\epsilon^2(\mathbf{s}_0)\right).$$

The details of the Bayesian kriging are as follows. In particular, we are interested in predicting the state of a GP, $W(\mathbf{s})$, at a new location $\mathbf{s}_0$. The joint distribution for $\mathbf{s} \in \{\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is a multivariate Gaussian distribution arising from the GP for $W(\mathbf{s})$, i.e.,

$$\begin{pmatrix} W(\mathbf{s}_0) \\ \{W(\mathbf{s}_i)\} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_0 \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma_{00} & \{\Sigma_{i0}\}^\top \\ \{\Sigma_{i0}\} & \boldsymbol{\Sigma} \end{pmatrix} \right).$$

Therefore, the conditional distribution of the process at $\mathbf{s}_0$ is

$$\left[W(\mathbf{s}_0) \mid \{W(\mathbf{s}_i)\} = \mathbf{w}, \cdots\right] \sim N\left(\mu_0 + \{\Sigma_{i0}\}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}), \Sigma_{00} - \{\Sigma_{i0}\}^\top \boldsymbol{\Sigma}^{-1}\{\Sigma_{i0}\}\right),$$

from which we would draw a sample. In particular, to obtain a sample for $\rho_Y(\mathbf{s}_0)$ or $\sigma_\epsilon^2(\mathbf{s}_0)$, it is enough to apply to the samples of their associated GPs the inverse function of the transformation applied to them.

S2.3 MCMC convergence diagnostics

In the MCMC fitting we run 10 chains and 200,000 iterations on each chain to obtain samples from the joint posterior distribution. The first 100,000 samples were discarded as burn-in and the remaining 100,000 samples were thinned (i) to retain 1,000 samples from each chain for computing the estimated potential scale reduction factor ($\hat{R}$; Gelman and Rubin, 1992), and (ii) to retain 100 samples from each chain for computing the effective sample size (ESS; Gong and Flegal, 2016) out of 1,000 samples and showing trace plots. The samples from (ii) were used for posterior inference.

We check the convergence and mixing of the MCMC algorithm for the full model based on diagnostics and trace plots for all the parameters, although we do not show the individual results for ψ's and γ's (a total of $T \times (n+1)$ parameters). Tables S2 and S3 show the ESS and the $\hat{R}$ for the main parameters and the GPs at the observed locations, respectively. The best ESS should be close to the actual sample size of 1,000, although an ESS of around 200 is considered sufficient. On the other hand, if the 10 chains have converged to the target posterior distribution, then $\hat{R}$ should be close to 1. In particular, if $\hat{R} < 1.2$ for all model parameters, one can be confident that convergence has been reached. The ESS is around 1,000 in most parameters, but it is particularly small for β_3 and $\beta_0(\mathbf{s}_i)$'s due to the high correlation between them, although their ESS is sufficient. The $\hat{R}$ is below 1.2 for all the parameters in the tables, ψ's and γ's (not shown), which suggests the adequate convergence of the chains. Figure S3 shows the trace plots of all 1,000 samples of the parameters.

Table S2: Convergence diagnostics for the main parameters of the full model.

|	ESS	$\hat{R}$
β_0	1,000	1.01
α	1,000	1.00
β_1	1,000	1.00
β_2	1,000	1.00
β_3	226	1.12
ρ_Y	1,000	1.00
σ_ϵ^2	1,000	1.00
σ_η^2	1,000	1.00
σ_λ^2	1,000	1.00
$\sigma_{\beta_0}^2$	1,000	1.00
σ_α^2	1,000	1.00
$\sigma_{\rho_Y}^2$	1,000	1.00
$\sigma_{\sigma_\epsilon^2}^2$	1,049	1.00

Table S3: Convergence diagnostics for the GPs at the observed locations of the full model.

Location	$\tilde{\beta}_0(\mathbf{s})$		$\tilde{\alpha}(\mathbf{s})$		$\rho_Y(\mathbf{s})$		$\sigma_Y^2(\mathbf{s})$	
	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$
Pamplona	247	1.07	1,000	1.00	1,000	1.00	1,000	1.00
Buñuel	194	1.11	1,000	1.00	1,000	1.00	1,000	1.00
El Bayo	215	1.09	1,000	1.00	1,000	1.00	1,000	1.00
Morella	190	1.10	1,000	1.00	1,000	1.00	1,000	1.00
Huesca	429	1.04	1,000	1.00	1,000	1.00	1,000	1.00
Tornos	219	1.10	1,000	1.00	1,000	1.00	1,000	1.00
Santa Eulalia	213	1.10	1,000	1.00	1,000	1.00	1,000	1.00
Calatayud	745	1.02	1,000	1.00	1,000	1.00	1,120	1.00
Panticosa	232	1.12	1,000	1.00	1,000	1.00	1,000	1.00
La Puebla de Híjar	220	1.10	1,000	1.00	1,076	1.00	1,000	1.00
Ansó	279	1.07	1,000	1.00	1,000	1.00	1,000	1.00
Daroca	627	1.04	915	1.00	1,000	1.00	1,000	1.00
Zaragoza	217	1.10	1,009	1.00	1,000	1.00	914	1.00
La Sotonera	247	1.08	1,000	1.00	1,000	1.00	1,000	1.00
Pallaruelo	218	1.09	1,000	1.00	1,000	1.00	1,000	1.00
Cueva Foradada	688	1.02	1,000	1.00	1,000	1.00	1,000	1.00
Sallent	181	1.12	1,000	1.00	1,000	1.00	1,000	1.00
Yesa	334	1.05	1,000	1.00	1,000	1.00	981	1.00
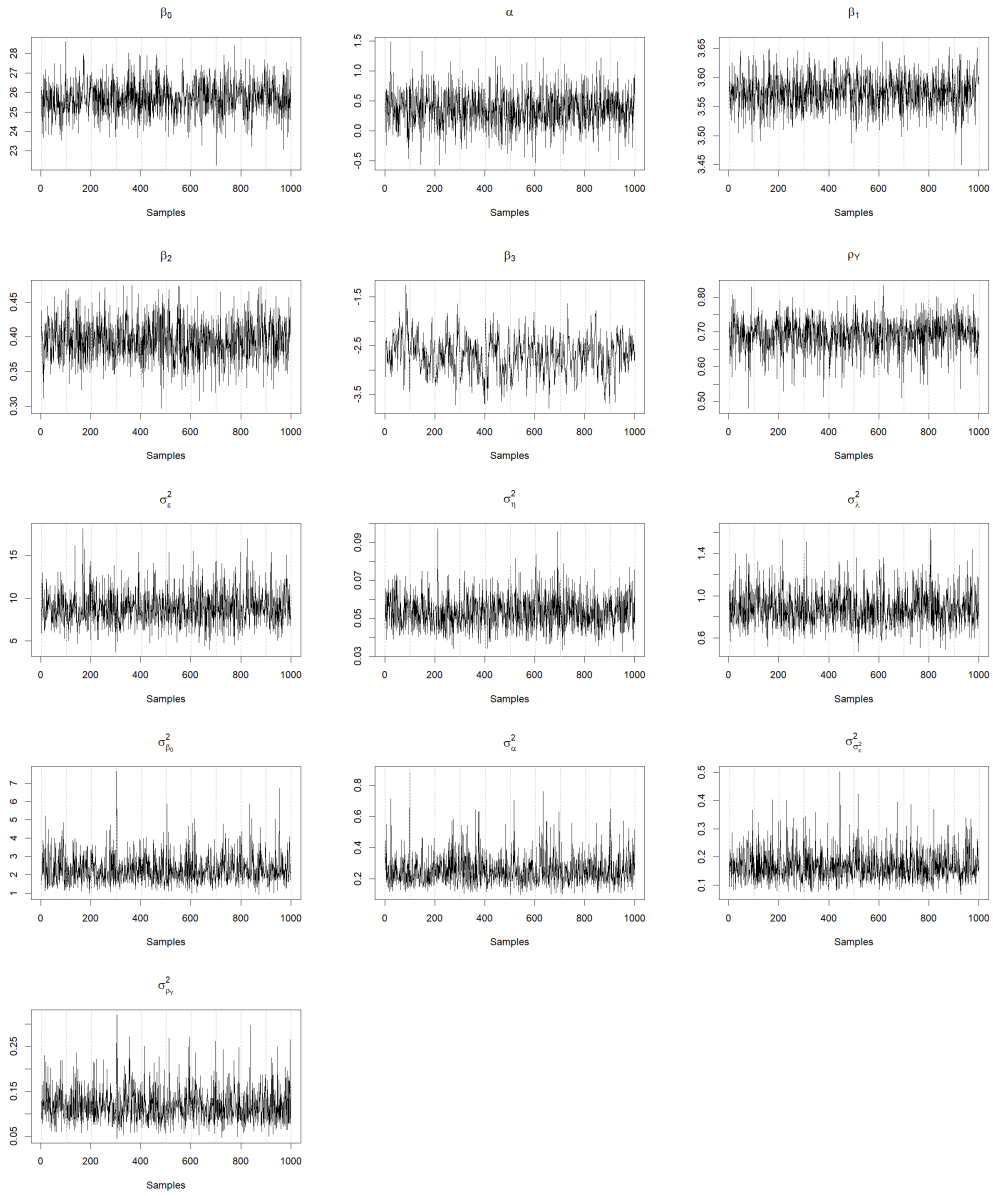
(a) Main parameters (not rescaled)

Figure S3: Trace plots for the parameters (1 of 5).
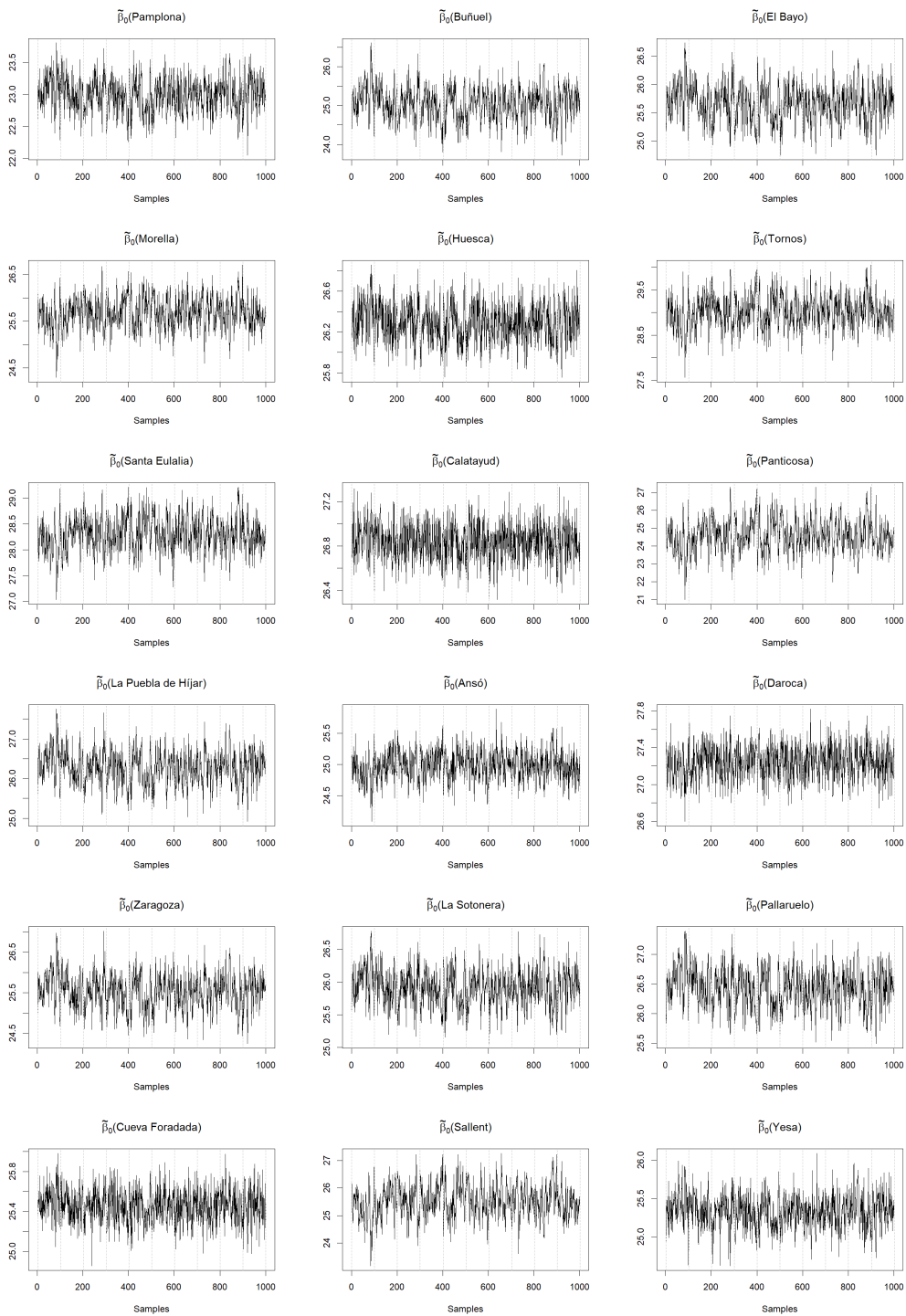
(b) $\tilde{\beta}_0(\mathbf{s})$

Figure S3: Trace plots for the parameters (2 of 5).
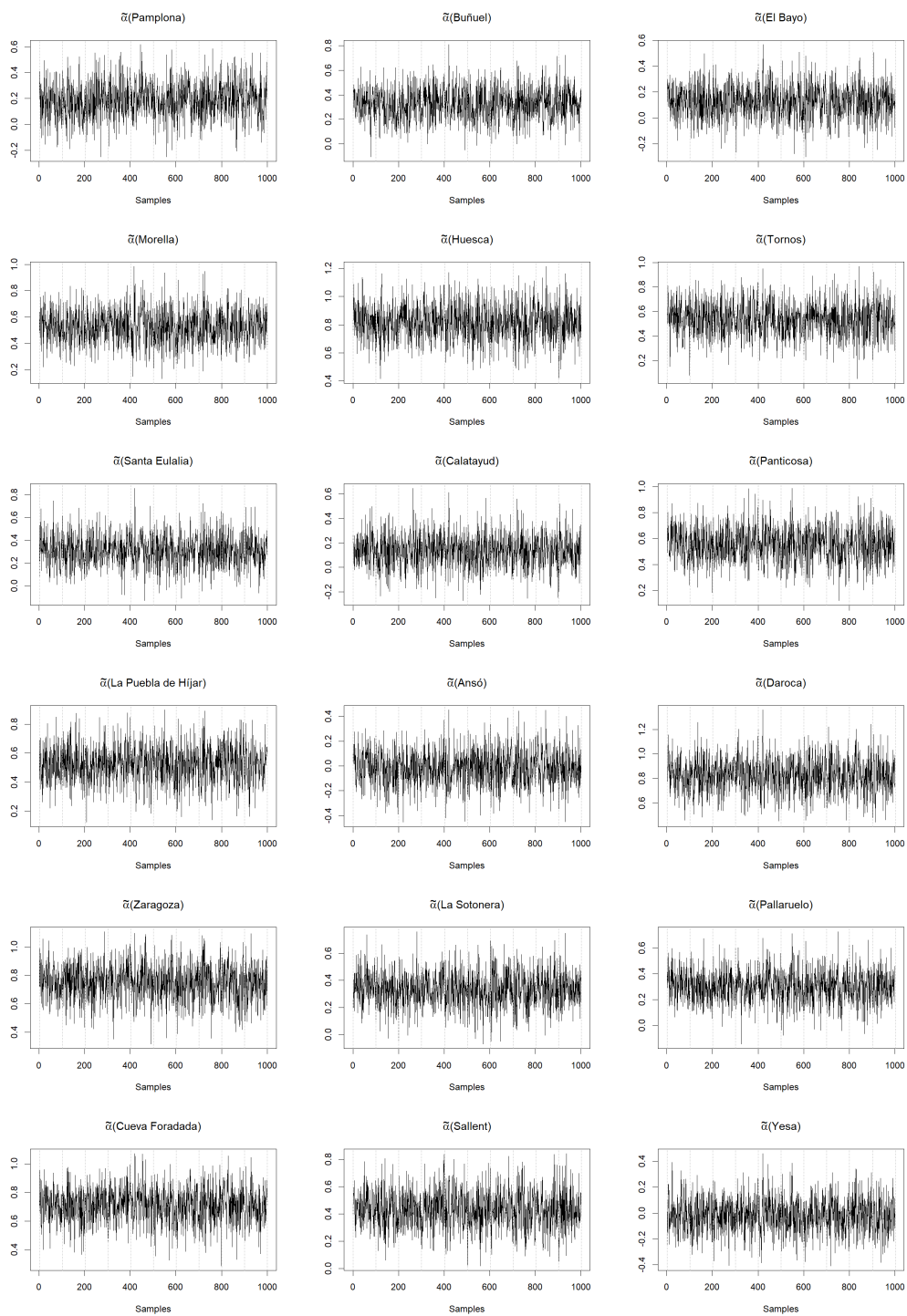
(c) $\tilde{\alpha}(\mathbf{s})$ (not rescaled)
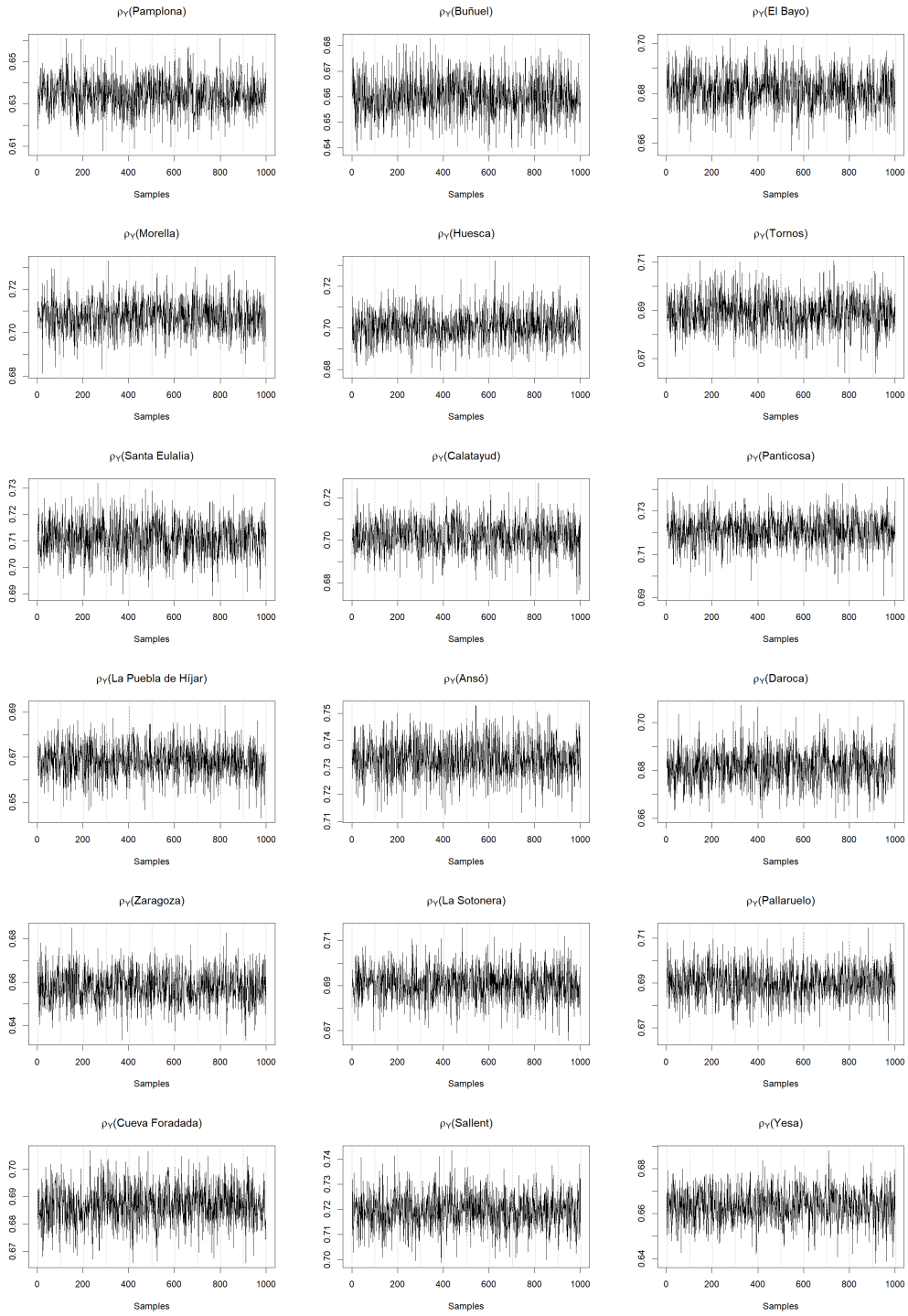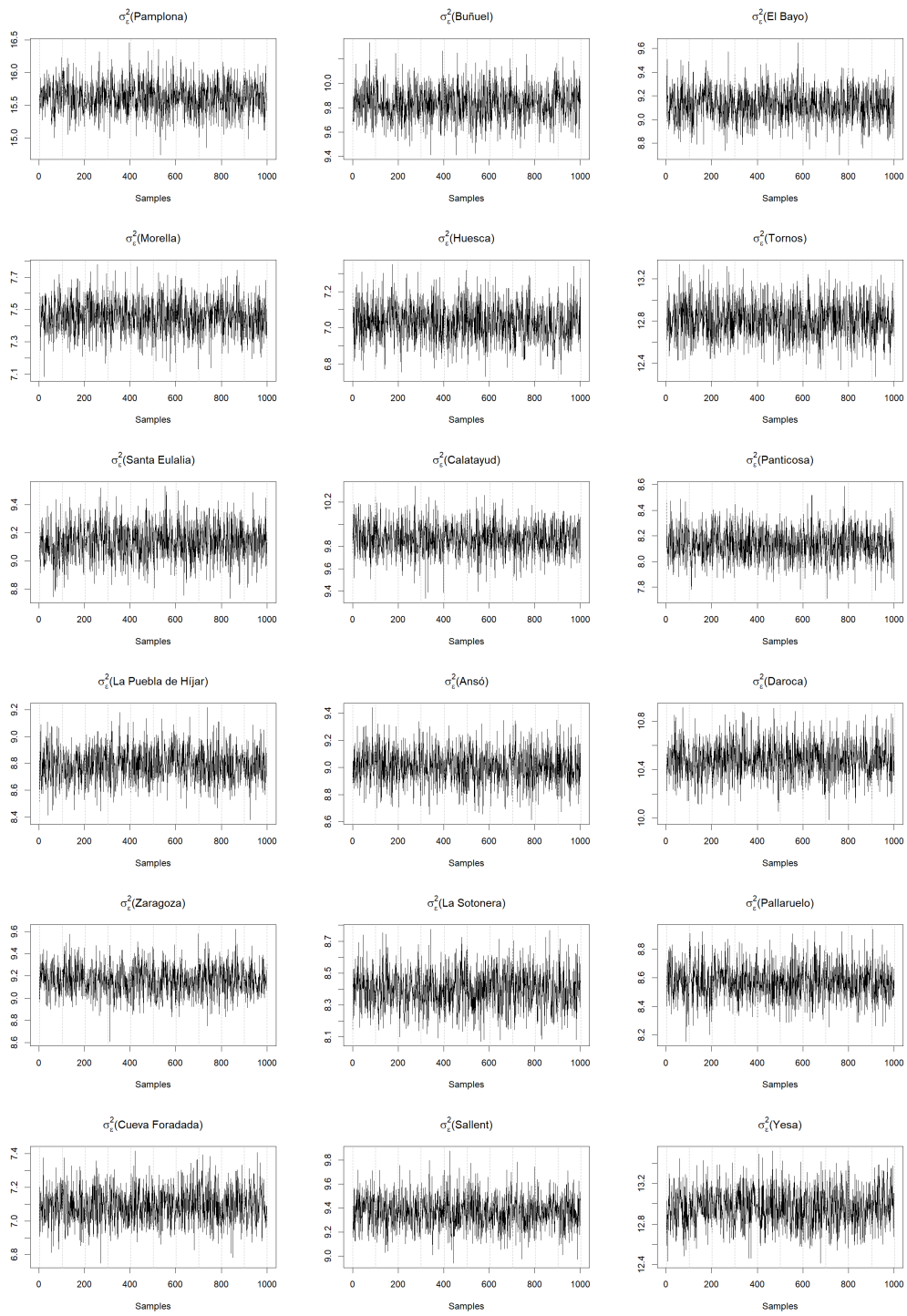
Figure S3: Trace plots for the parameters (3 of 5).

13

(d) $\rho_Y(\mathbf{s})$

Figure S3: Trace plots for the parameters (4 of 5).

(e) $\sigma_\epsilon^2(\mathbf{s})$

Figure S3: Trace plots for the parameters (5 of 5).

S3 Results

S3.1 Leave-one-out cross-validation

The performance of the models is compared using the approach in Section 3.4 of the Main Manuscript based on LOOCV for the 18 locations available. Table S4 summarizes for each site the four considered metrics.

S3.2 Comparison between local models and the full model

Figure S4 shows a comparison between the posterior distributions at the observed locations of the spatial processes $\tilde{\alpha}(\mathbf{s}), \rho_Y(\mathbf{s})$, and $\sigma_\epsilon(\mathbf{s})$, in M_4 (black), and the posterior distribution of the corresponding parameters in the local models shown in Section S1.1 (red). The results for M_4 show a good agreement with results of the local models. This agreement shows that M_4 has no systematic bias in the estimation of the parameters related to time trends, autocorrelations, or variances.

Note that β_0 expresses the baseline in local models, but M_4 also includes the term associated with elevation in the fixed effects, then the comparison of the posterior distribution of the intercept in local models and $\tilde{\beta}_0(\mathbf{s})$ is not of interest.

S3.3 Prediction at unobserved locations

Figure S5 shows for each unobserved location (Longares, Olite and Guara) the posterior densities of the four spatial processes in M_4. Figure S6 shows the posterior difference between average temperatures in both 30-year periods, 1956–1985 and 1986–2015.

Table S4: Value of the performance metrics for models with different spatial GPs for each location.

Model	Pamplona				Buñuel				El Bayo			
	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG
M_0	5.97	4.89	3.51	0.74	4.25	3.42	2.41	0.92	4.13	3.34	2.35	0.93
$M_1(\beta_0(\mathbf{s}))$	5.74	4.72	3.36	0.77	4.21	3.40	2.39	0.92	4.16	3.39	2.37	0.92
$M_1(\alpha(\mathbf{s}))$	5.97	4.89	3.51	0.74	4.25	3.43	2.41	0.92	4.12	3.33	2.34	0.93
$M_1(\rho_Y(\mathbf{s}))$	5.96	4.88	3.51	0.74	4.24	3.42	2.41	0.92	4.13	3.34	2.34	0.92
$M_1(\sigma_\epsilon(\mathbf{s}))$	5.97	4.89	3.48	0.77	4.26	3.43	2.42	0.92	4.13	3.34	2.35	0.94
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	5.77	4.74	3.35	0.79	4.21	3.40	2.39	0.92	4.17	3.39	2.37	0.94
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	5.76	4.73	3.35	0.80	4.20	3.40	2.39	0.92	4.15	3.39	2.37	0.94
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	5.95	4.87	3.48	0.77	4.25	3.43	2.41	0.91	4.12	3.33	2.34	0.93
M_4	5.75	4.73	3.36	0.79	4.20	3.40	2.38	0.91	4.15	3.39	2.36	0.93

Model	Morella				Huesca				Tornos			
	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG
M_0	3.89	3.05	2.20	0.94	3.75	3.01	2.14	0.95	5.92	4.95	3.50	0.75
$M_1(\beta_0(\mathbf{s}))$	3.89	3.06	2.21	0.95	3.75	3.00	2.14	0.95	5.23	4.34	3.04	0.81
$M_1(\alpha(\mathbf{s}))$	3.88	3.05	2.20	0.94	3.76	3.02	2.15	0.95	5.92	4.95	3.50	0.75
$M_1(\rho_Y(\mathbf{s}))$	3.89	3.05	2.20	0.94	3.76	3.02	2.15	0.96	5.92	4.96	3.50	0.76
$M_1(\sigma_\epsilon(\mathbf{s}))$	3.88	3.05	2.19	0.92	3.75	3.01	2.13	0.94	5.92	4.95	3.49	0.77
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	3.87	3.05	2.19	0.94	3.75	3.01	2.13	0.93	5.22	4.33	3.03	0.83
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	3.91	3.07	2.21	0.93	3.75	3.01	2.13	0.93	5.23	4.34	3.03	0.83
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	3.88	3.05	2.19	0.92	3.76	3.02	2.13	0.94	5.92	4.96	3.49	0.78
M_4	3.90	3.06	2.20	0.94	3.75	3.01	2.13	0.93	5.23	4.34	3.03	0.83

Model	Santa Eulalia				Calatayud				Panticosa			
	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG
M_0	4.93	4.11	2.84	0.87	4.53	3.71	2.59	0.90	5.15	4.05	2.91	0.83
$M_1(\beta_0(\mathbf{s}))$	4.38	3.56	2.49	0.91	4.42	3.60	2.52	0.90	4.44	3.49	2.50	0.88
$M_1(\alpha(\mathbf{s}))$	4.94	4.11	2.85	0.87	4.55	3.72	2.59	0.90	5.15	4.05	2.91	0.83
$M_1(\rho_Y(\mathbf{s}))$	4.93	4.10	2.84	0.87	4.54	3.71	2.59	0.89	5.20	4.09	2.94	0.85
$M_1(\sigma_\epsilon(\mathbf{s}))$	4.96	4.13	2.85	0.88	4.53	3.71	2.58	0.91	5.06	3.97	2.86	0.83
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.38	3.56	2.49	0.92	4.43	3.60	2.52	0.92	4.44	3.49	2.50	0.88
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.38	3.57	2.49	0.92	4.45	3.61	2.53	0.91	4.48	3.51	2.52	0.87
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.96	4.13	2.85	0.89	4.55	3.72	2.60	0.91	5.09	4.00	2.88	0.84
M_4	4.38	3.57	2.49	0.92	4.45	3.62	2.53	0.91	4.46	3.50	2.51	0.89

Model	La Puebla de Híjar				Ansó				Daroca			
	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG
M_0	4.01	3.24	2.28	0.94	4.56	3.63	2.57	0.89	4.61	3.83	2.65	0.90
$M_1(\beta_0(\mathbf{s}))$	4.02	3.24	2.28	0.93	4.44	3.58	2.52	0.90	4.50	3.64	2.56	0.88
$M_1(\alpha(\mathbf{s}))$	4.01	3.23	2.28	0.94	4.55	3.63	2.57	0.89	4.61	3.82	2.65	0.89
$M_1(\rho_Y(\mathbf{s}))$	4.02	3.25	2.28	0.94	4.56	3.63	2.58	0.89	4.61	3.83	2.65	0.90
$M_1(\sigma_\epsilon(\mathbf{s}))$	4.00	3.23	2.26	0.91	4.56	3.63	2.58	0.91	4.61	3.83	2.64	0.93
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.03	3.25	2.28	0.90	4.45	3.58	2.52	0.92	4.49	3.64	2.55	0.92
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.01	3.23	2.27	0.90	4.43	3.58	2.51	0.92	4.50	3.64	2.55	0.92
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.01	3.23	2.27	0.90	4.54	3.62	2.57	0.91	4.61	3.83	2.64	0.93
M_4	4.03	3.25	2.28	0.90	4.43	3.58	2.51	0.92	4.50	3.64	2.55	0.92

Model	Zaragoza				La Sotonera				Pallaruelo			
	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG
M_0	4.03	3.28	2.29	0.94	4.00	3.21	2.27	0.94	4.09	3.30	2.33	0.93
$M_1(\beta_0(\mathbf{s}))$	4.04	3.28	2.30	0.93	4.00	3.20	2.27	0.93	4.04	3.24	2.29	0.93
$M_1(\alpha(\mathbf{s}))$	4.02	3.27	2.30	0.94	4.01	3.21	2.28	0.94	4.11	3.30	2.33	0.93
$M_1(\rho_Y(\mathbf{s}))$	4.02	3.28	2.29	0.94	4.00	3.21	2.28	0.94	4.10	3.31	2.33	0.93
$M_1(\sigma_\epsilon(\mathbf{s}))$	4.03	3.28	2.29	0.93	4.00	3.21	2.27	0.91	4.08	3.29	2.31	0.90
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.04	3.28	2.30	0.93	4.00	3.20	2.26	0.90	4.04	3.24	2.29	0.91
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.04	3.28	2.30	0.92	4.00	3.21	2.26	0.90	4.06	3.25	2.29	0.90
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	4.02	3.27	2.29	0.92	4.00	3.21	2.27	0.91	4.10	3.30	2.32	0.90
M_4	4.04	3.28	2.30	0.92	4.00	3.21	2.27	0.91	4.06	3.25	2.30	0.90

| Model	Cueva Foradada				Sallent				Yesa			
	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG	RMSE	MAE	CRPS	CVG
M_0	3.69	2.94	2.11	0.95	4.47	3.56	2.52	0.90	4.85	3.96	2.78	0.86
$M_1(\beta_0(\mathbf{s}))$	3.85	3.05	2.18	0.94	4.48	3.67	2.56	0.89	4.86	4.03	2.80	0.86
$M_1(\alpha(\mathbf{s}))$	3.68	2.93	2.10	0.96	4.47	3.56	2.52	0.90	4.83	3.95	2.76	0.87
$M_1(\rho_Y(\mathbf{s}))$	3.69	2.94	2.11	0.95	4.49	3.57	2.53	0.92	4.85	3.96	2.78	0.86
$M_1(\sigma_\epsilon(\mathbf{s}))$	3.69	2.94	2.10	0.95	4.46	3.56	2.52	0.87	4.86	3.96	2.77	0.89
$M_2(\beta_0(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	3.85	3.05	2.18	0.93	4.48	3.67	2.57	0.86	4.87	4.03	2.79	0.88
$M_3(\beta_0(\mathbf{s}), \alpha(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	3.84	3.04	2.17	0.93	4.48	3.67	2.56	0.86	4.84	4.02	2.78	0.89
$M_3(\alpha(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s}))$	3.68	2.93	2.09	0.95	4.46	3.56	2.52	0.89	4.83	3.95	2.76	0.89
M_4	3.84	3.03	2.17	0.93	4.48	3.66	2.55	0.88	4.85	4.03	2.79	0.88
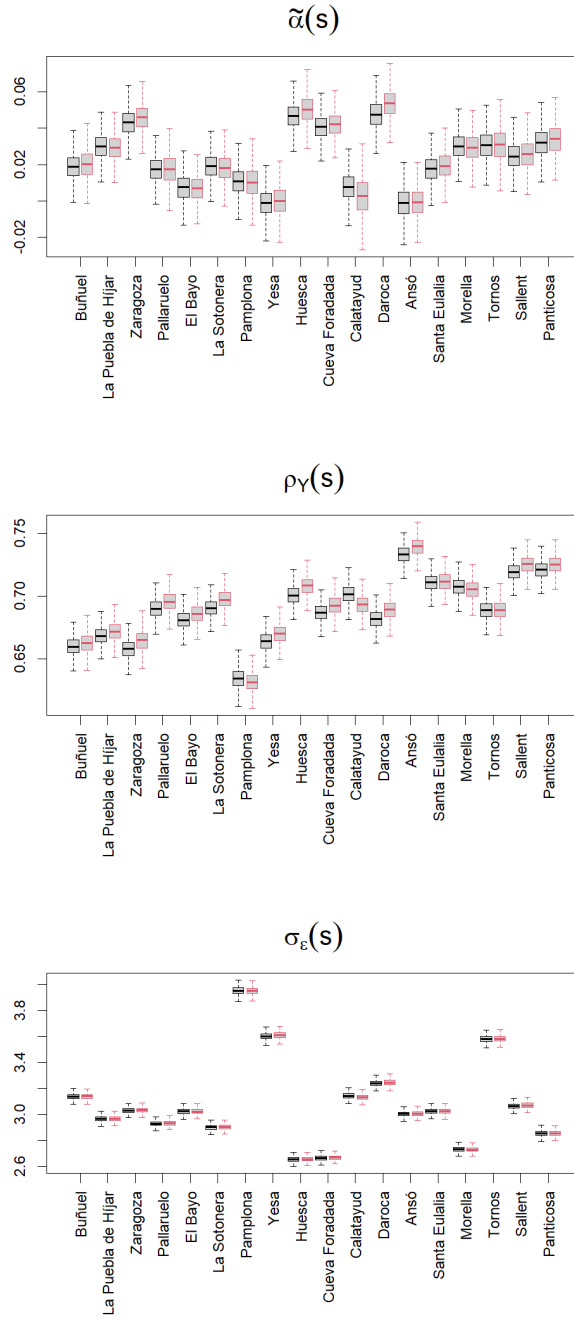
Figure S4: Boxplots of the posterior distributions of the spatial random effects in M_4 (black) and of the corresponding coefficients in the local models (red). Locations are sorted by elevation, from lowest to highest.
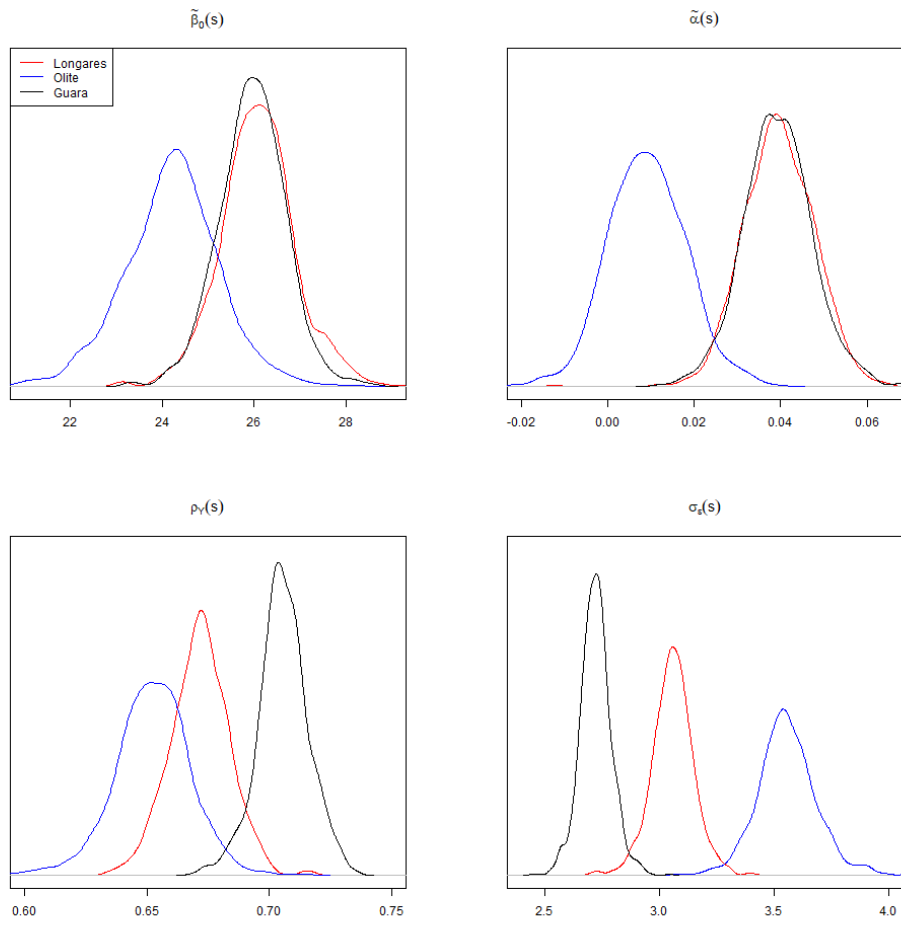
Figure S5: Posterior densities of the spatial random effects, $\tilde{\beta}_0(\mathbf{s}), \tilde{\alpha}(\mathbf{s}), \rho_Y(\mathbf{s}), \sigma_\epsilon(\mathbf{s})$, at the unobserved locations Longares, Olite and Guara.
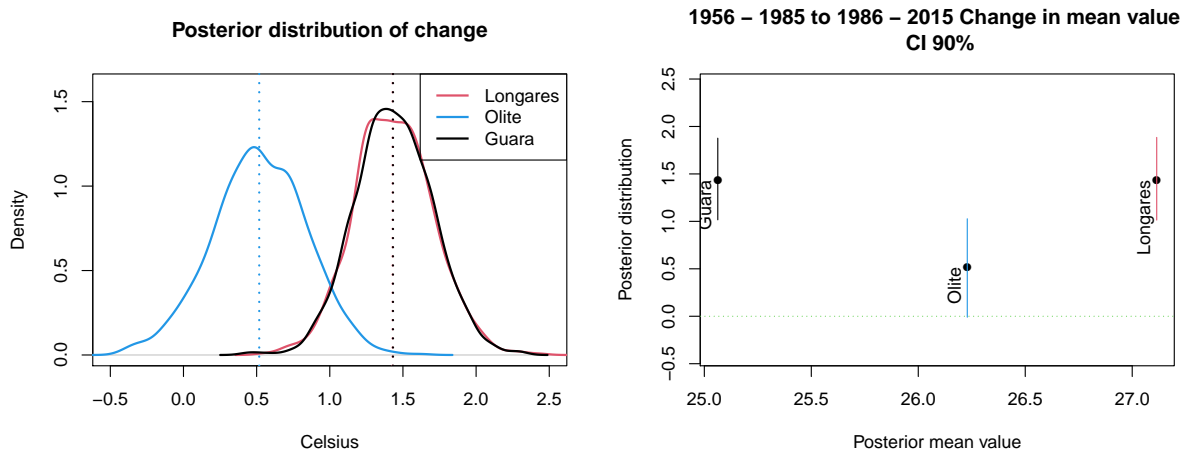
Figure S6: Left: Posterior densities of the difference between the mean value of the daily maximum temperature between both 30-year periods, 1956–1985 and 1986–2015, at the unobserved locations Longares, Olite and Guara. Right: Posterior mean and 90% credible intervals of the previous difference against posterior mean value of the entire period.

References

Gelfand AE (2012) Hierarchical modeling for spatial data problems. Spatial Statistics 1(1):30–39

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7(4):457–472

Gelman AG, Roberts GO, Gilks WR (1996) Efficient Metropolis jumping rules. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian Statistics 5, Oxford University Press, pp 599–607

Gong L, Flegal JM (2016) A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 25(3):684–700

3.2 Spatio-temporal analysis of the extent of an extreme heat event

This manuscript was published in:

Cebrián, A. C., Asín, J., Gelfand, A. E., Schliep, E. M., **Castillo-Mateo, J.**, Beamonte, M. A., & Abaurrea, J. (2022). Spatio-temporal analysis of the extent of an extreme heat event. *Stochastic Environmental Research and Risk Assessment, 36*(9), 2737–2751. https://doi.org/10.1007/s00477-021-02157-z

And it was disseminated (speaker emphasized) in:

- *Gelfand, A. E.*, **Castillo-Mateo, J.**, Cebrián, A. C., Asín, J., Schliep, E. M., Beamonte, M. A., & Abaurrea, J. (2022, June 7–10). *Long term spatial modeling for exploring incidence of extreme heat events* [Plenary talk]. XXXIX Congreso Nacional de Estadística e Investigación Operativa, Granada, Spain. (See also in Section 3.3.)

"The best thing about being a statistician is that you get to play in everyone's backyard."

John W. Tukey

ORIGINAL PAPER

Spatio-temporal analysis of the extent of an extreme heat event

Ana C. Cebrián[1] · Jesús Asín[2] · Alan E. Gelfand[3] · Erin M. Schliep[4] · Jorge Castillo-Mateo[1] · María A. Beamonte[5] · Jesús Abaurrea[1]

Abstract
Evidence of global warming induced from the increasing concentration of greenhouse gases in the atmosphere suggests more frequent warm days and heat waves. The concept of an extreme heat event (EHE), defined locally based on exceedance of a suitable local threshold, enables us to capture the notion of a period of persistent extremely high temperatures. Modeling for extreme heat events is customarily implemented using time series of temperatures collected at a set of locations. Since spatial dependence is anticipated in the occurrence of EHE's, a joint model for the time series, incorporating spatial dependence is needed. Recent work by Schliep et al. (J R Stat Soc Ser A Stat Soc 184(3):1070–1092, 2021) develops a space-time model based on a point-referenced collection of temperature time series that enables the prediction of both the incidence and characteristics of EHE's occurring at any location in a study region. The contribution here is to introduce a formal definition of the notion of the spatial extent of an extreme heat event and then to employ output from the Schliep et al. (J R Stat Soc Ser A Stat Soc 184(3):1070–1092, 2021) modeling work to illustrate the notion. For a specified region and a given day, the definition takes the form of a block average of indicator functions over the region. Our risk assessment examines extents for the Comunidad Autónoma de Aragón in northeastern Spain. We calculate daily, seasonal and decadal averages of the extents for two subregions in this comunidad. We generalize our definition to capture extents of persistence of extreme heat and make comparisons across decades to reveal evidence of increasing extent over time.

Keywords Bayesian inference · Block average · Monte Carlo integration · Spatial autoregression · Stochastic integral

1 Introduction

There is strong evidence of global warming due to the increasing concentration of greenhouse gases in the atmosphere (Lai and Dzombak 2019). This global warming suggests more frequent warm days, more frequent and persistent heat waves (Lemonsu et al. 2014; Alexander 2016) as well as events that break previous records by much larger margins (Fischer et al. 2021; Cebrián et al. 2021). The analysis of heat waves is particularly important due to the potential for serious anthropogenic, environmental, and economic impacts (Amengual et al. 2014; Campbell et al. 2018). Extreme heat raises significant health concerns in humans as it can result in death, change the range or niche for plants and animals, and lead to heat-driven peaks in electricity demands or lost crop income.

A challenge in analyzing heat waves stems from on-going debate over its exact definition. According to the World Meteorological Office (WMO), a period persisting at least three consecutive days of marked unusual hot weather (maximum, minimum and daily average temperature) over a region with thermal conditions above given thresholds based on local climatological conditions can be considered a heat wave. This definition suggests that

✉ Ana C. Cebrián
 acebrian@unizar.es

1 Dpto. Métodos Estadísticos, Matemáticas, Universidad de Zaragoza, Zaragoza, Spain

2 Dpto. Métodos Estadísticos, EINA, Universidad de Zaragoza, Zaragoza, Spain

3 Department of Statistical Science, Duke University, Durham, NC, USA

4 Department of Statistics, University of Missouri, Columbia, MO, USA

5 Facultad de Economía y Empresa, Universidad de Zaragoza, Zaragoza, Spain

analyses of heat waves require temperature series at a daily scale, but offers no operational guidance with regard to the various choices of implementation. Khaliq et al. (2005) and Reich et al. (2014) used only maximum temperature, while Keellings and Waylen (2014) and, more recently, Abaurrea et al. (2018) considered both maximum and minimum temperatures.

While there is lack of agreement on heat wave definition in the literature (Perkins and Alexander 2013; Smith et al. 2013), the concept of an extreme heat event (EHE) is explicitly defined, capturing the notion of a period (number of consecutive days) of persistent extremely high temperatures. Specifically, EHE's are defined locally and are based on exceedance of a suitable local threshold. That is, useful thresholds should be based on local conditions; in the sequel, we adopt the 95th percentile of local daily maximum temperatures, derived using a ten year period. In the context of persistence of exceedance, it is evident that we need to model daily max temperatures since an EHE is defined at a given location as a run of consecutive daily temperature observations exceeding the *given* threshold for that location. Additionally, daily modeling enables assessment of important behaviors describing the nature of an EHE such as the duration, average exceedance, and maximum exceedance above the threshold. As a result, we define extent in terms of incidence of exceedance on a given day, developed from a daily max temperature model that is able to adequately represent both the central part and the upper tail of the temperature distribution. It is not a model only for the tails or observations above a threshold, i.e., left-truncated data (peaks over thresholds models) or for the extremes (generalized extreme value distribution models). Such modeling addresses different objectives.

Modeling for extreme heat events is customarily implemented using time series of temperatures over a window of time collected at particular locations. However, spatial aspects of this phenomenon are also of interest and should be introduced into the modeling process, particularly with interest in predicting EHE behavior at locations without available time series of temperatures. Since spatial dependence is anticipated in the occurrence of EHE's, a joint model for time series at different locations that incorporates spatial dependence is needed.

Recent work by Schliep et al. (2021) develops a space-time model based on a point-referenced collection of temperature time series that enables the prediction of both the incidence and the characteristics of the EHE's occurring at any location in the region. Specifically, it offers direct spatial modeling for daily maximum temperatures which can then be used to characterize the EHE's. The challenge is that, while the bulk of the distribution, i.e., where most of the data is observed, is not extreme, the main interest for the model lies in the upper tail when

attempting to characterize EHE's (Keellings and Waylen 2015; Shaby et al. 2016). To address this, a specification incorporating thresholding is introduced, i.e., a model which switches between two observed states, one that defines extreme heat days (those above the temperature threshold) and the other that defines non-extreme heat days (those below the temperature threshold). Again, these thresholds are obtained locally and assumed fixed. Importantly, this two-state structure allows temporal dependence of the observations but also permits the parameters controlling the effects of covariates and the spatial dependence to differ between the two states. We briefly review details of this modeling in Sect. 2.4.

With regard to risk assessment, the contribution of this work is to formalize the notion of the spatial "extent" of an extreme heat event and then to illustrate it using output from the Schliep et al. (2021) modeling work. Interest in characterizing the extent of heat waves is clear. For example, Lhotka and Kyselý (2015) proposed an extremity index that captures joint effects of spatial extent, temperature and duration of heat waves. Keellings and Moradkhani (2020) also developed a spatial metric combining heat wave frequency, magnitude, duration and areal extent to analyze the evolution of heat waves across the United States. Rebetez et al. (2009) compared the heatwave extent in Europe in 2003 and 2006. Khan et al. (2019) found an increase of 1.36% of the affected area having both maximum and minimum temperature above the 95th percentile per decade in Pakistan. Lyon et al. (2019) projected the increase in the spatial extent of contiguous US summerheat waves using the CMIP5 archive under RCP4.5 and RCP8.5 scenarios. They found a substantial increase in spatial extent climate model employing projections for 2031–2055. However, all this work analyzes the extent using descriptive approaches and using observed or gridded data with no formal inference from probabilistic modeling.

Some more formal definitions related to the concept of area under extreme conditions have been introduced in the statistical literature. For example, French and Sain (2013) present a method for constructing confidence regions for Gaussian processes that contain the true exceedance regions with some predefined probability. Extending this methodology, Hazra and Huser (2021) obtain confidence regions that contain joint threshold exceedances of surface sea temperatures in the Red Sea, using a semiparametric Bayesian spatial mixed-effects linear model. Bolin and Lindgren (2015) consider excursion sets, which are sets of points in an area where a spatial function is above a given threshold. Sommerfeld et al. (2018) develop confidence regions for these spatial excursion sets with an application to climate and Romero-Béjar et al. (2018) develop quantile-based spatiotemporal risk assessment of exceedances using this concept. Zhong et al. (2020) analyze spatial

extent of heatwaves using a model based on max-infinitely divisible processes for annual temperature maxima. However, to our knowledge, the notion of the extent of an EHE has not been explicitly defined as a stochastic object. So, we take this up both conceptually and practically. Given a threshold surface over a specified region, for a given day, the extent of an EHE over the region is the proportion of the area of the region whose daily max temperature for the day is above the associated threshold surface. In the sequel we adopt a static threshold surface in order to assess change in extent over time. With different intentions, time-varying thresholds could be employed.

The basic idea is an extension of the so-called spatial cumulative distribution function (cdf) following the work of Lahiri et al. (1999) and Short et al. (2005). It is also discussed in Section 15.3 of Banerjee et al. (2014). The extent arises as a stochastic integral, i.e., an integral over a realization of a stochastic process. It is a random object as well as a conceptual object in the sense that it can not be observed and it can not be calculated in closed form. However, its moment properties can be calculated and approximate realizations can be obtained through Monte Carlo integration.

Attractively, we can study extents employing realizations from any spatio-temporal model fitted for daily max temperatures. Here, we adopt the model from Schliep et al. (2021) and do not do any additional modeling work. We can provide extents directly from the output of that model fitting. In different words, assessment of extent of an EHE is a post model fitting activity.

The region over which we study extents is the Comunidad Autónoma de Aragón region in northeastern Spain, located in the Ebro basin (85,362 km^2). The Ebro river flows from the NW to the SE through a valley bordered by the Pyrenees and the Cantabrian Range in the north and the Iberian System in the southwest. The maximum elevation is approximately 3400 m in the Pyrenees, 2600 m in the Cantabrian Range and Iberian System, and between 200-400 m in the central valley. In general, the area is characterized by a Mediterranean-continental dry climate with irregular rainfall, and a large temperature range. However, several climate subareas can be distinguished due to the heterogeneous orography and other influences, such as the Mediterranean sea to the east, and the continental conditions of the Iberian central plateau in the southwest. Zaragoza, the largest city in the region, is located in the central part of the valley, and experiences more extreme temperatures and drier conditions.

Our data are observational series from AEMET (the Spanish Meteorological Office). Only long term series with limited missing observations were considered, resulting in daily maximum temperatures for 18 sites across and around

the Comunidad Autónoma de Aragón region for the years 1953-2015. The names and locations of the 18 sites are shown spatially in Fig. 1. Data from the years 1953-1962 were used to obtain the location-specific thresholds for extreme heat events; the data for the years 1963-2015 were used in the modeling. Specifically, thresholds were computed as the 95th percentile of daily maximum temperature for the months June, July, and August during the 10 years 1953-1962. Again, our objective is to see change in extent over time which requires a fixed $q(\mathbf{s})$ threshold surface. While we could use fixed thresholds associated with any time window, using thresholds prior to the start of our modeling, using data not employed in our fitting, seemed to provide a sensible baseline.

The format of the paper is as follows. In Sect. 2, we present the details of the spatial extent and briefly review the Schliep et al. (2021) model. Section 3 shows the rich scope of extents we can calculate and compare. Section 4 provides a brief summary and future work.

2 Formalizing extent of extreme heat

2.1 The definition of extent

The definition of the spatial cdf (Lahiri et al. 1999) is attached to a realization of a stochastic process taking values in $\mathbb{R}$, $\mathcal{Z} = \{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$, over a region $\mathcal{D}$ and, for a given w, is

$$F_{\mathcal{Z}}(w) = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \mathbf{1}(Z(\mathbf{s}) \le w) d\mathbf{s}$$

where $|\mathcal{D}|$ is the area of $\mathcal{D}$. That is, it is the proportion of the realization over $\mathcal{D}$ that lies below w. It behaves like a cdf in the sense that it is nondecreasing and goes to 0 as $w \to -\infty$ and 1 as $w \to \infty$. However, since it is a function of the process realization, it is a random variable and arises as a stochastic integral. It differs from the marginal cdf associated with the variable $Z(\mathbf{s})$ at location $\mathbf{s}$, $P(Z(\mathbf{s}) \le w)$ which is a constant (as a feature of the distribution of $Z(\mathbf{s})$). Like stochastic integrals in general, it can not be calculated explicitly (but it can be approximated using Monte Carlo integration). However, some distributional properties, e.g., mean and variance, can be calculated.

To define the object of interest here, a spatial extent, let $Y_t(\mathbf{s})$ denote the daily max temp for day t at location $\mathbf{s}$. Suppose we consider a subregion $B \subseteq \mathcal{D}$ of the study domain. Then, for a given w, the extent of the EHE in subregion B on day t is:

$$Ext_t(w; B) = \frac{1}{|B|} \int_{B} \mathbf{1}(Y_t(\mathbf{s}) - q(\mathbf{s}) \ge w) d\mathbf{s}. \tag{1}$$
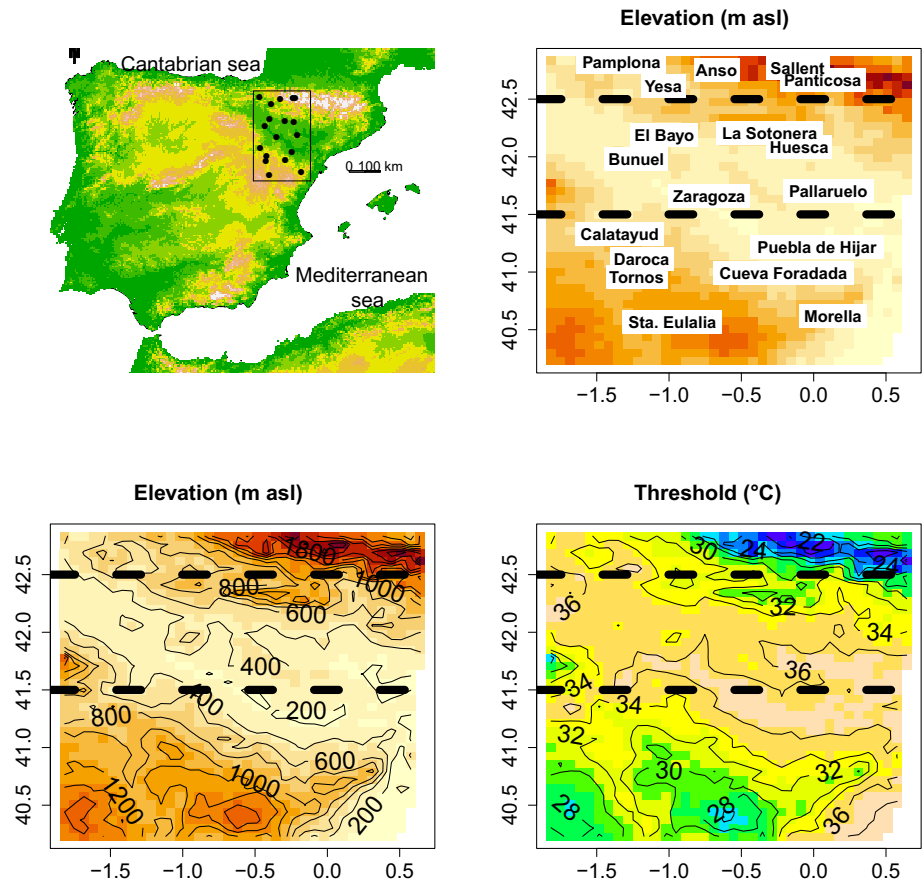
Here, $q(\mathbf{s})$ is the threshold surface over B. That is, it is the proportion of B which is experiencing extreme heat at least w degrees above or below (according to the sign of w) associated local thresholds on day t. In fact, it is referred to as a block average (Banerjee et al. 2014) in this case of indicator functions. Evidently, we can choose B as we wish. In Section 3 we consider two subregions of interest for comparison but, with interest in extent at a larger spatial scale, we might also consider the case where $B = \mathcal{D}$. Regardless, since extent captures proportion of incidence within a region, a larger region does not imply a larger extent. However, an alternative definition of extent would specify a region where extent is anticipated to be high (or low) and then calculate the extent in order to provide quantification. EHE is applicable to the extent when $w = 0$ and is probably of greatest interest but below, Sect. 3.4, we also look at the extent of more ($w > 0$) or less ($w < 0$) extreme heat events.

The posterior predictive distribution for $Ext_t(w; B)$ is needed for inference. We use a Monte Carlo integration to obtain an approximate realization of it by computing:

$$\widetilde{Ext_t}(w; B) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}(Y_t(\mathbf{s}_j) - q(\mathbf{s}_j) \geq w). \qquad (2)$$

Here, for a selected set of m locations in B with an associated set of thresholds, $\{q(\mathbf{s}_j), j = 1, 2, \ldots, m\}$, $\{Y_t(\mathbf{s}_j), j = 1, 2, \ldots, m\}$ is a posterior predictive realization of daily max temperatures for day t at the locations, $\mathbf{s}_j$. If we have a collection of these realizations, then we can obtain posterior samples of $\widetilde{Ext_t}(w; B)$ for any choices of w. In this way, with arbitrarily many posterior predictive realizations, we can learn arbitrarily well about the posterior predictive distribution for $Ext_t(w; B)$.

To compute (2), given t, we need a realization $\{Y_t(\mathbf{s}_j), j = 1, 2, \ldots, m\}$. To consider arbitrary days within arbitrary years within arbitrary decades, we need a posterior predictive realization of a 50 year daily max time series (1966-2015) for each $\mathbf{s}_j$. Employing the output of the model fitting of Schliep et al. (2021), we can obtain such posterior predictive realizations using composition sampling (Banerjee et al. 2014, Chapter 6). Below, we obtain a collection of 500 such realizations, enabling a posterior predictive distribution for $Ext_t(w; B)$ through $\widetilde{Ext_t}(w; B)$. These samples yield an empirical summary of the posterior predictive distribution for $Ext_t(w; B)$, hence, posterior inference regarding any features of $Ext_t(w; B)$ that are of interest.

Note that to compute (2) we first need to create the $q(\mathbf{s}_j)$. In the sequel, upon a regular gridding of B to fairly fine resolution, we obtained the centroids of the grid cells as our $\mathbf{s}_j$'s. Then, gathering elevations for these centroids from a digital terrain map (DTM), we kriged the $q(\mathbf{s}_j)$. That is, using the thresholds for the observed sites with their associated elevations, we employed a standard kriging model for the thresholds at the $\mathbf{s}_j$ using their associated DTM elevations. Furthermore, we identified an additional 37 stations which had temperature data available between 1953-1962, yielding a total of 55 stations from which we developed the $q(\mathbf{s})$ surface. In Figure OR.1 in the Online Resource we show the resulting $q(\mathbf{s})$ surface along with the 55 stations. Fancier kriging could be imagined but here and illustratively, we confine ourselves to the above. Furthermore, we treat all of the kriged $q(\mathbf{s}_j)$ as fixed in implementing the Monte Carlo integrations.

As a last remark, it is possible to create an *empirical* extent, employing the form in (2) but using only the available observed sites that are within B. With only 18 stations, only a few will be in B, yielding a single estimate that can assume only a few discrete values and with no uncertainty. In implementing the Monte Carlo integrations for (2) below, we employ $m \approx 6000 - 8000$, yielding much smoother extents and with replication enabling us to see the distribution.

2.2 Some technical details: a digression

Here, we offer some theoretical insight into the distribution of a spatial extent by calculating the first and second moments under an illustrative Gaussian spatial first order autoregression model, a simplified version of the model that we work with in Sect. 2.4. In particular, consider the model

$$Y_t(\mathbf{s}) = \mu_t(\mathbf{s}) + \eta(\mathbf{s}) + \rho(Y_{t-1}(\mathbf{s}) - (\mu_{t-1}(\mathbf{s}) + \eta(\mathbf{s}))) + \epsilon_t(\mathbf{s})$$

where $\mu_t(\mathbf{s})$ is a spatio-temporal drift term. Specific choices are adopted in Sect. 2.4. Here, $\eta(\mathbf{s})$ is a mean 0 Gaussian process with covariance $cov(\eta(\mathbf{s}), \eta(\mathbf{s}')) = \sigma^2 h(\mathbf{s} - \mathbf{s}')$ providing local spatial adjustment to the drift terms as well as spatial dependence across locations. The $\epsilon_t(\mathbf{s})$ are pure errors, independent and identically distributed as $N(0, \tau^2)$.

Let $Z_t(\mathbf{s}) = Y_t(\mathbf{s}) - \mu_t(\mathbf{s})$ so $Z_t(\mathbf{s}) = \rho Z_{t-1}(\mathbf{s}) + (1 - \rho)\eta(\mathbf{s}) + \epsilon_t(\mathbf{s})$. Marginalizing over $\eta(\mathbf{s})$, we obtain

$$Z_t(\mathbf{s}) \mid Z_{t-1}(\mathbf{s}) \sim N(\rho Z_{t-1}(\mathbf{s}), (1 - \rho)^2 \sigma^2 + \tau^2).$$

In fact, the joint distribution of $(Z_t(\mathbf{s}), Z_t(\mathbf{s}'))$ given $(Z_{t-1}(\mathbf{s}), Z_{t-1}(\mathbf{s}'))$ is bivariate normal with mean $\begin{pmatrix} \rho Z_{t-1}(\mathbf{s}) \\ \rho Z_{t-1}(\mathbf{s}') \end{pmatrix}$ and covariance matrix

$$\begin{pmatrix} (1 - \rho)^2 \sigma^2 + \tau^2 & (1 - \rho)^2 \sigma^2 h(\mathbf{s} - \mathbf{s}') \\ (1 - \rho)^2 \sigma^2 h(\mathbf{s} - \mathbf{s}') & (1 - \rho)^2 \sigma^2 + \tau^2 \end{pmatrix}.$$

Implementing the customary marginalization over $Z_{t-1}(\mathbf{s})$, at *equilibrium* (t large), we have $Z_t(\mathbf{s}) \sim N(0, \phi^2(\rho, \sigma^2, \tau^2))$ where $\phi^2(\rho, \sigma^2, \tau^2) = \frac{(1-\rho)^2 \sigma^2 + \tau^2}{1 - \rho^2}$. So, $Y_t(\mathbf{s}) \sim N(\mu_t(\mathbf{s}), \phi^2(\rho, \sigma^2, \tau^2))$. Similarly, we can obtain $cov(Z_t(\mathbf{s}), Z_t(\mathbf{s}')) = \frac{(1-\rho)^2 \sigma^2 h(\mathbf{s}-\mathbf{s}')}{1-\rho^2}$ and hence, the distribution, $[Y_t(\mathbf{s}), Y_t(\mathbf{s}')]$.

Returning to (1), $E(Ext_t(w; B)) =$

$$E\left(\frac{1}{|B|} \int_B \mathbf{1}(Y_t(\mathbf{s}) - q(\mathbf{s}) \geq w) d\mathbf{s}\right)$$
$$= \frac{1}{|B|} \int_B E(\mathbf{1}(Y_t(\mathbf{s}) - q(\mathbf{s}) \geq w)) d\mathbf{s}.$$

However,

$$E(\mathbf{1}(Y_t(\mathbf{s}) - q(\mathbf{s}) \geq w)) = P(Y_t(\mathbf{s}) \geq q(\mathbf{s}) + w) \equiv p_t(\mathbf{s}, w)$$
$$= \Phi\left(\frac{\mu_t(\mathbf{s}) - (q(\mathbf{s}) + w)}{\phi(\rho, \sigma^2, \tau^2)}\right).$$

So,

$$E(Ext_t(w; B)) = \frac{1}{|B|} \int_B \Phi\left(\frac{\mu_t(\mathbf{s}) - (q(\mathbf{s}) + w)}{\phi(\rho, \sigma^2, \tau^2)}\right) d\mathbf{s}. \qquad (3)$$

Monte Carlo approximation to (3) is straightforward.

Next, we obtain $var(Ext_t(w; B))$. Let $g(Y_t(\mathbf{s})) = \mathbf{1}(Y_t(\mathbf{s}) - q(\mathbf{s}) \geq w)$. Then, by familiar calculation for block averages (see Banerjee et al. 2014, Chapter 7),

$$var(Ext_t(w; B)) = \frac{1}{|B|} \int_B var(g(Y_t(\mathbf{s}))) d\mathbf{s}$$
$$+ \frac{1}{|B|^2} \int_B \int_B cov(g(Y_t(\mathbf{s})), g(Y_t(\mathbf{s}'))) d\mathbf{s}' d\mathbf{s}.$$

However, $var(g(Y_t(\mathbf{s}))) = p_t(\mathbf{s}, w)(1 - p_t(\mathbf{s}, w))$. Similarly,

$$cov(g(Y_t(\mathbf{s})), g(Y_t(\mathbf{s}'))) = p_t(\mathbf{s}, \mathbf{s}', w) - p_t(\mathbf{s}, w)p_t(\mathbf{s}', w)$$

where

$$p_t(\mathbf{s}, \mathbf{s}', w) = P(Y_t(\mathbf{s}) \geq q(\mathbf{s}) + w, Y_t(\mathbf{s}') \geq q(\mathbf{s}') + w)$$

, a double integral over the bivariate normal distribution for $Y_t(\mathbf{s})$ and $Y_t(\mathbf{s}')$ given above. So,

$$var(Ext_t(w; B)) = \frac{1}{|B|} \int_B p_t(\mathbf{s}, w)(1 - p_t(\mathbf{s}, w)) d\mathbf{s}$$
$$+ \frac{1}{|B|^2} \int_B \int_B p_t(\mathbf{s}, \mathbf{s}', w) - p_t(\mathbf{s}, w)p_t(\mathbf{s}', w) d\mathbf{s}' d\mathbf{s}. \qquad (4)$$

Monte Carlo integration for (4) can also be implemented.

2.3 Elaborating extents

In climate applications, such as attempting to assess the effect of global warming on extreme temperatures, the main interest is not usually to characterize the distribution of the extent in a given day $Ext_t(w; B)$ but rather to characterize the behavior of the extent across years or changes in its seasonal pattern. To this end, it is convenient to analyze the distribution of different averages of the daily extent over a given period of time, such as seasons or years. In defining average extents, we will use two time indexes, one for the day within year l and one for the year t,

$$Ext_{t,l}(w; B) = \frac{1}{|B|} \int_B \mathbf{1}(Y_{t,l}(\mathbf{s}) - q(\mathbf{s}) \geq w)d\mathbf{s}.$$

Then, we can average the daily extent over a period of days within the year, L, and over a period of years, T, as

$$Av_{t \in T, l \in L}Ext_{t,l}(w; B) = \frac{1}{n_T n_L} \sum_{t \in T} \sum_{l \in L} Ext_{t,l}(w; B)$$

where n_T and n_L are the number of observations in the periods T and L respectively.

In the analysis of EHE, we consider: (i) decadal averages for a given day l and decade D, $Av_{t \in D}Ext_{t,l}(w; B) = \frac{1}{10} \sum_{t \in D} Ext_{t,l}(w; B)$, (ii) the average extent over the summer months JJA for a given year t, $Av_{l \in JJA}Ext_{t,l}(w; B) = \frac{1}{92} \sum_{l \in JJA} Ext_{t,l}(w; B)$, and (iii) the average extent over the summer months and a decade, $Av_{t \in D, l \in JJA}Ext_{t,l}(w; B) = \frac{1}{920} \sum_{t \in D} \sum_{l \in JJA} Ext_{t,l}(w; B)$. In calculating these quantities, we replace all integrals by their Monte Carlo approximations, obtaining the values $Av_{t \in T, l \in L}\widetilde{Ext}_{t,l}(w; B)$. Further, using 500 posterior predictive samples of realizations of $\widetilde{Ext}_{t,l}(w; B)$, we obtain 500 samples of $Av_{t \in T, l \in L}\widetilde{Ext}_{t,l}(w; B)$ to supply an empirical summary of the posterior distribution of $Av_{t \in T, l \in L}Ext_{t,l}(w; B)$. In the cases where the average is carried out only over one time index, a bigger size sample is obtained if the distribution is the same in a given period of time. For example, if the distribution of $Av_{l \in JJA}Ext_{t,l}(w; B)$ is taken to be the same for all the years in a decade, $t \in D$, we can obtain a sample of 5000 realizations of $Av_{l \in JJA}\widetilde{Ext}_{t,l}(w; B)$, 500 for each of the 10 years $t \in D$, to consider empirically the posterior distribution of $Av_{l \in JJA}Ext_{t,l}(w; B)$ in D. When it is assumed that the distribution of the ten years in D is the same, we modify notation to $Av_{l \in JJA}^D Ext_{t,l}(w; B)$, and $Av_{l \in JJA}^D \widetilde{Ext}_{t,l}(w; B)$.

Persistence: In the context of global warming, a relevant feature in the analysis of extreme temperatures is persistence. We consider persistence as arising when the probability of being in an extreme heat state at day $l + 1$ is higher if we were in an extreme heat state at day l than if

we were not. To analyze this feature spatially through the extent, we consider the proportion of B at w degrees above threshold, for both of two consecutive days, l and $l + 1$, denoted by $^2Ext_{t,l}(w; B)$ and defined as

$$^2Ext_{t,l}(w; B) = \frac{1}{|B|} \int_B \mathbf{1}(Y_{t,l}(\mathbf{s}) - q(\mathbf{s}) \geq w)$$
$$\mathbf{1}(Y_{t,l+1}(\mathbf{s}) - q(\mathbf{s}) \geq w)d\mathbf{s}.$$

More generally, $^rExt_{t,l}(w; B)$ denotes an r-day EHE, that is an EHE persisting for r consecutive days starting at day l, with analogous definition. It is immediate that

$$^rExt_{t,l}(w; B) \leq {}^{r-1}Ext_{t,l}(w; B) \leq \ldots \leq Ext_{t,l}(w; B).$$

This accords with the fact that the extents of two-day EHE's will be smaller than the extents of one-day EHE's. Again, we will use a Monte Carlo integration to obtain a realization, e.g., for the $r = 2$ case:

$$^2\widetilde{Ext}_{t,l}(w; B) = \frac{1}{m} \sum_{j=1}^m \mathbf{1}(Y_{t,l}(\mathbf{s}) - q(\mathbf{s}) \geq w) \, \mathbf{1}(Y_{t,l+1}(\mathbf{s}) - q(\mathbf{s}) \geq w).$$

In the analysis for our dataset/study area we confine ourselves to $r = 1$ and $r = 2$ since the probability of observing runs with $r \geq 3$ is too small to show useful extents.

As with the daily extents, we are usually more interested in the average of $^rExt_{t,l}(w; B)$ over a time window of l's, t's, or both. These averages are defined as above but with $^2Ext_{t,l}(w; B)$. Monte Carlo integrations for these averages are analogous to those for averages of $Ext_{t,l}(w; B)$; we only have to substitute $^rExt_{t,l}(w; B)$ with $^r\widetilde{Ext}_{t,l}(w; B)$.

Useful displays: Displays we will develop for extent and persistence include the following. First, we will consider different subregions of Aragón in order to make comparison between regions. To do this we will examine evolution of extent or persistence across the JJA season. We do this averaged over a decade in order to enable decadal comparison. Further, with posterior predictive samples of extents for each day l within each year t, we will supply the entire posterior predictive distribution of some of these extents and persistences. We will also develop "time to" displays, showing the time to the first day in year t with extent or persistence greater than or equal to a specified v. Lastly, we will examine how extents and persistences vary over choices of w since there can be interest in different specification of local thresholds for extreme heat.

2.4 Reviewing the daily max temperature model

Returning to the notation at the start of this section, let $Y_t(\mathbf{s})$ denote the daily maximum temperature at day t and location $\mathbf{s}$. Schliep et al. (2021) propose a two-state model

where the state for a given day defines whether or not the location is experiencing an extreme heat event. Given a threshold, $q(\mathbf{s})$ at location $\mathbf{s}$, let $U_t(\mathbf{s}) \in \{0, 1\}$ denote the state at time t for location $\mathbf{s}$ where a value of 0 denotes the below threshold state and a 1 denotes the above threshold state. So, $U_t(\mathbf{s})$ is a spatial binary time series process reflecting times of transition or state-switching. It is observed for each t at a monitored site but is latent elsewhere. With regard to extents, we note that $Ext_t(0; B) = \frac{1}{|B|} \int_B U_t(\mathbf{s}) d\mathbf{s}$.

$U_t(\mathbf{s})$ is a Markov process where the state $U_t(\mathbf{s})$ depends only on the previous state $U_{t-1}(\mathbf{s})$. Then, the distribution of $Y_t(\mathbf{s})$ is specified explicitly given $U_t(\mathbf{s})$ and, given the threshold, $U_t(\mathbf{s})$ is a binary function of $Y_t(\mathbf{s})$. Furthermore, the transition probabilities between states are allowed to be a function of previous temperature. This specification ensures transition probabilities to be "local", i.e., to vary with location and to depend upon the previous day's maximum temperature at that location. The opposite would be expected if the maximum temperature of the previous day resulted in a non-extreme heat state.

The joint distribution for temperature and state is specified in a first order Markov fashion explicitly as follows. Given $Y_{t-2}(\mathbf{s})$, and thus, $U_{t-2}(\mathbf{s})$, for two consecutive time points, $t-1$ and t, the joint distribution $[U_{t-1}(\mathbf{s}), Y_{t-1}(\mathbf{s}), U_t(\mathbf{s}), Y_t(\mathbf{s})]$ is written as

$$[Y_t(\mathbf{s})|U_t(\mathbf{s}), Y_{t-1}(\mathbf{s})][U_t(\mathbf{s})|Y_{t-1}(\mathbf{s})][Y_{t-1}(\mathbf{s})|U_{t-1}(\mathbf{s}), Y_{t-2}(\mathbf{s})]$$
$$[U_{t-1}(\mathbf{s})|Y_{t-2}(\mathbf{s})].$$

This formulation requires three model specifications: (i) $[Y_t(\mathbf{s})|U_t(\mathbf{s}) = 0, Y_{t-1}(\mathbf{s})]$, (ii) $[Y_t(\mathbf{s})|U_t(\mathbf{s}) = 1, Y_{t-1}(\mathbf{s})]$, and (iii) $[U_t(\mathbf{s})|Y_{t-1}(\mathbf{s})]$. With $q(\mathbf{s})$ denoting the threshold (quantile) for location $\mathbf{s}$, truncated distributions are needed for (i) and (ii), i.e., $[Y_t(\mathbf{s}) = y]\mathbf{1}(y < q(\mathbf{s}))$ and $[Y_t(\mathbf{s}) = y]\mathbf{1}(y \geq q(\mathbf{s}))$, respectively. For (i), a truncated normal distribution is adopted, with autoregressive centering,

$$TN\big(\mu_t^0(\mathbf{s}) - \rho^0(Y_{t-1}(\mathbf{s}) - \mu_{t-1}^0(\mathbf{s})), \sigma^{2,0}(\mathbf{s})\big)\mathrm{I}(-\infty, q(\mathbf{s})).$$

Details for $\mu_t^0(\mathbf{s})$ are given below.

For (ii), a truncated t-distribution is adopted, with autoregressive centering,

$$Tt\big(\mu_t^1(\mathbf{s}) - \rho^1(Y_{t-1}(\mathbf{s}) - \mu_{t-1}^1(\mathbf{s})), \sigma^{2,1}(\mathbf{s})\big)\mathrm{I}(q(\mathbf{s}), \infty).$$

Details for $\mu_t^1(\mathbf{s})$ are given below. As an aside, unlike the multivariate normal, the multivariate t-distribution captures upper tail extreme dependence (Chan and Li 2008) which may be desirable in looking at extreme heat events. Further, with multivariate t-distributions in both time and space, tail dependence is inherited in space as well.

Exploratory analysis of the observed data suggested a smaller variance for the above threshold daily maximum temperature distribution than for the below threshold daily maximum temperature distribution. Further, spatially-varying variances are introduced, expecting that variation in say, Jaca (in the Pyrenees in the north of the region) would be different from variation in say, Zaragoza (flat and central in the region).

For (iii) a probit link is employed to define

$$\Phi^{-1}(p_t(\mathbf{s})) \equiv \Phi^{-1}(P(U_t(\mathbf{s}) = 1|Y_{t-1}(\mathbf{s}))) \equiv \eta_t(\mathbf{s})$$

with $\eta_t(\mathbf{s})$ given below. Putting (i), (ii), and (iii) together, a mixture distribution for $Y_t(\mathbf{s})$ results: (i) a truncated normal distribution for the bulk of the distribution, (ii) a truncated t-distribution for the upper tail of the distribution, and (iii) mixture weights according to $P(U_t(\mathbf{s}) = 0)$ or $P(U_t(\mathbf{s}) = 1)$, respectively.

As for the specifics of $\mu_t(\mathbf{s})$ and $\eta_t(\mathbf{s})$, compacting notation, for $\mu_t^{U_t(\mathbf{s})}(\mathbf{s})$, Schliep et al. (2021) consider

$$\mu_t^{U_t(\mathbf{s})}(\mathbf{s}) = \beta_0^{U_t(\mathbf{s})} + \beta_0^{U_t(\mathbf{s})}(\mathbf{s}) + \gamma_{[\frac{t}{365}]+1}^{U_t(\mathbf{s})} + \beta_1^{U_t(\mathbf{s})}\mathrm{elev}(\mathbf{s})$$
$$+ \beta_2^{U_t(\mathbf{s})}\mathrm{lat}(\mathbf{s}) + \lambda_1\sin(2\pi t/365) + \lambda_2\cos(2\pi t/365).$$

Here, $\beta_0^{U_t(\mathbf{s})}$ denotes a global (across the domain for our dataset) intercept and $\beta_0^{U_t(\mathbf{s})}(\mathbf{s})$ denotes a local spatial intercept, i.e., providing local adjustment to the global intercept. Each $\beta_0^{U_t(\mathbf{s})}(\mathbf{s})$ is modeled as a mean 0 Gaussian process with exponential covariance function. For $\gamma_{[\frac{t}{365}]+1}^{(1)}$, where [ ] denotes the greatest integer function, thus counting *years* with this subscript and, as a result, the γ's provide annual intercepts to allow for yearly shifts, i.e., for hotter or colder years. The *sin* and *cos* terms are introduced to capture annual seasonality with their coefficients reflecting associated amplitudes. This seasonality is critical to ensure that an annual daily maximum temperature trajectory over the course of a year at a location will provide sensible realizations. elev$(\mathbf{s})$ is the elevation at $\mathbf{s}$ and lat$(\mathbf{s})$ is the latitude. Finally, $\rho^{U_t(\mathbf{s})}$ provides a *centered* AR(1) specification, bringing in the previous day's temperature, $Y_{t-1}(\mathbf{s})$.

For $\eta_t(\mathbf{s})$, Schliep et al. (2021) propose

$$\eta_t(\mathbf{s}) = \phi_0 + \phi_0(\mathbf{s}) + \phi_1(Y_{t-1}(\mathbf{s}) - q(\mathbf{s}))$$
$$+ \phi_2((Y_{t-1}(\mathbf{s}) - q(\mathbf{s}))\mathbf{1}(Y_{t-1}(\mathbf{s}) - q(\mathbf{s}) \geq 0))$$
$$+ \phi_3\sin(2\pi t/365) + \phi_4\cos(2\pi t/365).$$

Here, centering by the threshold yields more sensible transition probabilities. The threshold, $q(\mathbf{s})$, can be moved over to the intercept term in order to provide a spatially varying offset. However, the inclusion of $\phi_0(\mathbf{s})$, modeled as a Gaussian process, allows for a richer spatially-varying intercept.

We note that, as other common mixture models with a "cut point" to capture skewness, the unconditional density obtained from this model is discontinuous at the threshold. No continuity constraints are needed to smooth the density since this discontinuity does not affect the calculation of the extents over the threshold. This model was satisfactorily validated using out-of-sample prediction of characteristics of extreme values in three locations, see Section 5.2 in Schliep et al. (2021) and a summary in Section OR.2 in the Online Resource.

3 Illustrative summaries

3.1 The data and subregions

We analyze the evolution of the extent over time in two different areas around Aragón, a region located in northeastern Spain. The areas are quite different in topography and temperature, see Fig. 1, where the elevation and the thresholds that define a local EHE are shown.

- *B*1: Pyrenees, with latitude between 42.5N and 43N and longitude between −1.9W and 0.7E. It is a mountainous area with a high variability in elevation. The elevation of the stations varies from 442 to 1645 m asl, but some points in the area are over 3000 m. This leads to a high variability in the thresholds $q(\mathbf{s})$, that vary from 25.5 to 36 °C. This region shows a high biodiversity, including the last glaciers in Spain.

- *B*2: Central Ebro valley with latitude between 41.5N and 42.5N and longitude between −1.9W and 0.7E. This region is more homogeneous both in elevation, ranging from 245 to 546 m, and in temperature. The thresholds $q(\mathbf{s})$ vary from 33.8 to 37 °C. This region is the most populated in Aragón, and the most important farming zones are located here.

With regard to computing extents over these regions, *B*1 was partitioned into $1km \times 1km$ grid cells yielding, with regard to (2), $m = 7881$. *B*2, a much larger region, was partitioned into $2km \times 2km$ grid cells yielding, with regard to (2), $m = 5841$.

The following five observed decades *D*1: 1966-1975, *D*2: 1976-1985, *D*3: 1986-1995, *D*4: 1996-2005 and *D*5: 2006-2015, are considered in most of the following analysis to quantify the evolution over time of the different features related to the extent.

Altogether, the posterior predictive time series result in very large data files from which extents and persistences are computed. For instance, for *B*1, we have 50 years by 92

days by 7881 grid centroids by 500 replicates yielding 1.81263×10^{10} points.

In the sequel we present some comparative analysis for the regions *B*1 and *B*2. We present in the text the analysis for both regions and displays for the Pyrenees (*B*1) region, with analogous displays for the Ebro Valley (*B*2) region in the Online Resource.

3.2 Analysis of the time evolution of the extent

To offer some quantification of global warming, we analyze the evolution of the extent of EHE's across years. In that regard, we also analyze persistence through the extent of two-day EHE's. We consider the averages of the extent in the summer period JJA, $Av_{l \in JJA}Ext_{t,l}(0; B)$ and $Av_{l \in JJA}{}^2Ext_{t,l}(0; B)$. We also compare the evolution in the regions *B*1 and *B*2.

Figure 2 shows the plot of the posterior mean of $Av_{l \in JJA}Ext_{t,l}(0; B)$ for *B*1 and *B*2 vs. year t. Both regions show an increasing trend; the magnitude of the slope is similar in both regions, although the mean extent in *B*2 is approximately 1% higher than in *B*1. The analogous plot for $Av_{l \in JJA}{}^2Ext_{t,l}(0; B)$ is also shown in Figure 2. Similar conclusions about the extent of two-day EHE's are obtained, although the magnitude is reduced almost to half, the increase over time is slower, and the difference between the mean extent in the two regions is reduced to 0.5%. These conclusions are in agreement with the analysis of the raw empirical EHE extents using the observed series of temperature, see Figure OR.4 in the Online Resource. However, that figure reveals the limitations of the empirical extent. The plots in Figure OR.4 show the high variability of the empirical extent, and the impossibility of inference to compare the two decades.

To study the entire distribution, we analyze $Av^D_{l \in JJA}Ext_{t,l}(0; B)$ using the 5000 realizations $Av^D_{l \in JJA}\widetilde{Ext_{t,l}}(0; B)$. The boxplots of the distribution for each decade in *B*1 are shown in Fig. 3, and Table 1 gives some summary measures in the two regions. Apart from a small dip in *D*4, an almost linear increase is observed. Similar conclusions are obtained in *B*2. With regard to the evolution of the extent of two-day EHE's, the boxplots of $Av^D_{l \in JJA}{}^2Ext_{t,l}(0; B)$ in Fig. 3, and Table OR.1 in the Online Resource, confirm that it is quite similar to the evolution of the extent of EHE's, but smaller and slower in magnitude.

To better illuminate the global increase of the extent in the summer season, the posterior density for the first and the last decades is shown in Fig. 3 and Figure OR.5, where a shift of the central location and a slightly higher

Fig. 2 Posterior mean of $Av_{l \in JJA}Ext_{t,l}(0;B)$ (*left*) and posterior mean of $Av_{l \in JJA}{}^2Ext_{t,l}(0;B)$ (*right*) for $B1$ (*red*) and $B2$ (*black*) and linear trend fitted to the means
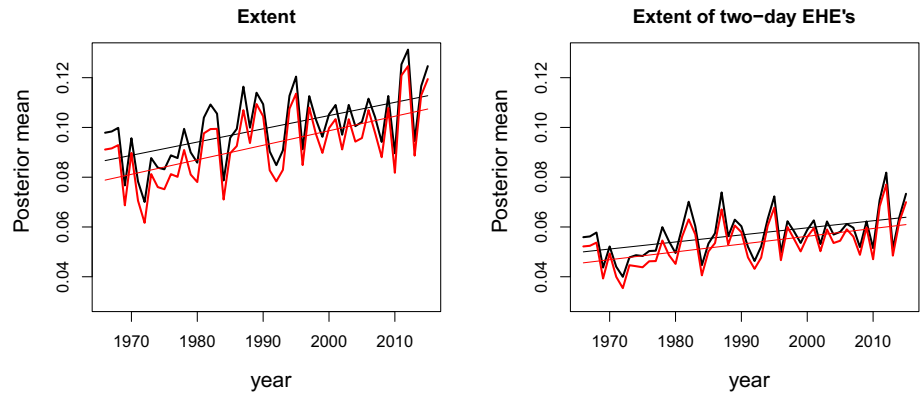

Fig. 3 *Top*: Boxplots of the posterior density of $Av_{l \in JJA}^D Ext_{t,l}(0;B1)$ (*left*) and of the posterior density of $Av_{l \in JJA}^D{}^2Ext_{t,l}(0;B1)$ (*right*) in the five decades. *Bottom*: Posterior density of $Av_{l \in JJA}^D Ext_{t,l}(0;B1)$ in $D1$ (*black*) and $D5$ (*red*). Vertical lines are the posterior means
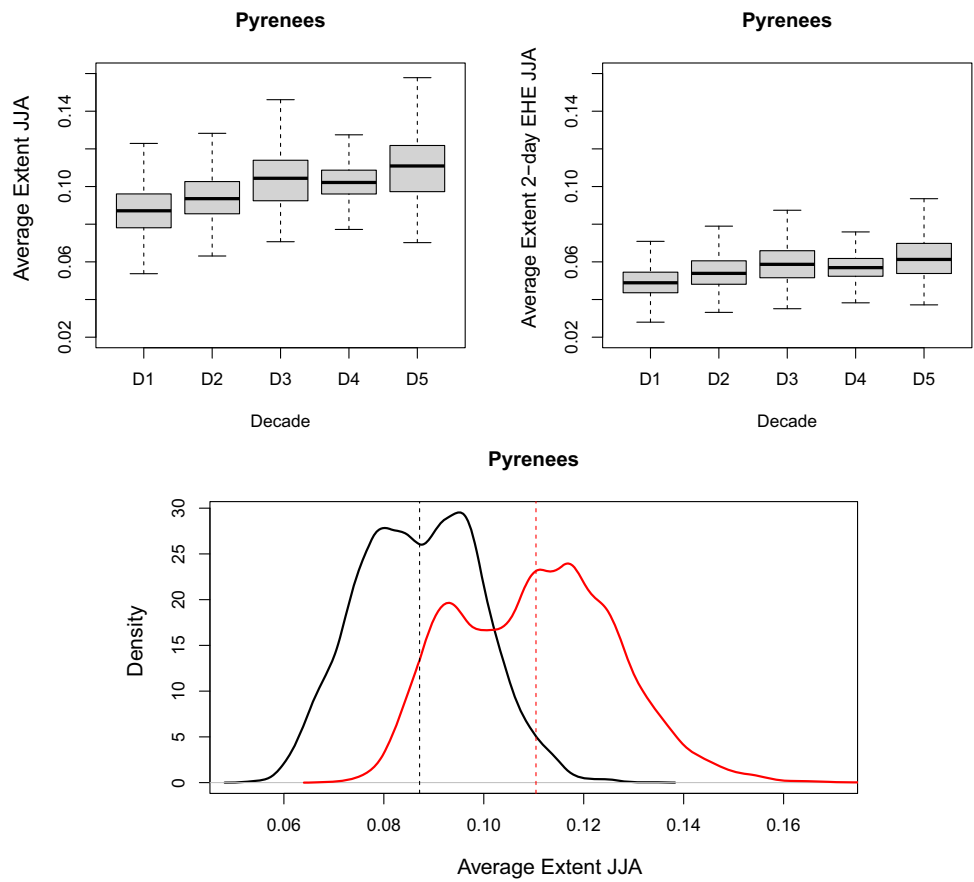
Table 1 Posterior mean, standard deviation, lower and upper 0.05 quantiles of $Av_{l \in JJA}^D Ext_{t,l}(0;B)$ for the five decades in regions $B1$ and $B2$

Decades	B1				B2			
	Mean	SD	p5	p95	Mean	SD	p5	p95
D1	0.087	0.012	0.068	0.107	0.080	0.012	0.060	0.098
D2	0.094	0.012	0.076	0.116	0.087	0.011	0.070	0.105
D3	0.104	0.014	0.082	0.126	0.097	0.014	0.076	0.118
D4	0.103	0.010	0.088	0.120	0.097	0.009	0.083	0.112
D5	0.110	0.016	0.086	0.137	0.105	0.016	0.080	0.128

Table 2 Posterior mean and 90% credible intervals of the increase between $D5$ and $D1$ of different averages of the extents, in regions $B1$ and $B2$

	B1		B2	
	Post. mean	90% CI	Post. mean	90% CI
$^r\Delta_{D1}^{D5}(B) = Av_{l \in JJA}^{D5}\,{}^rExt_{t,l}(0;B) - Av_{l \in JJA}^{D1}\,{}^rExt_{t,l}(0;B)$				
One-day EHE's ($r = 1$)	0.023	$(-0.009,\ 0.061)$	0.025	$(-0.010,\ 0.063)$
Two-day EHE's ($r = 2$)	0.013	$(-0.006,\ 0.041)$	0.014	$(-0.006,\ 0.041)$
$^r\Delta_{D5,D1}(B) = Av_{t \in D5, l \in JJA}\,{}^rExt_{t,l}(0;B) - Av_{t \in D1, l \in JJA}\,{}^rExt_{t,l}(0;B)$				
One-day EHE's ($r = 1$)	0.023	$(0.021,\ 0.026)$	0.025	$(0.023,\ 0.028)$
Two-day EHE's ($r = 2$)	0.013	$(0.007,\ 0.011)$	0.014	$(0.012,\ 0.016)$

variability is observed. Table 2 summarizes the posterior mean and 90% credible intervals (CI's) of the difference

$$^r\Delta_{D1}^{D5}(B) = Av_{l \in JJA}^{D5}\,{}^rExt_{t,l}(0;B) - Av_{l \in JJA}^{D1}\,{}^rExt_{t,l}(0;B)$$

for $r = 1, 2$ (superscript $r = 1$ is omitted for simplicity) and $B1$ and $B2$. In both regions, the interval for the one-day EHE's contains zero, showing that it is possible that the JJA average extent of one year in the last decade is lower than in the first decade. However, it is unlikely according to the posterior probability $P(\Delta_{D1}^{D5}(B) > 0 \mid y)$, which is 0.803 in $B1$ and 0.805 in $B2$. Similar conclusions are obtained for the two-day EHE's since $P(^2\Delta_{D1}^{D5}(B) > 0 \mid y)$ are 0.815 in $B1$ and 0.808 in $B2$.

Evidence of a significant increase over time in the decadal average is much stronger. Summary measures of the difference of the decadal average,

$$^r\Delta_{D5,D1}(B) = Av_{t \in D5, l \in JJA}\,{}^rExt_{t,l}(0;B)$$
$$- Av_{t \in D1, l \in JJA}\,{}^rExt_{t,l}(0;B)$$

for $r = 1, 2$ are also shown in Table 2. The CI's for the one-day EHE's are narrower than the corresponding CI's of the yearly averages and do not include zero. The posterior probabilities $P(^r\Delta_{D5,D1}(B) > 0 \mid y)$ are essentially 1 in both regions for $r = 1, 2$, indicating that the decadal average in $D5$ is significantly higher than in $D1$ in all the cases.

Another question of interest is the comparison of the increment of the extent between $D1$ and $D5$ in the two regions under study. Fig. 2 shows that the trends in $B1$ and $B2$ are quite parallel, and that the increase of the JJA average extent between $D1$ and $D5$ is quite similar in both regions. The posterior mean of the difference between $B1$ and $B2$ of the increase in the yearly average, $\Delta_{D1}^{D5}(B1) - \Delta_{D1}^{D5}(B2)$, is -0.002 and the 90% CI is $(-0.013, 0.009)$ while the CI of the decadal average, $\Delta_{D5,D1}(B1) - \Delta_{D5,D1}(B2)$ is $(-0.006, 0.002)$. That means that although there is a shift in the mean value of the average extent in both regions, there is no evidence of a significant difference in the increase of average extent, and both regions show a similar evolution over time.

3.3 Evolution of the seasonal pattern and the beginning of the summer

To analyze the evolution of the seasonal pattern in JJA of the extent $Ext_{t,l}(0;B)$, $l = 1, \ldots, 92$, and $^2Ext_{t,l}(0;B)$, Fig. 4 shows the plot vs. day within year of the posterior mean of $Av_{t \in D}Ext_{t,l}(0;B)$ in $D1$ and $D5$ in region $B1$. The seasonal pattern attains the maximum mean extent at the end of July in all cases, and a similar increase of the mean across decades, is observed in both regions. The analogous plot of $Av_{t \in D}\,{}^2Ext_{t,l}(0;B)$, see Fig. 4, shows a smoother seasonal pattern of the extent of the two-day EHE's, and a smaller increase between decades than in the extent of the one-day EHE's. The same conclusions are obtained in region $B2$, and the corresponding plots are shown in Figure OR.6 in the Online Resource.

The increase between decades is quantified in Table 3 that summarizes the posterior distribution of the monthly averages $Av_{l \in Jn}^D Ext_{t,l}(0;B)$, $Av_{l \in Jl}^D Ext_{t,l}(0;B)$ and $Av_{l \in Ag}^D Ext_{t,l}(0;B)$, in $D1$ and $D5$. There is an increase in the posterior mean in the three months with a similar seasonal pattern in both regions: the highest absolute increase, more than 3%, is observed in July and the lowest, around 1.4%, in June. This increase is higher in the upper tail: almost 2% in June and 4% in July. The posterior probability $P(Av_{l \in Jn}^{D5} Ext_{t,l}(0;B) > Av_{l \in Jn}^{D1} Ext_{t,l}(0;B) \mid y)$ is 0.813 in $B1$ and 0.818 in $B2$, and in July and August they are 0.805 and 0.804, in both regions.

3.3.1 Time to beginning of EHE's

An important feature related to the seasonality is the beginning of extreme temperatures in summer, and the analysis of its change across years. To that end, we define the variable $L_t(v;B)$, the number of days to the first day l within the period JJA in year t with an extent higher or equal to v, that is the first day l such that $Ext_{t,l}(0;B) \geq v$. If no extent over v is observed in a year, we set $L_t(v;B) = 92$.

Fig. 4 Posterior means of $Av_{t \in D}Ext_{t,l}(0;B1)$ (*left*) and $Av_{t \in D}{}^2Ext_{t,l}(0;B1)$ (right) in $D1$ (*black*) and $D5$ (*red*), Pyrenees
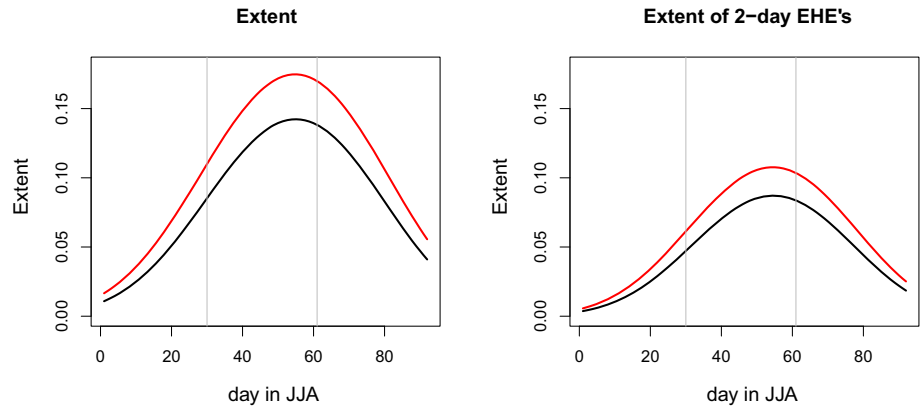

Table 3 Posterior mean, standard deviation and lower and upper 0.05 quantiles of $Av_{l \in Jn}Ext_{t,l}^D(0;B)$, $Av_{l \in Jl}Ext_{t,l}^D(0;B)$ and $Av_{l \in Ag}Ext_{t,l}^D(0;B)$ in $D1$ and $D5$, in regions $B1$ and $B2$

Decade	Month	B1				B2			
		mean	SD	p5	p95	mean	SD	p5	p95
D1	June	0.042	0.007	0.030	0.053	0.036	0.007	0.024	0.046
D5		0.056	0.009	0.041	0.071	0.051	0.009	0.036	0.065
D1	July	0.126	0.016	0.099	0.152	0.117	0.016	0.091	0.142
D5		0.157	0.021	0.125	0.192	0.151	0.021	0.118	0.182
D1	August	0.092	0.013	0.072	0.114	0.085	0.013	0.065	0.105
D5		0.117	0.018	0.090	0.147	0.111	0.017	0.084	0.138

Since extents over high thresholds are being analyzed, large extents are very rarely observed so we only consider here $v = 0.1$. ${}^2L_t(v;B)$, the first day with an extent of the two-day EHE's higher or equal than v, is defined analogously.

Figure 5 shows the posterior mean of $L_t(0.1;B)$ vs. year for $B1$ and $B2$, and the the analogous plot for ${}^2L_t(0.1;B)$. A decreasing trend is observed in both variables and in both regions. However, the decrease of the mean of ${}^2L_t(0.1;B)$ across years is much greater, almost 30 days, versus less than 10, but with much higher variability. This variability is likely due to the low incidence of the considered event.

Figure 6 shows the posterior density of the first day, estimated in a decade, $L_t^D(0.1;B)$, in $D1$ and $D5$ in $B1$; the
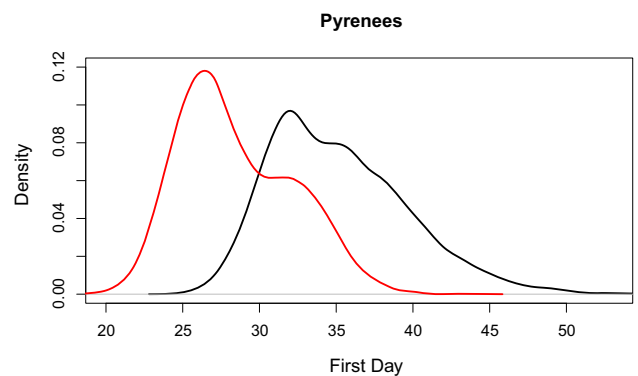
Fig. 6 Posterior density of $L_t^D(0.1;B1)$ for $D1$ (*black*) and $D5$ (*red*)

Fig. 5 Posterior mean of the first day $L_t(0.1;B)$ (*left*) and posterior mean of the first day ${}^2L_t(0.1;B)$ (*right*) vs. year, for $B1$ (*red*) and $B2$ (*black*)
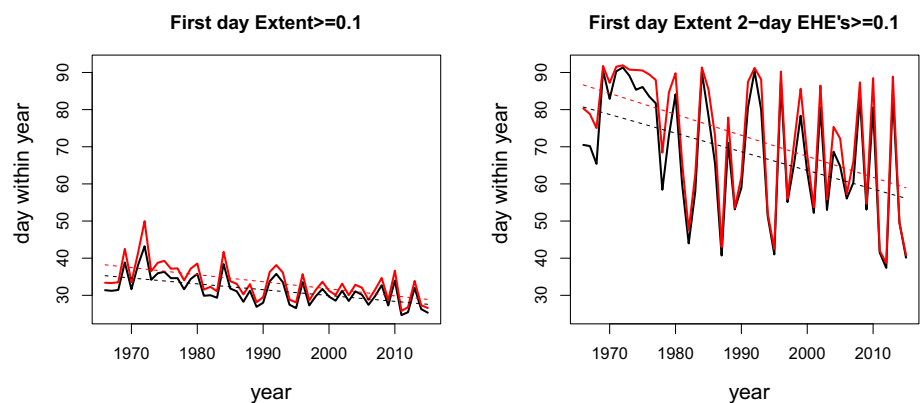
analogous plot in $B2$ is shown in Figure OR.7 in the Online Resource. Table 4 summarizes the posterior mean, standard deviation and 90% CI for $L_t^D(0.1; B)$, for the five decades in the observed period. The mean of the first day between the first and the last decade has decreased almost 7 days in $B1$ and 8 days in $B2$. The posterior probability $P(L_t^{D1}(0.1; B) > L_t^{D5}(0.1; B) \mid y)$, that is the probability that the first day of the summer where more than 10% of the region is under extreme temperatures occurs earlier in the last decade, is 0.775 in $B1$ and 0.781 in $B2$.

A summary of $^2L_t^D(0.1; B)$ analogous to that in Table 4 is presented in Table OR.2 in the Online Resource. Since the probability of a two-day EHE is quite low, it is likely that in a year all the extents of two-day EHE's are lower or equal than 0.1 and, consequently, the event $P(^2L_t^D(0.1; B) = 92 \mid y)$ has a positive probability mass. However, this posterior probability decreases over the observed period. In $B1$, the percentage of years with $L_t(0.1; B) = 92$ is 0.77 in $D1$ and 0.31 in $D5$, and 0.88 and 0.36 in $B2$. Further evidence that $L_t^D(0.1; B)$ is decreasing is that, conditionally to the fact that an extent higher than 0.1 occurs in a year, the posterior mean of $^2L_t^D(0.1; B)$ has decreased more than 5 days between $D1$ and $D5$ in both regions.

3.4 Behavior of extent across choices of *w*

The previous sections consider results for $Ext_{t,l}(w; B)$ only at $w = 0$, that is the extent corresponding to the threshold used to define an EHE. Here, we consider the effect on extent by adjustment of the local thresholds to lower extreme temperatures and to higher extreme temperatures. To that end, we consider $Ext_{t,l}(w; B)$ for a grid of values over and under the threshold, $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5$ °C.

To study the evolution across years and across threshold, Fig. 7 shows the posterior mean of $Av_{l \in JJA} Ext_{t,l}(w; B1)$ vs.

Table 4 Posterior mean, standard deviation and lower and upper 0.05 quantiles of $L_t^D(0.1; B)$ in $D1$ and $D2$, in regions $B1$ and $B2$

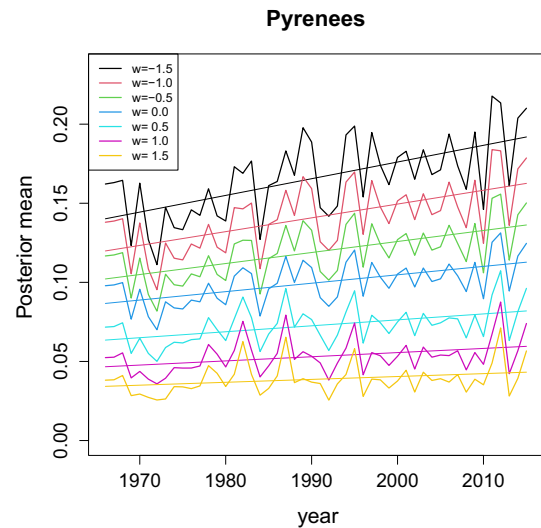| Decade	B1				B2			
	Mean	SD	p5	p95	Mean	SD	p5	p95
D1	35.230	4.848	29	43	38.214	6.560	31	48
D2	33.052	3.625	28	39	35.463	3.926	30	42
D3	30.252	3.705	25	37	32.146	4.001	27	39
D4	30.146	2.600	26	34	31.884	2.686	28	36
D5	28.490	3.748	23	35	30.067	4.068	25	37

Fig. 7 Posterior mean of $Av_{l \in JJA} Ext_{t,l}(w; B1)$ for $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5$ °C and linear trends fitted to the means

year for the grid of w values, and the corresponding linear trends fitted to the means. An increasing trend is observed for all the w values but the trend is stronger for smaller w's. More precisely, the ratio between the slope for $w = -1.5$ and $w = 1.5$ is $0.00106/0.00018 = 5.8$. The finding is that the incidence of extents associated with more extreme thresholds is increasing at a slower rate than that for less extreme thresholds.

This increase over time is not only observed in the mean, but in the entire distribution. As an example, we consider the decadal average of July, where Fig. 8 shows the posterior density of $Av_{t \in D, l \in Jl} Ext_{t,l}(w; B1)$ across w for $D1$ and $D5$; the analogous plot for $B2$ is shown in Figure OR.8 in the Online Resource. It is observed that the
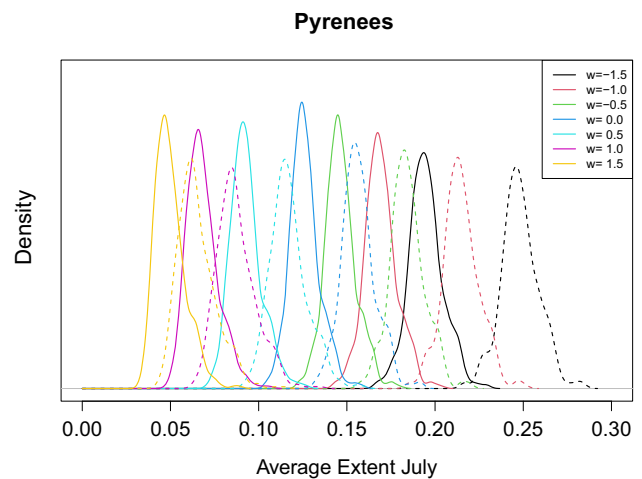

Fig. 8 Posterior density of $Av_{t \in D, l \in Jl} Ext_{t,l}(w; B1)$ for $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5$ and $D1$ (*solid line*) and $D5$ (*dashed line*)

shape of the distribution is very similar across w and across decades but, in both cases, there are clear shifts that are not homogeneous across w.

The different evolution across w is observed if we compare the mean difference between decades. For example, in $B1$, the posterior mean difference

$$E\left(Av_{l \in Jl}^{D5}Ext_{t,l}(w; B1) - Av_{l \in Jl}^{D1}Ext_{t,l}(w; B1) \mid y\right)$$

is 5% for $w = -1.5$, while for $w = 1.5$ is 1%. The change is also observed if we compare the mean differences between the range of w values across decades. Again in $B1$, the posterior mean difference

$$E\left(Av_{l \in Jl}^{D}Ext_{t,l}(-1.5; B1) - Av_{l \in Jl}^{D}Ext_{t,l}(1.5; B1) \mid y\right)$$

is 18% in $D5$, while in $D1$ is 14%. These values together with the counterparts for $B2$ are summarized in Table OR.3 in the Online Resource. The change in the entire distribution is quantified by the posterior probability of the extent for a given w in the last decade being higher than in the first,

$$P\left(Av_{l \in Jl}^{D5}Ext_{t,l}(w; B1) > Av_{l \in Jl}^{D1}Ext_{t,l}(w; B1) \mid y\right).$$

Table 5 summarizes these posterior probabilities and shows that there is a nonmonotonic decrease across w.

Figure 9 shows the posterior mean of the average $Av_{t \in D}Ext_{t,l}(w; B1)$ vs. day within year for $D1$ and $D5$ in order to compare the seasonal behavior of the extent across w; the analogous plot for $B2$ is shown in Figure OR.9 in the Online Resource. It can be seen that the seasonal pattern is smoother for larger w. In all the cases, the seasonal pattern is more pronounced in the last decade but the changes across w are not homogeneous, with stronger differences in smaller w values. More precisely, the seasonal pattern for $w = 1.5$ in $D5$ is more pronounced than its counterpart in $D1$, approaching that of $w = 1$ in $D1$, with probability $P\left(Av_{t \in D5}^{JJA}Ext_{t,l}(1.5; B1) > Av_{t \in D1}^{JJA}Ext_{t,l}(1.5; B1) \mid y\right)$, that is essentially 1, and $P\left(Av_{t \in D5}^{JJA}Ext_{t,l}(1.5; B1) > Av_{t \in D1}^{JJA}Ext_{t,l}(1; B1) \mid y\right) = 0.18$. For smaller w values, the differences are higher, with the posterior probabilities comparing the previous averages in $w = -1.5$ and $w = -1.5$, and in

Table 5 Posterior probabilities for a grid of w values, $P\left(Av_{l \in Jl}^{D5}Ext_{t,l}(w; B) > Av_{l \in Jl}^{D1}Ext_{t,l}(w; B) \mid y\right)$, for regions $B1$ and $B2$

w	−1.5	−1.0	−0.5	0.0	0.5	1.0	1.5
$B1$	0.808	0.806	0.804	0.805	0.817	0.779	0.732
$B2$	0.814	0.810	0.808	0.805	0.814	0.803	0.763

$w = -0.5$ and $w = -1.5$ equal to 1; in addition, the seasonal pattern of $w = -1.5$ in $D1$ is only slightly more pronounced than the pattern in $w = -0.5$ in $D5$.

4 Summary and future work

Notions of the spatial extent of heat waves and extreme heat events have been considered informally and descriptively in the climate community. Here we have introduced a formal probabilistic definition for extents of extreme heat events. For a specified region, for a given day, the definition of spatial extent takes the form of a block average over the region. It is an average of indicator variables which identify exceedance of a local threshold by the daily max temperature surface for the day at each location within the region. We demonstrate that extents can be calculated through Monte Carlo integration and can be obtained for realizations from arbitrary space-time autoregressive models for daily max temperatures. Using a dataset of daily max temperatures over 50 years, adopting a particular choice of model, working within a Bayesian framework, we obtained posterior predictive samples of daily temperature time series on a fairly fine grid scale to implement the Monte Carlo integrations.

With these samples, we calculated daily, seasonal and decadal averages of the extents for two regions around the Comunidad Autónoma de Aragón in Spain. We generalized these extents to capture extents of persistence of extreme heat. We made comparisons across decades to reveal evidence of increasing extent over time. We also studied other features related to the extent of EHE's, for example, the first day in the period JJA with an extent higher than a given percentage, and the behaviour of extent across choices of the threshold. Following our approach, other extents yielding other comparisons may be developed according to the interest of the user.

With regard to the regions under study, Pyrenees and Ebro Valley, we found that a clear increase of the extent of EHE's across time is observed. For example, the posterior probability of the yearly average extent across JJA in the decade 2006–2015 being higher than in 1966-1975 is higher than 0.8. It is also found that the first day of the summer where the extent of EHE's is higher than 10% has decreased around seven days. The extent of EHE's defined with all the considered thresholds is increasing, but the incidence of extents associated with more extreme thresholds is increasing at a slower rate than that for less extreme thresholds.

Future work will explore the temporal evolution of geographic extent for different regions to enable comparison. It will also examine the challenges of working with regions at larger spatial scales. Alternatively, with
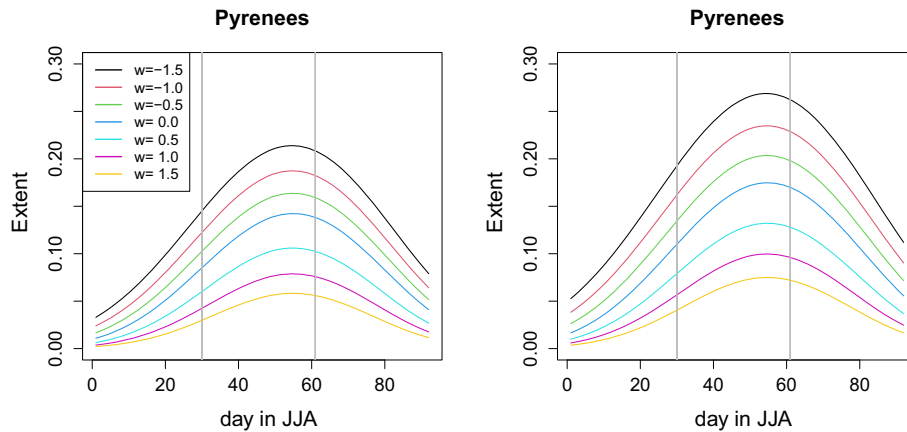
Fig. 9 Posterior mean of $Av_{t \in D}Ext_{t,l}(w; B1)$ for $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5$ for $D1$ (*left*) and $D5$ (*right*)

suitable computing capability, we can consider investigating extents at higher spatial resolution than done here. Further, while here we work with actual weather data, another goal is to consider projection of future spatial extent using future climate scenarios.

Availability of data and code Code (in R), temperature series used to fit the model, and simulated series used to compute the extents are available upon request.

Declarations

Conflicts of interest The authors declare that they do not have financial or personal interest that can inappropriately influence this work.

References

Abaurrea J, Asín J, Cebrián AC (2018) Modelling the occurrence of heat waves in maximum and minimum temperatures over Spain and projections for the period 2031–60. Global Planet Change 161:244–260. https://doi.org/10.1016/j.gloplacha.2017.11.015

Alexander LV (2016) Global observed long-term changes in temperature and precipitation extremes: a review of progress and limitations in IPCC assessments and beyond. Weather Clim Extrem 11:4–16

Amengual A, Homar V, Romero R, Brooks HE, Ramis C, Gordaliza M, Alonso S (2014) Projections of heat waves with high impact on human health in Europe. Global Planet Change 119:71–84

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data, 2nd edn. Chapman and Hall/CRC, New York, https://doi.org/10.1201/b17115

Bolin D, Lindgren F (2015) Excursion and contour uncertainty regions for latent Gaussian models. J R Stat Soc Ser B Stat Methodol 77(1):85–106

Campbell SL, Remenyi T, White CJ, Johnston F (2018) Heatwave and health impact research: a global review. Health Place 53:210–218

Cebrián AC, Castillo-Mateo J, Asín J (2021) Record tests to detect non-stationarity in the tails with an application to climate change. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-021-02122-w

Chan Y, Li H (2008) Tail dependence for multivariate t-copulas and its monotonicity. Insur Math Econ 42(2):763–770

Fischer E, Sippel S, Knutti R (2021) Increasing probability of record-shattering climate extremes. Nat Clim Change, 11. https://doi.org/10.1038/s41558-021-01092-9

French J, Sain S (2013) Spatio-temporal exceedance locations and confidence regions. Ann Appl Stat, 7. https://doi.org/10.1214/13-AOAS631

Hazra A, Huser R (2021) Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank semiparametric spatial model. Ann Appl Stat 15(2):572–596. https://doi.org/10.1214/20-AOAS1418

Keellings D, Moradkhani H (2020) Spatiotemporal evolution of heat wave severity and coverage across the United States. Geophys Res Lett 47(9):e2020GL087097. https://doi.org/10.1029/2020GL087097

Keellings D, Waylen P (2014) Increased risk of heat waves in Florida: characterizing changes in bivariate heat wave risk using extreme value analysis. Appl Geogr 46:90–97

Keellings D, Waylen P (2015) Investigating teleconnection drivers of bivariate heat waves in Florida using extreme value analysis. Clim Dyn 44(11):3383–3391

Khaliq MN, St-Hilaire A, Ouarda TBMJ, Bobée B (2005) Frequency analysis and temporal pattern of occurrences of southern Quebec heatwaves. Int J Climatol 25(4):485–504

Khan N, Shahid S, Ismail T, Ahmed K, Nawaz N (2019) Trends in heat wave related indices in Pakistan. Stochast Environ Res Risk Assess 33:287–302

Lahiri SN, Kaiser MS, Cressie N, Hsu NJ (1999) Prediction of spatial cumulative distribution functions using subsampling. J Am Stat Assoc 94(445):86–97

Lai Y, Dzombak D (2019) Use of historical data to assess regional climate change. J Clim 32:4299–4320. https://doi.org/10.1175/JCLI-D-18-0630.1

Lemonsu A, Beaulant AL, Somot S, Masson V (2014) Evolution of heat wave occurrence over the Paris basin (France) in the 21st century. Clim Res 61:75–91

Lhotka O, Kyselý J (2015) Characterizing joint effects of spatial extent, temperature magnitude and duration of heat waves and cold spells over Central Europe. Int J Climatol 35(7):1232–1244

Lyon B, Barnston AG, Coffel E, Horton RM (2019) Projected increase in the spatial extent of contiguous US summer heat waves and associated attributes. Environ Res Lett 14(11):114029

Perkins SE, Alexander LV (2013) On the measurement of heat waves. J Clim 26(13):4500–4517

Rebetez M, Dupont O, Gaillard M (2009) An analysis of the July 2006 heatwave extent in Europe compared to the record year of 2003. Theor Appl Climatol 95:1–7. https://doi.org/10.1007/s00704-007-0370-9

Reich BJ, Shaby BA, Cooley D (2014) A hierarchical model for serially-dependent extremes: a study of heat waves in the western US. J Agricu Biol Environ Stat 19(1):119–135

Romero-Béjar JL, Madrid A, Angulo J (2018) Quantile-based spatiotemporal risk assessment of exceedances. Stochast Environ Res Risk Assess 32:2275–2291. https://doi.org/10.1007/s00477-018-1562-9

Schliep EM, Gelfand AE, Abaurrea J, Asín J, Beamonte MA, Cebrián AC (2021) Long-term spatial modelling for characteristics of extreme heat events. J R Stat Soc Ser A Stat Soc 184(3):1070–1092. https://doi.org/10.1111/rssa.12710

Shaby BA, Reich BJ, Cooley D, Kaufman CG (2016) A Markov-switching model for heat waves. Ann Appl Stat 10(1):74–93

Short M, Carlin B, Gelfand A (2005) Bivariate spatial process modeling for constructing indicator or intensity weighted spatial CDFs. JABES 10(3):259–275

Smith T, Zaitchik B, Gohlke J (2013) Heat waves in the United States: definitions, patterns and trends. Clim Change 118(3):811–825

Sommerfeld M, Sain S, Schwartzman A (2018) Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. J Am Stat Assoc 113(523):1327–1340. https://doi.org/10.1080/01621459.2017.1341838

Zhong P, Huser R, Opitz T (2020) Modeling non-stationary temperature maxima based on extremal dependence changing with event magnitude. arXiv 2006.01569

.

Online Resource

Spatio-temporal Analysis of the Extent of an Extreme Heat Event

Authors: Ana C. Cebrián · Jesús Asín · Alan E. Gelfand · Erin M. Schliep · Jorge Castillo-Mateo · María A. Beamonte · Jesús Abaurrea

This supplementary material includes some complementary figures and tables. Most of them, except those in Section OR.1, are figures for *Section 3: Illustrative summaries*. For example, the preliminary analysis of the empirical extent and the displays of the analysis for the Ebro valley region ($B2$). These displays are analogous to the displays for the Pyrenees ($B1$) region presented in the text. The conclusions from these displays are also analogous to those described in the text.

OR.1 Comparison of kriged and empirical thresholds

In order to obtain a $q(\mathbf{s})$ surface, we employed a standard kriging model for the thresholds. It is based on the thresholds for the observed sites with their associated DTM elevations. We used the 18 stations in the database plus an additional 37 stations which had temperature data available between 1953 and 1962, yielding a total of 55 stations. The series are selected from the database of Spanish Meteorological Office (AEMET). Only series that fulfill some quality standards, and contain at least 70% observed values in each month (June, July and August) in the reference period 1953-62, were selected.

Figure OR.1 shows the resulting $q(\mathbf{s})$ surface. Black numbers are the levels of the contour curves, plotted at 2°C increments, and red numbers are the empirical thresholds in the 55 observed series. The agreement with the empirical thresholds shows that the obtained $q(\mathbf{s})$ surface is reasonable.
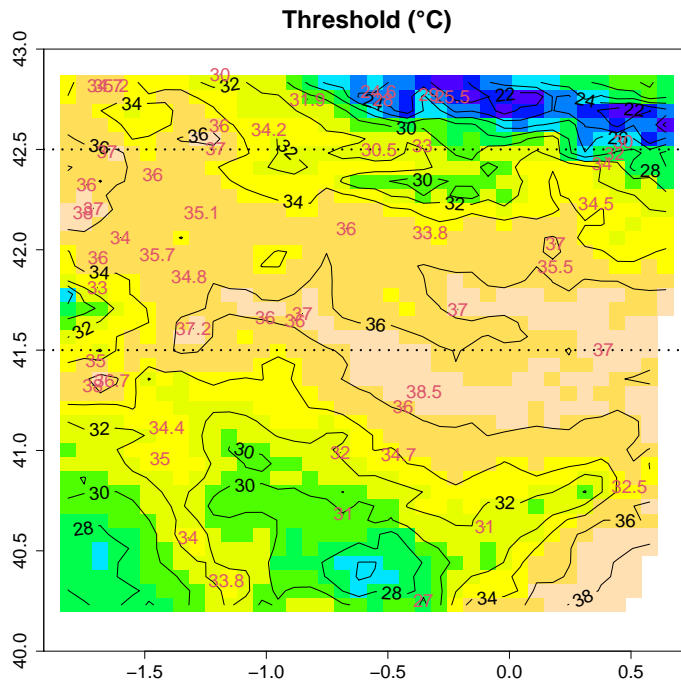
Fig. OR.1 Threshold map and empirical thresholds in the observed series (red numbers).

OR.2 Validation analysis of the daily max temperature model

This model was thoroughly validated, see Section 5.2 in Schliep et al (2021), using out-of-sample prediction of characteristics of extreme values in three locations, Tornos, Zaragoza and Yesa, that represent different climates in the region. The validation compares the posterior predictive distribution of exceedance days as well as EHE characteristics (duration, and intensity) with the observed empirical counterparts. Comparisons are made for the entire time window of the analysis, 1966-2015, as well as for two 10-year periods, 1976-1985 and 2006-2015 to examine the time evolution. A brief summary of that validation is shown here.

Figure OR.2 shows the mean and 90% credible interval of the probability density for events lasting 3 days, 4-5 days, 6-7 days, and 8 or more days together with the empirical probabilities. The results reveal that, for each site and for each duration bin, our predictive intervals always capture the observed/true proportion.

Figure OR.3 shows the mean and 90% credible of cumulative probabilities of the average (or maximum) excess being greater than or equal to a set of discrete values for the average. Our model appears to capture these cumulative probabilities well for both average and maximum exceedance.
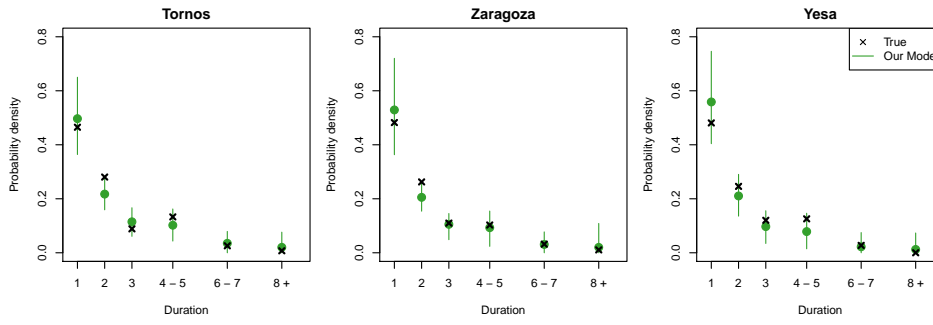
Fig. OR.2 Posterior predictive mean estimates and 90% credible intervals of the probability density for the durations of extreme heat events across the years 1966-2015. For each of the three out-of-sample locations, true duration density is plotted for the durations 3 days, 4-5 days, 6-7 days, and 8 or more days.
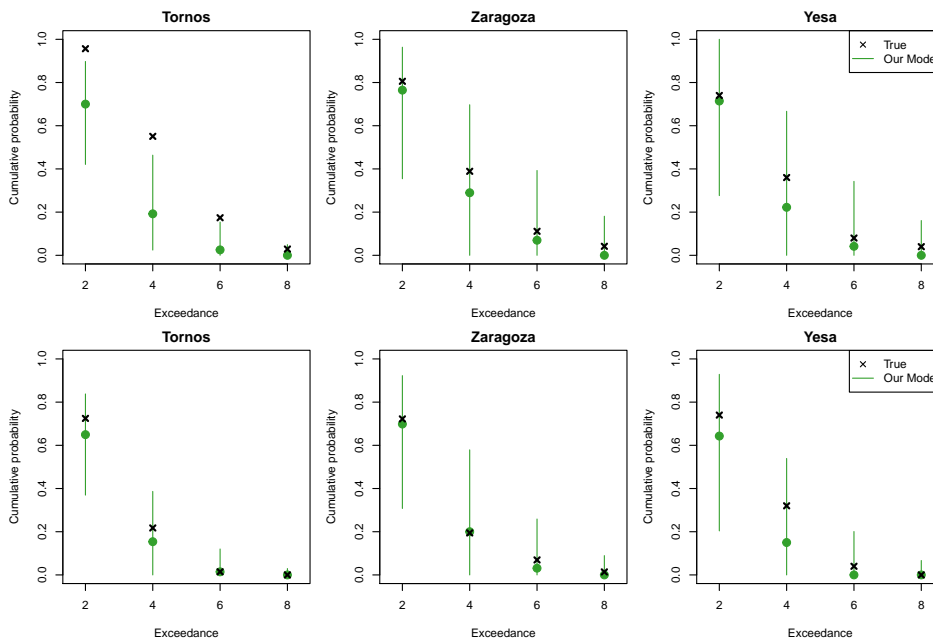

Fig. OR.3 Posterior predictive mean estimates and 90% credible intervals of the cumulative probability of the average (top) and maximum (bottom) exceedance being at or greater than the specified level during an EHE lasting 3 of more days.

OR.3 Analysis of the time evolution of the extent

Analysis of the empirical extent. The plots in the upper part of Figure OR.4 show the mean of the empirical extent vs. day within JJA in the first and the last decade of the observed period, for regions $B1$ and $B2$. Figure OR.4 (bottom part) shows the plot of the mean of the empirical extent over JJA for each year

and the linear trend fitted to that mean in $B1$ and $B2$ (left plot), and the mean of the empirical extent vs. day within JJA (right plot). Although the comparison between the decades is not clear, the plots in the bottom part suggest that the mean level of the empirical extent of EHE's in $B2$ tends to be higher than in $B1$. A possible reason of this behaviour is that area $B1$ is more heterogeneous in elevation and topography than $B2$.

Figure OR.5 shows the plots analogous to those in Figure 3 in the text, for region $B2$. Table OR.1 shows the summary measures analogous to those in Table 1 in the text, for $Av_{l \in JJA} \, ^2Ext_{t,l}(0; B)$.
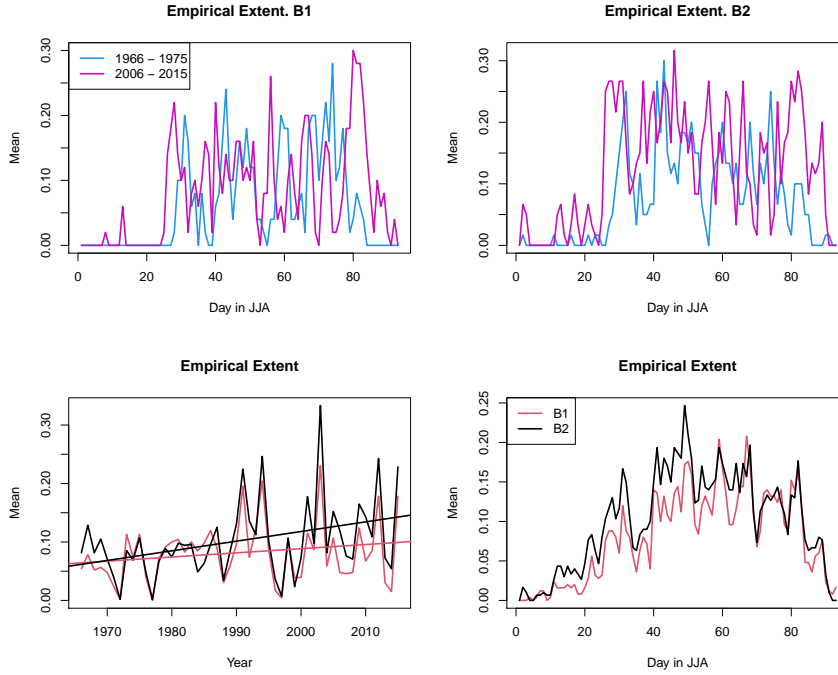

Fig. OR.4 Top: Mean of the empirical extent in the first and the last decade vs. day within JJA in $B1$ (left) and $B2$ (right). Bottom: Mean of the empirical extent over JJA and linear trend fitted to than mean in $B1$ and $B2$ (left) and mean of the empirical extent vs. day within JJA (right).
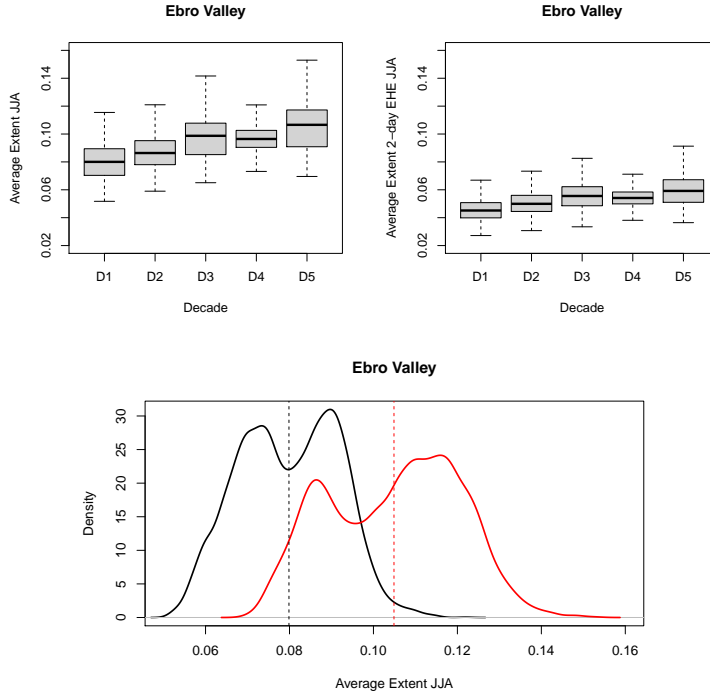
Fig. OR.5 Top: Boxplots of the posterior density of $Av^D_{l \in JJA} Ext_{t,l}(0; B2)$ (left) and of the posterior density of $Av^D_{l \in JJA}{}^2 Ext_{t,l}(0; B2)$ (right) in the five decades. Bottom: Posterior density in $D1$ (black) and $D5$ (red) in $B2$. Vertical lines are the posterior means.

Table OR.1 Posterior mean, standard deviation and lower and upper 0.05 quantiles of $Av^D_{l \in JJA}{}^2 Ext_{t,l}(0; B)$, by decades in areas $B1$ and $B2$.

Decades	B1				B2			
	Mean	sd	p5	p95	Mean	sd	p5	p95
$D1$	0.049	0.008	0.037	0.064	0.046	0.008	0.034	0.058
$D2$	0.055	0.010	0.041	0.074	0.051	0.008	0.039	0.066
$D3$	0.060	0.011	0.044	0.080	0.056	0.010	0.041	0.072
$D4$	0.058	0.008	0.046	0.071	0.054	0.007	0.044	0.066
$D5$	0.063	0.012	0.046	0.085	0.060	0.011	0.044	0.079

OR.4 Evolution of the seasonal pattern and the beginning of the summer

Figure OR.6 shows the plots analogous to those in Figure 4 in the text, for region $B2$. Figure OR.7 shows the plots analogous to those in Figure 6 in the text, for region $B2$. Table OR.2 shows the summary measures analogous to those in Table 4 in the text, for $^2L_t^D(0.1; B)$.
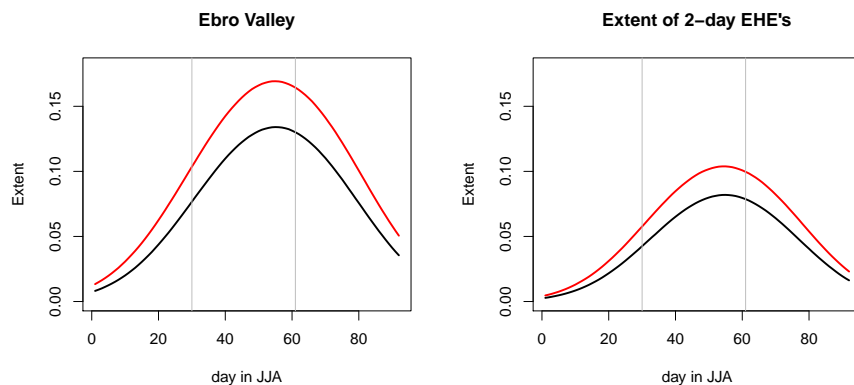

Fig. OR.6 Posterior means of $Av_{t \in D} Ext_{t,l}(0; B2)$ (left) and $Av_{t \in D} {}^2Ext_{t,l}(0; B2)$ (right) in $D1$ (black) and $D5$ (red), Ebro valley.
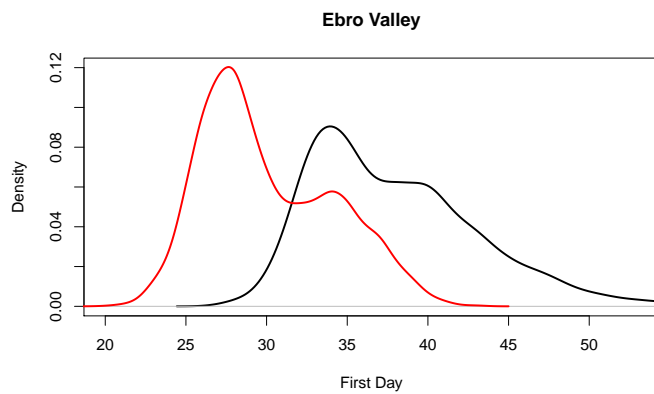

Fig. OR.7 Posterior density of $L_t^D(0.1; B2)$ for $D1$ (black) and $D5$ (red).

Table OR.2 Posterior mean, standard deviation and lower and upper 0.05 quantiles of $^2L_t^D(0.1; B)$, for decades $D1$ and $D2$ and regions $B1$ and $B2$.

Decade	B1				B2			
	Mean	sd	p5	p95	Mean	sd	p5	p95
$D1$	82.173	18.344	45	92	86.852	13.830	49	92
$D2$	71.122	23.037	40	92	77.308	20.850	43	92
$D3$	63.262	23.552	37	92	67.059	23.135	40	92
$D4$	66.827	22.616	40	92	71.413	21.873	43	92
$D5$	58.615	23.057	35	92	61.370	23.272	37	92

OR.5 Behavior of extent across choices of w

Figure OR.8 shows the plots analogous to those in Figure 7 in the text, for region $B2$. Table OR.3 summarizes the posterior mean of some increments of the average extents $Av^D_{l \in Jl} Ext_{t,l}(w; B)$ to compare the evolution across a grid of w values. Figure OR.9 shows the plots analogous to those in Figure 8 in the text, for region $B2$. Figure OR.10 shows the plots analogous to those in Figure 9 in the text, for region $B2$.
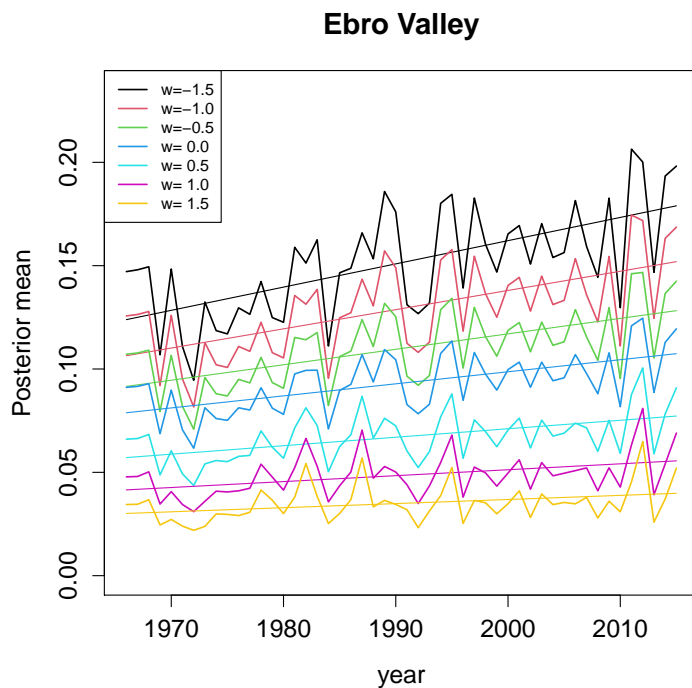

Fig. OR.8 Posterior mean of $Av_{l \in JJA} Ext_{t,l}(w; B2)$ for $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5^oC$ and linear trends fitted to the means.

Table OR.3 Posterior mean of some increments of the average extents to compare the evolution across w.

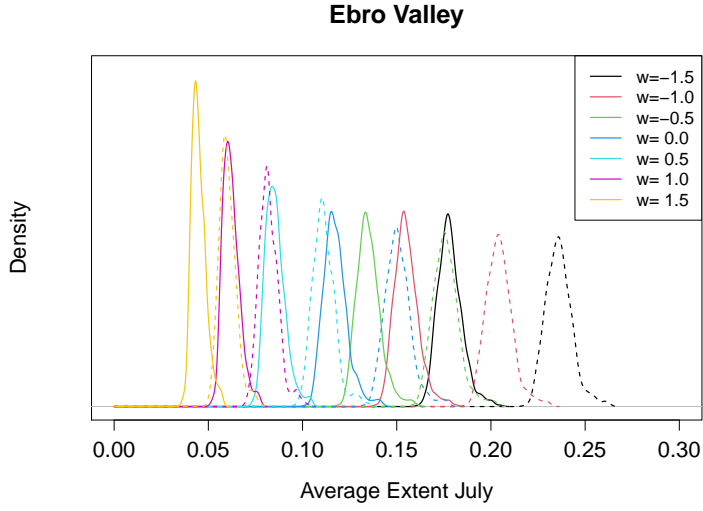|	B1		B2	
w	-1.5	1.5	-1.5	1.5
$E\left(Av_{l\in Jl}^{D5}Ext_{t,l}(w;B) - Av_{l\in Jl}^{D1}Ext_{t,l}(w;B) \mid y\right)$	0.053	0.015	0.058	0.015
Decade	D1	D5	D1	D5
$E\left(Av_{l\in Jl}^{D}Ext_{t,l}(-1.5;B) - Av_{l\in Jl}^{D}Ext_{t,l}(1.5;B)) \mid y\right)$	0.145	0.183	0.134	0.176

Fig. OR.9 Posterior density of $Av_{t\in D,l\in Jl}Ext_{t,l}(w;B2)$ for a grid of values $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5$ and $D1$ (solid line) and $D5$ (dashed line).
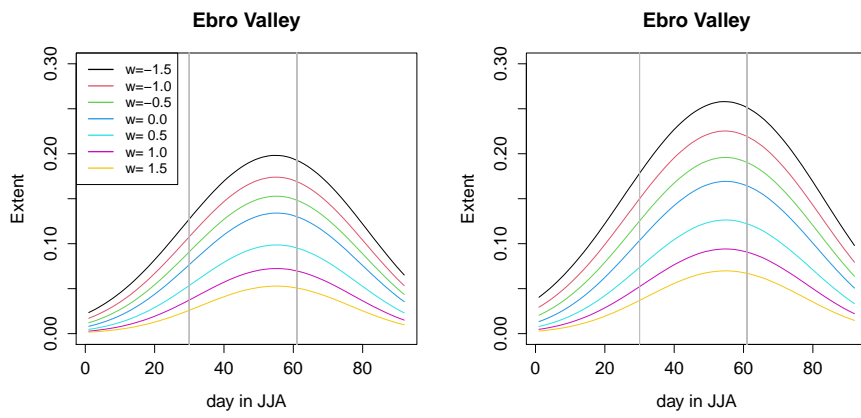
Fig. OR.10 Posterior mean of $Av_{t \in D} Ext_{t,l}(w; B2)$ for a grid of values $w = -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5$ for $D1$ (left) and $D5$ (right).

3.3 Spatial quantile autoregression for season within year daily maximum temperature data

This manuscript was published in:

And it was disseminated (speaker emphasized) in:

- ***Castillo-Mateo, J.***, Asín, J., Cebrián, A. C., & Gelfand, A. E. (2023, January 24–26). *Spatio-temporal modeling of the trend in daily maximum temperature quantiles: A case study in Aragón, Spain* [Poster session]. CLIVAR 2023: Towards an integrated view of climate, Madrid, Spain.

- ***Castillo-Mateo, J.***, Cebrián, A. C., Asín, J., & Gelfand, A. E. (2022, December 17–19). *Mixed effects quantile autoregressive modeling for point-referenced daily maximum temperatures in Aragon, Spain* [Contributed talk]. CFE-CMStatistics 2022, London, UK.

- *Gelfand, A. E.*, **Castillo-Mateo, J.**, Cebrián, A. C., Asín, J., Schliep, E. M., Beamonte, M. A., & Abaurrea, J. (2022, June 7–10). *Long term spatial modeling for exploring incidence of extreme heat events* [Plenary talk]. XXXIX Congreso Nacional de Estadística e Investigación Operativa, Granada, Spain. (See also in Section 3.2.)

- ***Castillo-Mateo, J.***, Gelfand, A. E., Asín, J., & Cebrián, A. C. (2022, June 1–3). *Detecting climate change in daily temperatures with a space-time quantile autoregressive model* [Contributed talk]. 10th International Workshop on Spatio-Temporal Modelling, Lleida, Spain.

- ***Castillo-Mateo, J.***, Gelfand, A. E., Asín, J., & Cebrián, A. C. (2022, May 25–27). *Spatio-temporal quantile autoregression for detecting changes in daily temperature in northeastern Spain* [Contributed talk]. XVIII Congreso de Biometría, Madrid, Spain.

> *"the average man, that hypothetical man ...who is comfortable with his head in the oven and his feet in a bucket of ice."*
>
> —————————————
>
> Roger W. Koenker, in Koenker (2017)

SPATIAL QUANTILE AUTOREGRESSION FOR SEASON WITHIN YEAR DAILY MAXIMUM TEMPERATURE DATA

BY JORGE CASTILLO-MATEO[1,a] ⓘ, JESÚS ASÍN[1,b] ⓘ, ANA C. CEBRIÁN[1,c] ⓘ,
ALAN E. GELFAND[2,e] ⓘ AND JESÚS ABAURREA[1,d]

[1]*Department of Statistical Methods, University of Zaragoza,* [a]*jorgecm@unizar.es,* [b]*jasin@unizar.es,* [c]*acebrian@unizar.es,*
[d]*abaurrea@unizar.es*

[2]*Department of Statistical Science, Duke University,* [e]*alan@stat.duke.edu*

Regression is the most widely used modeling tool in statistics. Quantile regression offers a strategy for enhancing the regression picture beyond customary mean regression. With time-series data, we move to quantile autoregression and, finally, with spatially referenced time series, we move to space-time quantile regression. Here, we are concerned with the spatiotemporal evolution of daily maximum temperature, particularly with regard to extreme heat. Our motivating data set is 60 years of daily summer maximum temperature data over Aragón in Spain. Hence, we work with time on two scales—days within summer season across years—collected at geocoded station locations. For a specified quantile, we fit a very flexible, mixed-effects autoregressive model, introducing four spatial processes. We work with asymmetric Laplace errors to take advantage of the available conditional Gaussian representation for these distributions. Further, while the autoregressive model yields conditional quantiles, we demonstrate how to extract marginal quantiles with the asymmetric Laplace specification. Thus, we are able to interpolate quantiles for any days within years across our study region.

1. Introduction. Quantile regression (QR) has a rich history by now, dating to Koenker and Bassett (1978), with much seminal work by Koenker and colleagues (see, e.g., Koenker and Machado (1999), Koenker (2005), Koenker and Xiao (2006)). Many facets are considered in the literature including choice of optimization function (equivalently error distribution), dependence through autoregression, and quantile crossing. We review this literature briefly below. Here, our contribution is to consider QR in the context of a complex spatiotemporal model. This model specifies temporal dependence through autoregression, adopting two time scales, and introduces needed spatial dependence through four Gaussian processes (GPs). We are motivated by mean modeling work developed in Castillo-Mateo et al. (2022) but now seeking quantiles associated with time series of daily *maximum* temperature during the summer season over a period of 60 years. We use data obtained from monitoring stations in the Comunidad Autónoma de Aragón, Spain. Our interest is in extreme heat; specifically, we work with daily maximum temperatures and primarily inferential focus on the $\tau = 0.95$ quantile. Throughout the paper, when we refer to temperature it is a daily maximum temperature. However, our model could be applied to arbitrary quantiles of, for example, daily average temperature or daily minimum temperature.

More precisely, we specify a spatial conditional autoregression model on a daily scale using the asymmetric Laplace (AL) distribution. Our quantile autoregression is an AR(1) form, producing conditional temperature quantiles given the previous day's temperature. The specification enables spatial autoregression at daily and annual scale. We first present the

inference associated with the conditional model, discussing the resulting conditional quantiles and employing model performance assessment by location. The conditional quantiles facilitate assessment of persistence, for example, according to yesterday's temperature, what are today's temperature quantiles? Next, we offer an attractive approach to obtain marginal quantiles at daily scale. The marginal quantiles enable interpolation. We can show the spatial surface for a given marginal quantile. We can also consider averaging to provide marginal quantiles associated with say, 7-day average temperatures. We consider these both spatially and also dynamically. In fact, we show how to provide marginal quantiles associated with 7-day averages over a specified region. We work in a hierarchical Bayesian framework, enabling full posterior inference for all of the quantiles we develop.

There are two modeling approaches for QR in the literature. The first follows the original ideas by Koenker and Bassett (1978) and offers a separate regression model for each of the quantiles of interest. This approach is usually called *multiple* QR, and inference typically proceeds by minimizing a check loss function or assuming an AL error term. Examples of multiple QR with AL errors appear in Yu and Moyeed (2001) while Kozumi and Kobayashi (2011) present a Gibbs sampler for a Bayesian QR model. The second approach, which is usually called *joint* QR, specifies an appropriate joint model for all quantiles (see, e.g., Tokdar and Kadane (2012), Yang and Tokdar (2017), Das and Ghosal (2017a)). Broad implementation for joint QR has proven challenging.

We can also classify the models in terms of whether they incorporate temporal, spatial, or spatiotemporal dependence. Koenker and Xiao (2006) established the basis for joint quantile autoregression (QAR) models in time series. A detailed overview of the different strands of time-series QR modeling can be found in Peters (2018). Recently, spatial quantiles have been an active area of research. Hallin, Lu and Yu (2009) introduce spatial multiple QR that is nonparametric, focusing on asymptotic behavior using assumptions associated with time-series asymptotics. Reich, Fuentes and Dunson (2011) develop a spatial joint QR model that incorporates spatial dependence through spatially varying regression coefficients, which are expressed as a weighted sum of Bernstein basis polynomials where the weights are constrained spatial GPs. Lum and Gelfand (2012) consider spatial multiple QR with AL errors and then extend it to capture spatial dependence by introducing the AL process. Yang and He (2015) consider a nonparametric approach based on Bayesian spatial QR using empirical likelihood as a working likelihood and spatial priors. Chen and Tokdar (2021) specify a spatial joint QR based on the so-called constraint-free reparametrization by generalizing the model of Yang and Tokdar (2017) and characterizing spatial dependence via a Gaussian or t-copula process on the underlying quantile levels of the observation units. Spatiotemporal quantile models are the most challenging and little work has been done in that regard. For example, Reich (2012) follows Reich, Fuentes and Dunson (2011), but allows for residual correlation via a spatiotemporal copula model. Neelon et al. (2015) propose a multiple QR model for areal data. They model the random effects via intrinsic conditionally autoregressive priors, and they adopt the Bayesian approach based on the AL errors. Das and Ghosal (2017b) develop a joint QR model with a single explanatory variable following the representation of quantile functions given by Tokdar and Kadane (2012) and Das and Ghosal (2017a). The explanatory variable is a linear trend over time and spatial dependence is captured by a B-spline basis expansion prior.

The primary advantage of joint QR models is that they avoid the possibility of quantile crossing. This can occur in methods that estimate and infer about quantiles separately. However, joint methods have the disadvantage of restrictive assumptions on covariates and very demanding computation. Further, suppose we work with quantiles of an error distribution such as the AL, that is, we model $Y = \mu_\tau + \epsilon_\tau$ where the distribution for ϵ_τ has zero as the τ quantile. Therefore, $P(\epsilon_\tau \leq 0) = \tau$ so μ_τ provides a τ-QR for Y. If we model in this

way, we impose a *soft* (stochastic) order on the quantiles. More precisely, if we also write $Y = \mu_{\tau^*} + \epsilon_{\tau^*}$ with $\tau < \tau^*$, then ϵ_{τ^*} will be stochastically smaller than ϵ_τ, equivalently μ_{τ^*} will be stochastically larger than μ_τ, yielding the stochastic order.

Recently, QR models have become widely used in climate studies (see, e.g., Haugen et al. (2018), McKinnon and Poppick (2020)). Gao and Franzke (2017) fit a local joint QR to analyze the spatial and temporal pattern of extreme daily temperature. The Bayesian spatiotemporal quantile model in Reich (2012) is used by Tan, Gan and Chen Shu Liu (2019) to identify climate changes in accumulated precipitation in Canada. None of those models include serial dependence in daily variables, but Yang, Li and Xu (2018) propose a semiparametric autoregressive QR model including lagged data to estimate the thresholds to define quantile-based temperature extreme indices.

Our interest focuses on the analysis of the temporal evolution of the distributional changes in the daily maximum temperatures during the summer periods from 1956 to 2015 around the Comunidad Autónoma de Aragón, in the northeast of Spain. The region includes part of the Ebro Valley in the center, with mountainous areas in the south (Iberian System) and the north (Pyrenees). Despite its relatively small size, the region shows a diverse orography, with a warm homogeneous climate in the center and greater climatic variability in the mountainous areas.

Lastly, in this article we discuss many different quantiles—empirical and modeled, marginal and conditional (perhaps conditioned on a marginal quantile). So, we offer some notation to hopefully help in what follows. Our primary model is for a spatial QAR in the form of conditional quantiles. So, we define $Q_V(\tau \mid y)$ as the τ quantile for variable V given y. Hence, for example, $Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$ is the τ conditional quantile for the daily temperature variable for day ℓ in year t at location $\mathbf{s}$ given the previous day's temperature at that site. Also, y might be an empirical quantile as we clarify below. Marginal quantiles, extracted from our spatial QAR (employing adjustment), are denoted by $\tilde{q}_V(\tau)$. So, $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau)$ is the τ marginal quantile for the daily temperature variable for day ℓ in year t at location $\mathbf{s}$. Empirical quantiles are denoted by $q_V^{\mathrm{emp}}(\tau^*)$. They may be indexed by site, averaged over days within a year, years for a given day, or both. When they appear, an explicit definition is clarified in the associated text. They may be used in specifying a conditional quantile and, in this case, τ^* need not equal τ.

The format of the paper is as follows. Section 2 describes the data set with some descriptive work. Section 3 presents our spatial QAR model with the results of the model fitting and some model adequacy assessment. Section 4 develops a strategy for extracting marginal quantiles from our QAR model and then develops interpolation and averaging over time and space for them. Section 5 concludes with a summary and future work.

2. The data. The analyses presented here consider daily maximum temperature (°C) data at $n = 18$ sites around the Comunidad Autónoma de Aragón (see Figure 1) provided by the Agencia Estatal de Meteorología (AEMET) in Spain. The data are available at a daily scale from 1956 to 2015, but the focus of the analyses is in the warm months of June, July, and August (denoted as JJA); in this regard, we fit the models with data in an extended period from May 1 to September 30 to avoid boundary issues.

The region of interest, $\mathcal{D}$, is around Aragón, located in the Ebro Basin in northeastern Spain, with a spatial extent of roughly 53,000 km². In particular, $\mathcal{D}$ has corners at approximately (40.5°N, 1.7°W), (42.9°N, 1.7°W), (42.9°N, 0.0°E), and (40.5°N, 0.0°E). Figure 1 shows the name, location, and elevation of the 18 sites. The Ebro River flows from the northwest to the southeast through a valley bordered by the Pyrenees and the Cantabrian Range in the north and the Iberian System in the southwest. The maximum elevation is around 3000 m in the Pyrenees, 2000 m in the Cantabrian Range and the Iberian System, while the elevation
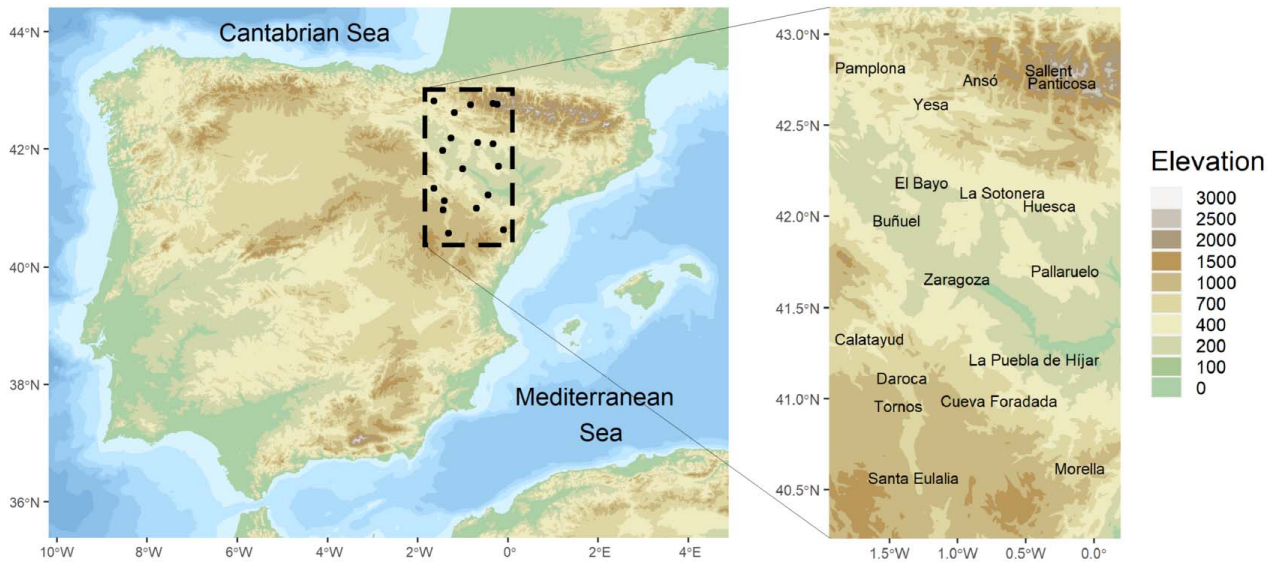
FIG. 1.   *Location of the* 18 *sites around the Comunidad Autónoma de Aragón in northeastern Spain.*

in the Central Valley varies between 200–400 m. In summary, roughly 62% of the area is above 500 m and 28% above 1000 m.

According to AEMET (2011), the central part of the Valley is characterized by a Mediterranean-continental dry climate with irregular rainfall and a large temperature range. However, several climate subareas can be distinguished due to the heterogeneous orography and other influences. Consequently, the region presents a wide variety of climate conditions in a relatively small area, bringing interest in studying it and challenge in modeling it.

With regard to showing maps over this region, $\mathcal{D}$ was partitioned with a resolution of 4 km × 4 km grid cells yielding $K = 2342$ cells.

2.1. *Descriptive analysis.*   Figure 2 describes the distribution across sites of three features related to the empirical quantiles of daily temperatures in JJA months for a grid of quantiles. Each boxplot corresponds to a quantile and is based on 18 points, one point for each observed site. The first plot shows the empirical quantiles calculated with 60 (years) × 92 (days) observations, that is, each one of the 18 points corresponds to the empirical quantile across 60 × 92 observations, and each boxplot is across 18 empirical quantiles from the 18 sites. The range of each boxplot varies around 10°C to 15°C, and the difference between the median of the 0.95 and 0.05 quantiles is around 15°C, indicating a similar variability across sites and across quantiles. The second plot shows the difference between the empirical quantiles of the 30-year periods 1986–2015 and 1956–1985 so that each quantile is now calculated with 30 × 92
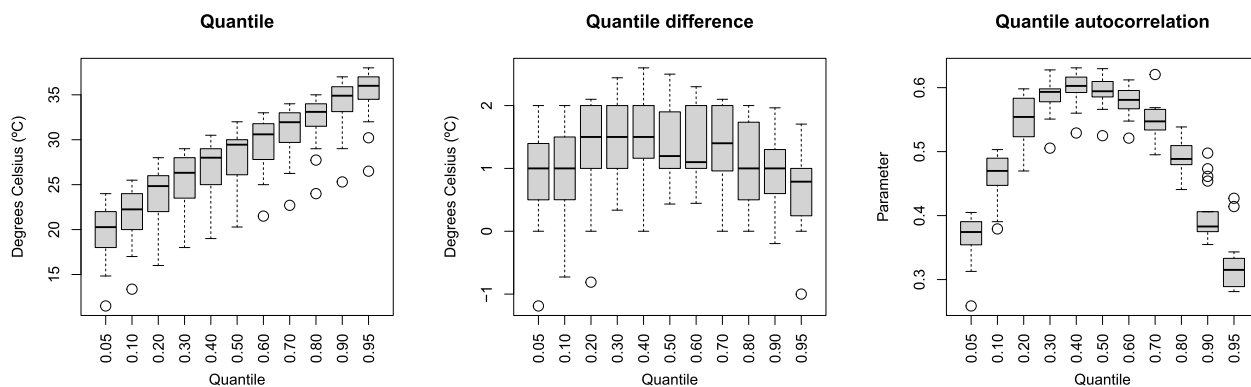

FIG. 2.   *Boxplots of the* 18 *sites across quantiles to describe empirical quantile features.*

observations. Most of the observed increases vary from around $-0.5°C$ to $2.5°C$, but the observed warming is higher in central quantiles. The variability of this increase across quantiles is lower than across sites. The third plot shows the quantile autocorrelation, a measure of quantile dependence based on the correlation between a binary variable indicating whether the temperature in a day is higher than its empirical quantile and the previous day's temperature; see Li, Li and Tsai (2015) and Section S1.1 in the Supplementary Material (Castillo-Mateo et al. (2023)) for the details. This autocorrelation is strong in all sites and quantiles, but it clearly decreases in both tails. The variability across quantiles is larger than across sites. Section S1.1 also includes a descriptive analysis of a second-order quantile autocorrelation, with little or no evidence in favor of including it in the model.

A thorough analysis of the temporal pattern and the effect of elevation and latitude in the quantiles of temperatures is offered in Section S1.2 of the Supplementary Material. This analysis shows that elevation of the site has a roughly negative linear relationship in the empirical quantiles and that this effect is slightly higher in the left tail. As expected, given the size of the region, no latitudinal gradient is observed. Lastly, the data show a strong seasonal component, with no evidence of changes across quantiles and across time. The increase observed between both 30-year periods is homogeneous during the summer, although relevant differences are observed across sites and quantiles.

3. The spatial quantile autoregression model.

3.1. *Review of the asymmetric Laplace distribution.* We introduce the AL distribution as an error distribution for multiple QR models using the following parametrization. We denote by $\epsilon \sim \text{AL}(\mu, \sigma, \tau)$ a random variable with probability density function (pdf),

$$f(\epsilon \mid \mu, \sigma, \tau) = \sigma \tau (1 - \tau) \begin{cases} \exp\{-(1-\tau)\sigma|\epsilon - \mu|\} & \text{if } \epsilon < \mu, \\ \exp\{-\tau\sigma|\epsilon - \mu|\} & \text{if } \epsilon \geq \mu. \end{cases}$$

The cumulative distribution function is

$$F(\epsilon \mid \mu, \sigma, \tau) = \begin{cases} \tau \exp\{-(1-\tau)\sigma|\epsilon - \mu|\} & \text{if } \epsilon < \mu, \\ 1 - (1-\tau)\exp\{-\tau\sigma|\epsilon - \mu|\} & \text{if } \epsilon \geq \mu. \end{cases}$$

Here, μ is a location parameter, $\sigma > 0$ is a scale parameter, and $\tau \in (0, 1)$ is an asymmetry parameter. In particular, it is easily checked that μ is the τ quantile of the distribution and we will typically set $\mu = 0$ so that $P(\epsilon \leq 0) = \tau$.

The pdf above can be rewritten as $f(\epsilon \mid \mu, \sigma, \tau) = \sigma \tau (1 - \tau) \exp\{-\sigma \delta_\tau(\epsilon - \mu)\}$ where $\delta_\tau(u) = u(\tau - \mathbf{1}(u < 0))$ is the *check loss* function (Koenker and Bassett (1978)). For a sample $\{x_i : i = 1, \ldots, n\}$, finding $\arg\min_\mu \sum \delta_\tau(x_i - \mu)$ returns the τ empirical quantile. Just as minimizing the sum of squares loss is associated with normal errors, minimizing check loss is associated with AL errors.

A convenient strategy for generating $\epsilon \sim \text{AL}(0, \sigma, \tau)$ variables is to use the following representation proven by comparing moment generating functions (see, e.g., Kotz, Kozubowski and Podgórski (2001)). We can express ϵ in terms of

$$\epsilon = \sqrt{\frac{2U}{\sigma^2 \tau (1-\tau)}} Z + \frac{1 - 2\tau}{\sigma \tau (1-\tau)} U,$$

where $Z \sim N(0, 1)$ and $U \sim \text{Exp}(1)$. So,

$$(1) \qquad \epsilon \mid \sigma, U \sim N\left(\frac{1 - 2\tau}{\sigma \tau (1-\tau)} U, \frac{2U}{\sigma^2 \tau (1-\tau)}\right)$$

is normally distributed enabling us to use all of the familiar Gaussian theory.

3.2. *The space-time model.* Let $\tau \in (0, 1)$ denote a quantile order, where each quantile is modeled separately. Our general form for a spatiotemporal τ-QAR with two time scales is given by

$$
\begin{aligned}
Y_{t\ell}(\mathbf{s}) &= Q_{Y_{t\ell}(\mathbf{s})}\big(\tau \mid Y_{t,\ell-1}(\mathbf{s})\big) + \epsilon_{t\ell}^{\tau}(\mathbf{s}) \\
&= q_{t\ell}^{\tau}(\mathbf{s}) + \rho^{\tau}(\mathbf{s})\big(Y_{t,\ell-1}(\mathbf{s}) - q_{t,\ell-1}^{\tau}(\mathbf{s})\big) + \epsilon_{t\ell}^{\tau}(\mathbf{s}),
\end{aligned}
$$

(2)

where $Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$ is the τ conditional quantile of $Y_{t\ell}(\mathbf{s})$ given $Y_{t,\ell-1}(\mathbf{s})$ and the error term is $\epsilon_{t\ell}^{\tau}(\mathbf{s}) \sim$ ind. $\mathrm{AL}(0, \sigma^{\tau}(\mathbf{s}), \tau)$. Here, $q_{t\ell}^{\tau}(\mathbf{s})$ contains fixed and random effects as below. In addition, $\rho^{\tau}(\mathbf{s})$ is a spatially varying autoregression coefficient and $\sigma^{\tau}(\mathbf{s})$ is a spatially varying pure error scale parameter at location $\mathbf{s}$.

Based upon the foregoing exploratory analysis along with that developed in Castillo-Mateo et al. (2022), we adopt an analogue of their spatiotemporal mean autoregression model. Here, $Y_{t\ell}(\mathbf{s})$ denotes the daily maximum temperature for day ℓ, $\ell = 2, \ldots, L$ of year t, $t = 1, \ldots, T$ at location $\mathbf{s}$, $\mathbf{s} \in \mathcal{D}$, the study region. We specify $\rho^{\tau}(\mathbf{s})$ to capture spatial autoregession dependence through the GP $Z_{\rho}^{\tau}(\mathbf{s}) = \log\{(1 + \rho^{\tau}(\mathbf{s}))/(1 - \rho^{\tau}(\mathbf{s}))\}$ with mean Z_{ρ}^{τ} and exponential covariance function having variance parameter $\sigma_{\rho}^{2,\tau}$ and decay parameter ϕ_{ρ}^{τ}. In the same manner, we specify $\sigma^{\tau}(\mathbf{s})$ to capture spatial scale dependence through the GP $Z_{\sigma}^{\tau}(\mathbf{s}) = \log\{\sigma^{\tau}(\mathbf{s})\}$ with mean Z_{σ}^{τ} and exponential covariance function having variance parameter $\sigma_{\sigma}^{2,\tau}$ and decay parameter ϕ_{σ}^{τ}.

As for $q_{t\ell}^{\tau}(\mathbf{s})$, we adopt

$$
q_{t\ell}^{\tau}(\mathbf{s}) = \beta_0^{\tau} + \alpha^{\tau}t + \beta_1^{\tau}\sin(2\pi\ell/365) + \beta_2^{\tau}\cos(2\pi\ell/365) + \beta_3^{\tau}\,\mathrm{elev}(\mathbf{s}) + \gamma_t^{\tau}(\mathbf{s}),
$$

where $\gamma_t^{\tau}(\mathbf{s}) = \beta_0^{\tau}(\mathbf{s}) + \alpha^{\tau}(\mathbf{s})t + \psi_t^{\tau} + \eta_t^{\tau}(\mathbf{s})$. The *fixed effects* are given by β_0^{τ}, a global intercept, $\alpha^{\tau}t$, a global long-term linear trend, sin and cos terms that provide the annual seasonal component, and $\mathrm{elev}(\mathbf{s})$, the elevation at $\mathbf{s}$. The *random effects* given by $\gamma_t^{\tau}(\mathbf{s})$ capture space-time dependence through GPs. In particular, $\beta_0^{\tau}(\mathbf{s})$ is a GP with zero mean and exponential covariance function having variance parameter $\sigma_{\beta_0}^{2,\tau}$ and decay parameter $\phi_{\beta_0}^{\tau}$, and it provides a local spatial adjustment to the intercept. The $\alpha^{\tau}(\mathbf{s})$ are a GP, with zero mean and exponential covariance function having variance parameter $\sigma_{\alpha}^{2,\tau}$ and decay parameter ϕ_{α}^{τ}, to provide a local slope adjustment to the linear trend. Together, $\gamma_t(\mathbf{s})$ supplies a *locally linear* trend, an exceptionally rich spatial specification. With the inclusion of seasonality, it is difficult to imagine that the data could inform about a higher-order local choice. Continuing, $\psi_t^{\tau} \sim$ i.i.d. $N(0, \sigma_{\psi}^{2,\tau})$ provides annual intercepts to allow for yearly shifts (i.e., for hotter or colder years),[1] and $\eta_t^{\tau}(\mathbf{s}) \sim$ i.i.d. $N(0, \sigma_{\eta}^{2,\tau})$ provides local annual intercepts to allow for local yearly shifts.

We make two further points here. Focusing on performance of conditional quantiles, we have investigated departures from our first-order regression but have found no reason to adopt a more elaborate ARMA specification. In the exploratory and residual analysis, no evidence of second-order correlation has been found while dependence explained by MA terms becomes confusing with regard to our conditional objective. Second, we do not introduce spatial variability in the seasonality in our study. We are interested in the warmest time of the year, JJA. Our region is small enough to assume that the daily amplitude of solar incidence, that is, seasonality, during JJA is almost equivalent between the northernmost and southernmost sites. A brief residual analysis to validate the seasonal term used in the model is given in Section S2 of the Supplementary Material.

---

[1]An autoregression could be considered for ψ_t^{τ}. However, Castillo-Mateo et al. (2022) found no such autocorrelation in their means model.

We conclude with a technical remark regarding the validity of the Bayesian analysis for individual quantiles based on the working AL likelihood since this likelihood differs from that of the data generating process. Sriram, Ramamoorthi and Ghosh (2013) established sufficient conditions for the posterior consistency of model parameters under the AL working likelihood (at a single quantile level). The situation is more complicated when multiple quantile levels are considered, as this means different likelihoods will be utilized in each of the Bayesian analyses. When the AL likelihood differs from the data generating process, there is some literature providing posterior adjustments under the linear quantile regression model for independent data (see, e.g., Chernozhukov and Hong (2003), Yang, Wang and He (2016)). While it is useful to be aware of these issues, addressing them is beyond the scope of our complex spatiotemporal modeling employed here.

3.2.1. *Prior distributions and model fitting.* Model inference is implemented in a Bayesian framework. To complete the model, we specify prior distributions for all model parameters. In this setting, diffuse and, when available, conditionally conjugate prior distributions are chosen. Recall that the model adopts a conditional AL distribution for all $Y_{t\ell}(\mathbf{s})$, and that this distribution can be expressed as normal when it is conditioned on $U_{t\ell}^\tau(\mathbf{s}) \sim \text{Exp}(1)$. Therefore, the coefficient parameters β_0^τ, α^τ, β_1^τ, β_2^τ, and β_3^τ are each assigned independent normal prior distributions with mean 0 and standard deviation 100. The variance parameters, $\sigma_\psi^{2,\tau}$ and $\sigma_\eta^{2,\tau}$, are assigned independent Inverse-Gamma(0.1, 0.1) prior distribution.

The specification of the GPs is as follows. First, Z_ρ and Z_σ are each given a normal prior distribution with mean 0 and standard deviation 100. Second, the variance parameter for each of the four spatial covariances, $\sigma_{\beta_0}^{2,\tau}$, $\sigma_\alpha^{2,\tau}$, $\sigma_\rho^{2,\tau}$, and $\sigma_\sigma^{2,\tau}$, is assigned an independent Inverse-Gamma(0.1, 0.1) prior distribution. With an exponential covariance function, the product of the variance and the decay parameter is identified but the individual parameters are not (Zhang (2004)). With stronger interest in the spatial variability, we adopt weak priors and let the data inform about the σ^2's. We are more precise with regard to the decay parameters. In fact, with information well informed by the spatial scale of our study region, we fix the decay parameters $\phi \equiv \phi_{\beta_0}^\tau = \phi_\alpha^\tau = \phi_\rho^\tau = \phi_\sigma^\tau = 3/d_{\max}$, where $d_{\max}$ is the maximum distance between any pair of spatial locations. That is, with an exponential covariance function, the decay parameter is 3/range, and it is set to the value associated with the largest spatial range for the observed data locations.

We develop a Metropolis-within-Gibbs algorithm to obtain Markov chain Monte Carlo (MCMC) samples from the joint posterior distribution. In particular, we derive full conditional distributions for each of the parameters, including the $n \times T \times (L-1)$ latent exponential variables $\xi_{t\ell}^\tau(\mathbf{s}) = U_{t\ell}^\tau(\mathbf{s})/\sigma^\tau(\mathbf{s})$. This parametrization is adopted to obtain a gamma full conditional distribution for the scale parameter in the AL (Kozumi and Kobayashi (2011)). However, in our case, with a GP prior for the log-scale parameter, the full conditional is still nonstandard so we do not benefit from this parametrization. For the fitting, we introduce $\tilde{\beta}_0^\tau(\mathbf{s}) = \beta_0^\tau + \beta_0^\tau(\mathbf{s})$ and $\tilde{\alpha}^\tau(\mathbf{s}) = \alpha^\tau + \alpha^\tau(\mathbf{s})$ within $\gamma_t^\tau(\mathbf{s})$ to enable the benefits of hierarchical centering in the model fitting (Gelfand, Sahu and Carlin (1995)). Details of the Gibbs sampler used for the model fitting are provided in Section S3 of the Supplementary Material. All the covariates have been centered and scaled to have mean zero and standard deviation one to improve the mixing of the algorithm.

3.3. *Results of model fitting.* This section summarizes the results of the QAR models for $\tau \in \{0.05, 0.10, 0.20, \ldots, 0.80, 0.90, 0.95\}$ quantiles fitted to the temperature series described in Section 2. The parameters α^τ, β_1^τ, β_2^τ, β_3^τ, $\alpha^\tau(\mathbf{s})$, and σ_α^τ have been rescaled to interpret them in terms of the original scales of the covariates. For each τ, using MCMC and the Gibbs sampling algorithm (see Section S3), we ran three chains, each with different initial
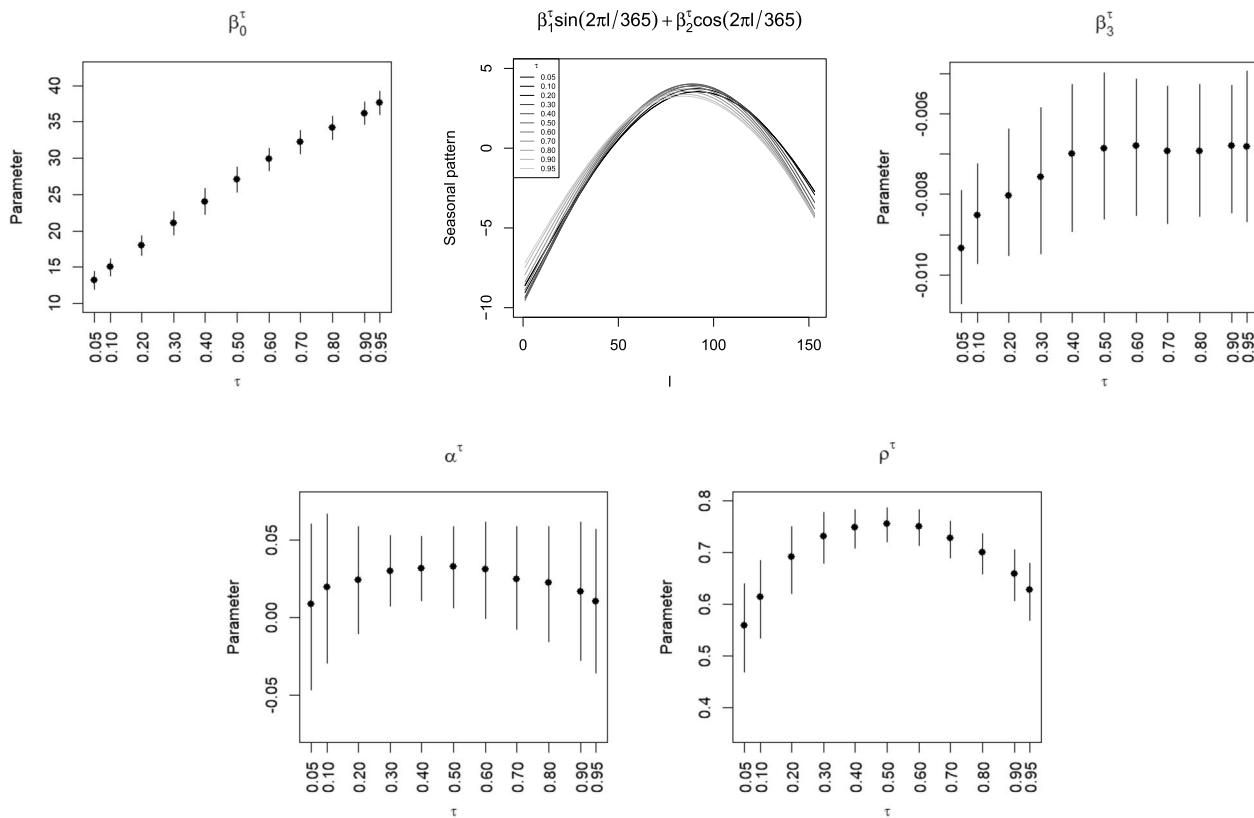
FIG. 3.   *Posterior median and 90% credible interval of the main parameters and posterior mean of the centered seasonal pattern captured by the harmonic terms (first row central plot) across quantiles.*

values, out to 1,000,000 iterations for each chain, to obtain samples from the joint posterior distribution. The first 100,000 samples were discarded as burn-in and the remaining 900,000 samples were thinned to retain 1000 samples from each chain for posterior inference. Experiments were run on a computer with an Intel Core i9-10900K processor running at 3.70 GHz using 64 GB of RAM, running Windows 10 Pro version 21H2. Under this setup, fitting eleven models in parallel, we were able to fit the 3 (chains) $\times$ 11 (τ's) models in 3×36 hours. Convergence was monitored by usual trace plots (not shown), and the marginal and multivariate potential scale reduction factors (Brooks and Gelman (1998)).

Figure 3 shows a summary across quantiles for the fixed effects parameters and for ρ^τ, the mean of the spatial process $\rho^\tau(\mathbf{s})$. Figure S7 of the Supplementary Material provides the rest of the model parameters. As expected, β_0^τ increases monotonically with the quantiles, but it is not a linear function of τ since the slope varies from $\tau = 0.60$. The seasonal pattern obtained from the harmonic terms is very similar across the τ's. The elevation coefficient β_3^τ is close to $-7°C/km$, the environmental lapse rate (Navarro-Serrano et al. (2018)), for τ greater than or equal to 0.40. This value decreases below $-9°C/km$ for τ's close to the extreme cold. The posterior median of α^τ is above $0.30°C/decade$ in the central quantiles, close to the trend of $0.27°C/decade$ estimated by Peña-Angulo et al. (2021) for the daily temperature in Spain in the summer period 1956–2015, and decreases to about $0.10°C/decade$ in the tails. The results of ρ^τ show a strong autoregression in all the quantiles, varying from 0.55 for $\tau = 0.05$ to 0.75 in the central quantiles.

Figure 4 shows the posterior mean of the spatial random effects, $\tilde{\beta}_0^\tau(\mathbf{s})$, $\tilde{\alpha}^\tau(\mathbf{s})$, $\rho^\tau(\mathbf{s})$, and $\sigma^\tau(\mathbf{s})$ at each observed site across quantiles. Figure 5 shows spatially the posterior mean of the previous spatial processes (showing $\beta_0^\tau(\mathbf{s})$ instead of $\tilde{\beta}_0^\tau(\mathbf{s})$) for $\tau = 0.05, 0.50, 0.95$. See also Figure S8 of the Supplementary Material for posterior boxplots of these processes at the observed sites and Figure S9 to get an idea of the uncertainties of these estimates. These figures show wide spatial variability in the four spatial processes at all quantiles. In particular,
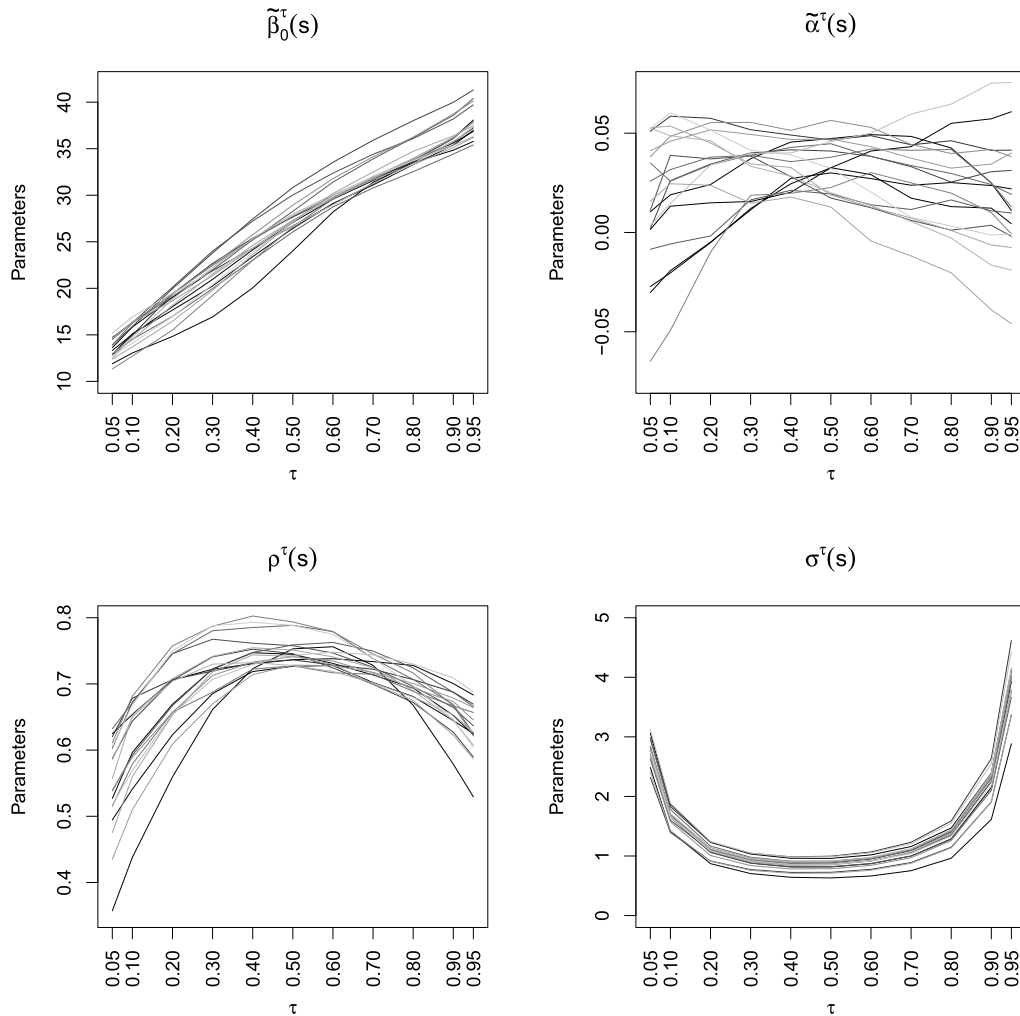
FIG. 4. *Posterior mean of the four spatial random effects in the* 18 *observed sites across quantiles.*

the spatially varying intercept $\beta_0^\tau(\mathbf{s})$ captures climate variability not explained by elevation. The similarity of the spatial pattern for $\beta_0^{0.50}(\mathbf{s})$ and $\beta_0^{0.95}(\mathbf{s})$ versus $\beta_0^{0.05}(\mathbf{s})$ is noteworthy. There are clear differences across quantiles in the spatial pattern of the linear trend $\tilde{\alpha}^\tau(\mathbf{s})$. For $\tau = 0.50$, the trends are positive in all of the regions while negative values are observed in the northwest for $\tau = 0.05$ and in the central part and northeast for $\tau = 0.95$. The areas where the 80% credible intervals do not contain a null trend are shown in Figure S10 of the Supplementary Material. The range and spatial pattern of the autoregression term $\rho^\tau(\mathbf{s})$ also vary across quantiles. The posterior mean varies from 0.72 to 0.80 in $\tau = 0.50$, from 0.53 to 0.69 in $\tau = 0.95$, and from 0.36 to 0.64 in $\tau = 0.05$.

The random evolution across years of the posterior distribution of the temporal random effects ψ_t^τ (see Figure S11 in the Supplementary Material) confirms that a more complex trend would not improve the fit of the linear trend. Again, we note the similarity of the distribution of $\psi_t^{0.50}$ and $\psi_t^{0.95}$, while $\psi_t^{0.05}$ is slightly different and has a greater variance.

3.3.1. *Conditional quantiles.* As an illustration of the output provided by the model, Figure 6 shows the posterior mean of the conditional quantiles $Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid y)$ on July 15, 2015 ($t = 60$ and $\ell = 76$) under three different situations for y and for $\tau = 0.05, 0.50, 0.95$. An important feature of daily temperature is its high persistence, evidenced by the autoregression coefficients $\rho^\tau(\mathbf{s})$ of the model. The selected values of y correspond to location dependent cold, mild, and warm situations, that is, we use local empirical quantiles $q_{Y(\mathbf{s})}^{\text{emp}}(\tau^*)$ for $\tau^* = 0.05, 0.50, 0.95$. First, the empirical quantiles at the observed sites are obtained using the
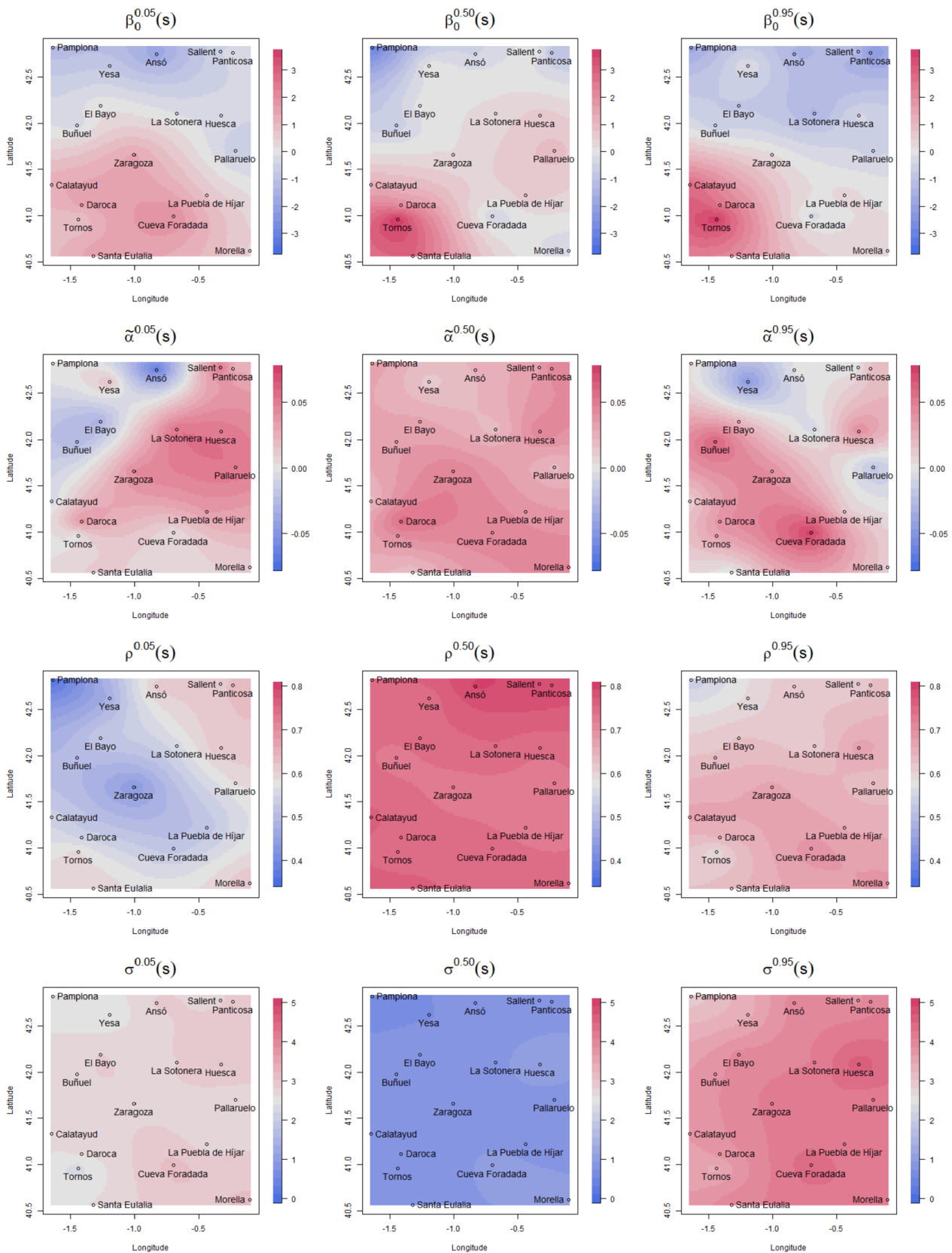
FIG. 5. *Maps of the posterior mean of the four spatial random effects for $\tau = 0.05, 0.50, 0.95$.*

30×92 observations of the JJA months in the reference period 1981–2010. Later, to obtain a value for each **s**, the observed values are interpolated by means of a simple kriging.

Though the posterior mean level changes, the spatial pattern in all the conditional quantiles is similar. However, as a consequence of the different persistence across quantiles, the posterior mean of the difference between $Q_{Y_{t\ell}(\mathbf{s})}(0.50 \mid y)$ and $Q_{Y_{t\ell}(\mathbf{s})}(0.05 \mid y)$ varies noticeably depending on the value of the previous day's temperature. For example, in Zaragoza, it is
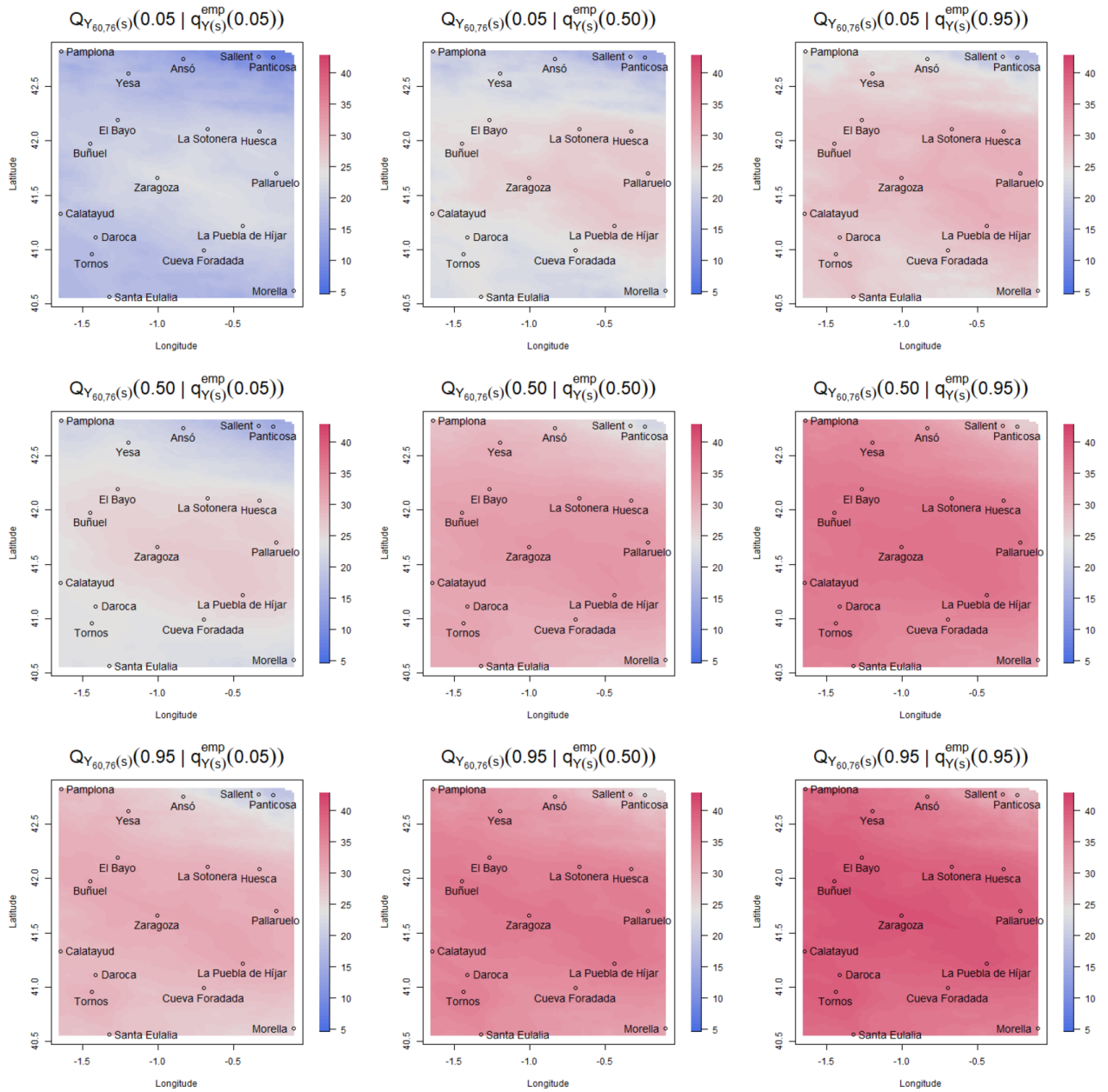
FIG. 6. *Maps of the posterior mean of the conditional quantiles $Q_{Y_{60,76}(\mathbf{s})}(\tau \mid q_{Y(\mathbf{s})}^{emp}(\tau^*))$ on July 15, 2015, for $\tau, \tau^* = 0.05, 0.50, 0.95$.*

2.8°C, 5.7°C and 8.6°C for $y = 20, 30, 40$°C, respectively. Analogously, the posterior mean of the difference between $Q_{Y_{t\ell}(\mathbf{s})}(0.95 \mid y)$ and $Q_{Y_{t\ell}(\mathbf{s})}(0.50 \mid y)$ is 4.3°C, 3.7°C, and 3.1°C for the previous values of y. Another consequence is that an increase of around 1°C in the previous day's temperature, y, yields an increase of around 0.44°C in the posterior mean of the 0.05 conditional quantile, 0.73°C in the conditional median, and around 0.66°C in the 0.95 conditional quantile.

3.4. *Model assessment through cross-validation.* Here, we take up model assessment in the context of performance across the n locations. That is, we are not implementing model comparison; rather, we are looking at local and global adequacy of the model employing three different quantiles, $\tau = 0.05, 0.50, 0.95$. In particular, a leave-one-out cross-validation is carried out. That is, each site is held out one at a time and subsequently the model is fitted with the remaining $n-1$ sites. Then the conditional quantiles are obtained using one-step ahead prediction since, for held-out site, we know the $Y_{t,\ell-1}(\mathbf{s})$ to condition on.

We consider two types of residuals. The first type is defined as $R_{t\ell}^{(b)}(\tau; \mathbf{s}) = Y_{t\ell}(\mathbf{s}) - Q_{Y_{t\ell}(\mathbf{s})}^{(b)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$ for $b = 1, 2, \ldots, B$, where b denotes an MCMC posterior realization of the τ conditional quantile on day ℓ, year t at site $\mathbf{s}$. A simpler type based on the mean of those realizations yields the single residual value $R_{t\ell}(\tau; \mathbf{s}) = Y_{t\ell}(\mathbf{s}) - E[Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s})) \mid \text{data}]$. The simpler version only looks at $Y_{t\ell}(\mathbf{s})$ relative to the associated mean of the conditional quantile. The replicated version looks at $Y_{t\ell}(\mathbf{s})$ relative to the distribution of the associated conditional quantile.

Now, we define three measures of model assessment. First, we propose an approximation to the probability (an integral) that the observed $Y_{t\ell}(\mathbf{s})$ is less than the conditional quantile on day ℓ, year t, site $\mathbf{s}$. This probability is estimated as $p_{t\ell}(\tau; \mathbf{s}) = \sum_b \mathbf{1}(R_{t\ell}^{(b)}(\tau; \mathbf{s}) < 0)/B$. The mean value of these probabilities should be τ. As a result, a global measure is defined as

$$p(\tau) = \frac{1}{nT(L-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{\ell=2}^{L} p_{t\ell}(\tau; \mathbf{s}_i).$$

Under an adequate model, this should take a value close to τ. Analogous versions without averaging over days, years, or sites are denoted by $p_\ell(\tau)$, $p_t(\tau)$, or $p(\tau; \mathbf{s})$, respectively.

Second, to evaluate potential overfit in terms of out of sample performance, we employ the asymmetrically τ-weighted mean absolute error (Koenker and Bassett (1978)), which we denote by WMAE(τ). Using the check loss function $\delta_\tau(u) = u(\tau - \mathbf{1}(u < 0))$ defined in Section 3.1, this measure is given by

$$\text{WMAE}(\tau) = \frac{1}{nT(L-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{\ell=2}^{L} \delta_\tau\big(R_{t\ell}(\tau; \mathbf{s}_i)\big).$$

Equivalently, this measure calculates the mean value of the absolute errors weighted by $1 - \tau$ if $R_{t\ell}(\tau; \mathbf{s}) < 0$ or τ otherwise. The smaller its value, the better the model performance. The analogous site-level version is denoted by WMAE($\tau; \mathbf{s}$).

The third measure is calculated as $R^1(\tau)$ according to Koenker and Machado (1999). This goodness-of-fit measure is viewed as an analogue of R^2 for the classical residual sum of squares; the check loss function for quantiles replaces the least-squares loss function and the τ empirical quantile $q_{Y(\mathbf{s})}^{\text{emp}}(\tau)$ replaces the sample mean. In this section, $q_{Y(\mathbf{s})}^{\text{emp}}(\tau)$ is calculated with the $T \times (L-1)$ observations of the held-out site. Thus, this measure is given by

$$R^1(\tau) = 1 - \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{\ell=2}^{L} \delta_\tau(R_{t\ell}(\tau; \mathbf{s}_i))}{\sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{\ell=2}^{L} \delta_\tau(Y_{t\ell}(\mathbf{s}_i) - q_{Y(\mathbf{s}_i)}^{\text{emp}}(\tau))}.$$

Note that in-sample, but not out-of-sample, $R^1(\tau)$ would fall between 0 and 1. In both cases, it measures the relative success of the corresponding QR models at a specific quantile in terms of an appropriately weighted sum of absolute residuals. Thus, $R^1(\tau)$ provides a local measure of goodness-of-fit for a particular quantile rather than a global measure of goodness-of-fit over the entire conditional distribution. The analogous site-level version is denoted by $R^1(\tau; \mathbf{s})$.

Table 1 shows global performance metrics and Table 2 shows the results by site. The performance is good for the three values of τ according to $p(\tau)$. Further, for most of the stations, $p(\tau; \mathbf{s})$ is within $\tau \pm 0.02$. The goodness-of-fit considering $R^1(\tau)$ is around 0.367 for $\tau = 0.05$, with values above 0.4 in some sites. This criteria is around 0.464 and 0.442 for $\tau = 0.50$ and 0.95, respectively, but it is above 0.5 in some sites. Quantiles 0.50 and 0.95 perform better than the left tail.

TABLE 1
Performance metrics for the models with $\tau = 0.05, 0.50, 0.95$

τ	$p(\tau)$	WMAE(τ)	$R^1(\tau)$
0.05	0.056	0.378	0.367
0.50	0.505	1.192	0.464
0.95	0.944	0.272	0.442

Less accurate results appear in sites located in the western part of the region. Tornos has a climate that is affected by plateau conditions, and presents the worst values in $p(\tau; \mathbf{s})$ and $R^1(\tau; \mathbf{s})$ for $\tau = 0.95$. Pamplona has colder conditions related to Cantabrian Sea effects, and the actual $\tau = 0.05$ quantile is much colder than expected by the model. Furthermore, its median is the most poorly captured across the sites. The performance of the other sites is reasonably homogeneous based on the three measures for the three values of τ.

Figure 7 shows the evolution of $p_t(\tau)$ and $p_\ell(\tau)$ across t and ℓ, respectively. It reveals that the temporal evolution of days-within-years and years is well captured and there seems to be no bias in the estimates.

4. Marginal quantiles. Here, we present a general strategy for obtaining marginal quantiles from the conditional quantiles. Implementation is straightforward working with AL errors, employing the conditional Gaussian representation.

Marginal quantiles enjoy direct interpretation as well as the benefit of spatial interpolation. That is, to be locally appropriate, conditional quantiles would require the local previous day's temperature, which will not be available at unobserved locations.[2] We present the approach

TABLE 2
Performance metrics for models with $\tau = 0.05, 0.50, 0.95$ for the 18 sites

Location	$p(\tau; \mathbf{s})$			WMAE$(\tau; \mathbf{s})$			$R^1(\tau; \mathbf{s})$		
	0.05	0.50	0.95	0.05	0.50	0.95	0.05	0.50	0.95
Pamplona	0.140	0.582	0.933	0.501	1.619	0.359	0.144	0.333	0.369
Buñuel	0.069	0.519	0.954	0.368	1.227	0.270	0.345	0.422	0.430
El Bayo	0.036	0.485	0.952	0.367	1.162	0.262	0.361	0.457	0.444
Morella	0.035	0.531	0.953	0.342	1.046	0.250	0.418	0.488	0.460
Huesca	0.040	0.485	0.954	0.341	1.005	0.225	0.398	0.520	0.515
Tornos	0.051	0.446	0.876	0.436	1.392	0.349	0.386	0.458	0.347
Santa Eulalia	0.037	0.463	0.951	0.390	1.137	0.257	0.393	0.510	0.480
Calatayud	0.044	0.504	0.944	0.372	1.221	0.279	0.411	0.473	0.451
Panticosa	0.042	0.568	0.976	0.346	1.110	0.267	0.402	0.491	0.419
La Puebla de Híjar	0.063	0.465	0.925	0.377	1.118	0.252	0.342	0.462	0.449
Ansó	0.045	0.524	0.969	0.370	1.160	0.262	0.392	0.490	0.463
Daroca	0.053	0.523	0.968	0.380	1.275	0.274	0.394	0.454	0.456
Zaragoza	0.066	0.505	0.957	0.365	1.176	0.244	0.350	0.446	0.484
La Sotonera	0.055	0.506	0.941	0.361	1.094	0.251	0.378	0.493	0.470
Pallaruelo	0.058	0.481	0.932	0.373	1.106	0.250	0.375	0.488	0.463
Cueva Foradada	0.046	0.573	0.976	0.326	1.027	0.246	0.396	0.477	0.452
Sallent de Gállego	0.060	0.465	0.914	0.373	1.167	0.276	0.389	0.487	0.429
Yesa	0.062	0.464	0.910	0.423	1.414	0.316	0.324	0.416	0.411

---

[2]Of course, we can always interpolate conditional quantiles given illustrative choices of previous day's temperature.
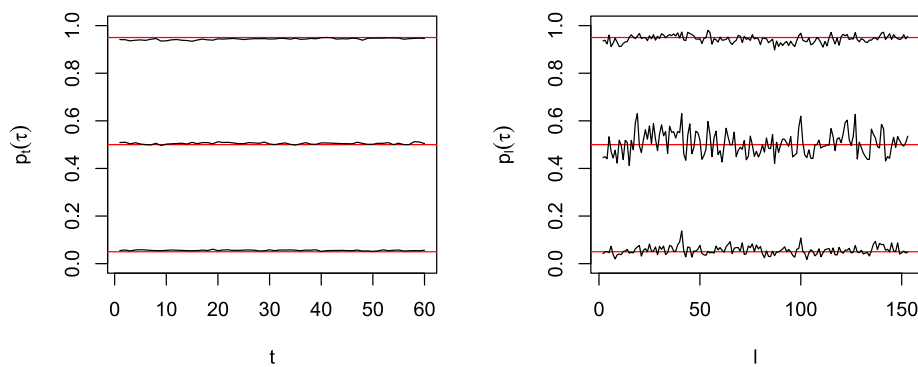
FIG. 7.   *Evolution of $p_t(\tau)$ across t (left) and $p_\ell(\tau)$ across ℓ (right) for $\tau = 0.05, 0.50, 0.95$.*

for extracting a marginal quantile from a conditional quantile *after* fitting the conditional quantile model. In the setting of an AL specification for the errors, first we show how to implement this procedure for a single day. Next, we clarify how the spatial kriging is implemented using these marginal quantiles and supply a spatial map of the marginal τ-quantile temperature surface in our study region for a given day. Then we move to the marginal quantile associated with an average over days. Such averaging is helpful in providing say, weekly quantiles; in our application, 7-day average temperatures quantiles in the summer may be useful, particularly for comparison over years. Such quantiles can be kriged over a spatial region to reveal the *quantile surface*. Lastly, such a daily or weekly quantile surface can be averaged over a region of interest to obtain areal daily or weekly quantiles. The tool here is block averaging (Banerjee, Carlin and Gelfand (2015), Chapter 7).

4.1. *Obtaining marginal quantiles from conditional quantiles in an autoregression.*   Considering expression (2), it is attractive to think about $q_{t\ell}^\tau(\mathbf{s})$ as a version of a marginal τ quantile for $Y_{t\ell}(\mathbf{s})$. However, $P(Y_{t\ell}(\mathbf{s}) \le q_{t\ell}^\tau(\mathbf{s})) \ne \tau$. So, we seek an additive adjustment to $q_{t\ell}^\tau(\mathbf{s})$, which depends upon the model for $Y_{t\ell}(\mathbf{s})$, so that it adjusts this probability to τ. Then $q_{t\ell}^\tau(\mathbf{s})$ plus this adjustment becomes the marginal τ quantile we want.

To present the idea in its simplest form, we ignore space and years and suppress the superscript τ in the parameters. So, we have $Y_\ell = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1}) + \epsilon_\ell$ where the $\epsilon_\ell \sim$ i.i.d. AL$(0, \sigma, \tau)$. In this notation, $Q_{Y_\ell}(\tau \mid Y_{\ell-1}) = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1})$ is the τ quantile of the QAR. For convenience, write this model as $W_\ell = \rho W_{\ell-1} + \epsilon_\ell$ with $W_\ell = Y_\ell - q_\ell$. Upon substitution, we have $W_\ell = \rho^\ell W_0 + \sum_j \rho^j \epsilon_{\ell-j}$. Using the conditional normal form for ϵ_ℓ in (1), we have

$$\tilde\epsilon_\ell \mid \rho, \sigma, U_\ell, U_{\ell-1}, \ldots, U_1 \equiv \sum_{j=0}^{\ell-1} \rho^j \epsilon_{\ell-j} \mid \rho, \sigma, U_\ell, U_{\ell-1}, \ldots, U_1$$

$$\sim N\left(\frac{1-2\tau}{\sigma\tau(1-\tau)} \sum_{j=0}^{\ell-1} \rho^j U_{\ell-j}, \; \frac{2}{\sigma^2\tau(1-\tau)} \sum_{j=0}^{\ell-1} \rho^{2j} U_{\ell-j} \right).$$

We want the τ quantile of W_ℓ, call it $d_\ell^\tau(\rho, \sigma)$, so that $W_\ell - d_\ell^\tau(\rho, \sigma)$ has 0 as its τ quantile and, therefore, Y_ℓ has $q_\ell + d_\ell^\tau(\rho, \sigma)$ as the τ marginal quantile. The τ quantile of W_ℓ is $\rho^\ell W_0$ plus the τ quantile of $\tilde\epsilon_\ell$. While 0 is the τ quantile of ϵ_ℓ, the τ quantile of $\tilde\epsilon_\ell$ will not be 0. That is, each term in $\tilde\epsilon_\ell$ has τ quantile 0 but the sum will not.

Though $\tilde\epsilon_\ell$ does not have an AL distribution, we can find its τ quantile. For any d, we seek $P(\tilde\epsilon_\ell < d \mid \rho, \sigma)$. However,

$$P(\tilde\epsilon_\ell < d \mid \rho, \sigma)$$

$$= \int \int \cdots \int P(\tilde\epsilon_\ell < d \mid \rho, \sigma, \{U_j : j = 1, 2, \ldots, \ell\})[\{U_j\}] \, dU_1 \, dU_2 \cdots dU_\ell.$$

But given $\{U_j : j = 1, 2, \ldots, \ell\}$, we have the distribution for $\tilde{\epsilon}_\ell$ above. So, we can calculate $P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma, \{U_j : j = 1, 2, \ldots, \ell\})$. In fact, we can do a Monte Carlo integration to calculate $P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma)$ by generating many sets $\{U_j : j = 1, 2, \ldots, \ell\}$, all i.i.d., all distributed as $\mathrm{Exp}(1)$. We can do this for any value d, in fact, using the same Monte Carlo samples. Then, using a simple search, we can find $d_\ell^\tau(\rho, \sigma)$. Moreover, we can use the same Monte Carlo samples for any ρ and σ so computation is not demanding.

In our modeling setting, we need $d_{t\ell}^\tau(\rho^\tau(\mathbf{s}), \sigma^\tau(\mathbf{s}))$ to accompany $q_{t\ell}^\tau(\mathbf{s})$. From the model fitting described in Section 3.2.1, we have random samples of $\rho^\tau(\mathbf{s})$ and $\sigma^\tau(\mathbf{s})$, which can produce random samples of $d_{t\ell}^\tau(\rho^\tau(\mathbf{s}), \sigma^\tau(\mathbf{s}))$. These can be used with posterior samples of $q_{t\ell}^\tau(\mathbf{s})$ to create the posterior distribution of the τ marginal quantile for any year, day, and site. In the sequel, we denote this as $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) \equiv q_{t\ell}^\tau(\mathbf{s}) + d_{t\ell}^\tau(\rho^\tau(\mathbf{s}), \sigma^\tau(\mathbf{s}))$. Again, we can use the same sets of $\{U_j\}$'s. Moreover, we can do this for any τ. Generating the entire collection of marginal quantiles of interest is straightforward.

4.2. *Kriging marginal quantiles and block averaging of marginal spatial quantiles.* As above, for a given τ, year, and day within year, marginal quantiles enable kriging to unobserved locations. More precisely, at new site $\mathbf{s}_0$, we can obtain the posterior distribution of $\tilde{q}_{Y_{t\ell}(\mathbf{s}_0)}(\tau)$. To obtain a sample of $\tilde{q}_{Y_{t\ell}(\mathbf{s}_0)}(\tau)$, we need a sample of the model part plus a sample of the adjustment part, $d_{t\ell}^\tau(\rho^\tau(\mathbf{s}_0), \sigma^\tau(\mathbf{s}_0))$. The model part is a function of the parameters and process realizations while the correction part is a function of just process realizations. Posterior samples for the parameters are available from the model fitting. Posterior samples for the GPs are available, using posterior samples of the parameters, through usual Bayesian kriging (Banerjee, Carlin and Gelfand (2015), Chapter 6). Therefore, we can interpolate marginal quantiles to any desired location in the study region. If we do this to a sufficiently spatially resolved grid, we can obtain the posterior mean at each point and "see" the posterior τ quantile surface for the given day within year.

Further, we might seek the average of the τ quantile over some subregion $\mathcal{B} \subseteq \mathcal{D}$ for day ℓ in year t. This becomes a block average, $\tilde{q}_{Y_{t\ell}(\mathcal{B})}(\tau) \equiv \int_\mathcal{B} \tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) \, d\mathbf{s}/|\mathcal{B}|$. As is customary, we approximate this integral by Monte Carlo integration of the form $\sum_{k=1}^K \tilde{q}_{Y_{t\ell}(\mathbf{s}_k)}(\tau)/K$ for $\{\mathbf{s}_k \in \mathcal{B}\}$, drawing the $\mathbf{s}_k$ from above grid. Lastly, we note that in the above, we are interpolating a parameter, not an observation. We are obtaining posterior distributions, not posterior predictive distributions.

4.2.1. *Marginal quantiles for one day at unobserved locations.* For each of the 3000 posterior samples of $\rho^\tau(\mathbf{s}_i)$ and $\sigma^\tau(\mathbf{s}_i)$ $(i = 1, \ldots, n)$ stored in Section 3.3, we use Bayesian kriging to obtain 3000 samples of $\rho^\tau(\mathbf{s}_k)$ and $\sigma^\tau(\mathbf{s}_k)$ $(k = 1, \ldots, K)$ in a grid of the study region $\mathcal{D}$ (see Section 2 for grid details). For each of them, we sample 1000 sets of $\{U_j\}$'s to calculate $P(\tilde{\epsilon}_\ell < d \mid \rho, \sigma)$, and obtain a sample of $d_{t\ell}^\tau(\rho^\tau(\mathbf{s}_k), \sigma^\tau(\mathbf{s}_k))$ through a one-dimensional rootfinder. In particular, $\rho^\ell W_0$ drops rapidly to zero as ℓ tends to infinity, so it becomes negligible.

As an illustration of the output provided by the marginal quantiles, Figure 8 shows maps of the posterior mean of $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau)$ on July 15, 2015 ($t = 60$ and $\ell = 76$) for $\tau = 0.05, 0.50, 0.95$. For all quantiles, the maximum temperature is reached in the Valley, center and southeast, and the minimum in the Pyrenees in the northeast. For $\tau = 0.05$, the temperature range goes from 6.0°C to 27.9°C, for $\tau = 0.50$ from 19.7°C to 37.1°C, and for $\tau = 0.95$ from 23.3°C to 41.1°C.

We use the marginal quantiles to analyze climate change for $\tau = 0.05, 0.50, 0.95$ in Figure 9. The first row shows spatially

$$(3) \qquad E\left(\frac{1}{10} \sum_{t \in D6} \tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) - \frac{1}{10} \sum_{t' \in D1} \tilde{q}_{Y_{t'\ell}(\mathbf{s})}(\tau) \,\Big|\, \text{data}\right),$$
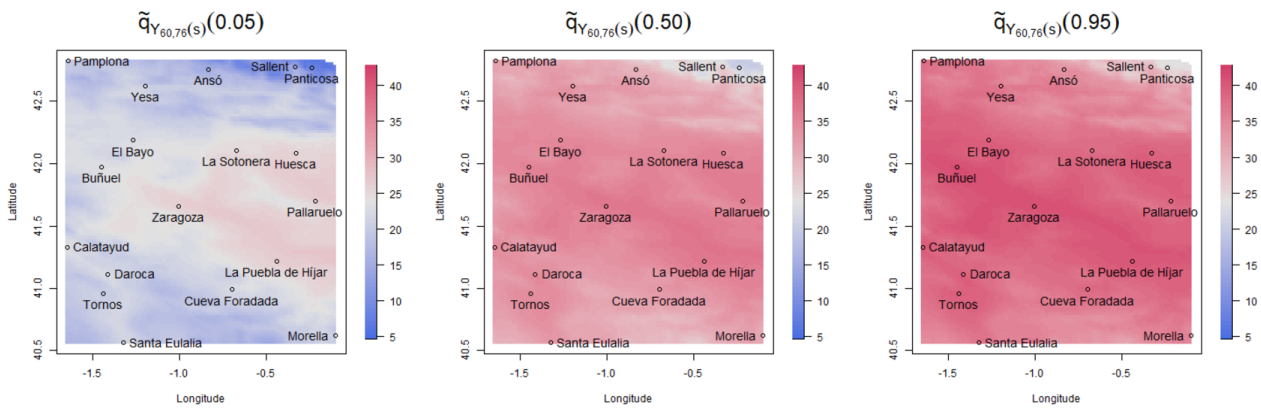
FIG. 8. *Maps of the posterior mean of the marginal quantiles $\tilde{q}_{Y_{60,76}(\mathbf{s})}(\tau)$ on July 15, 2015, for $\tau = 0.05, 0.50, 0.95$.*

where $D1$ is the first decade (1956–1965) and $D6$ the last (2006–2015). The result does not depend on ℓ since it can be summed up to the change in the terms $\alpha^{\tau} t + \gamma_t^{\tau}(\mathbf{s})$. The posterior mean in (3) supplies the mean change in temperature between the marginal quantile of a day averaged over the first decade and the marginal quantile of that same day averaged over the last decade, that is, we use averages of daily quantiles. The spatial pattern appears different across quantiles, with a smaller range of change for the median than for the extreme quantiles. Warming is general, exceeding 3°C in the southwest for $\tau = 0.95$. But cooling patterns also appear in the northwest for $\tau = 0.05$.

The second row shows spatially

$$(4) \qquad P\big(\tilde{q}_{Y_{t'\ell}(\mathbf{s})}(\tau) < \tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau) \mid t' \in D1, t \in D6, \text{data}\big).$$
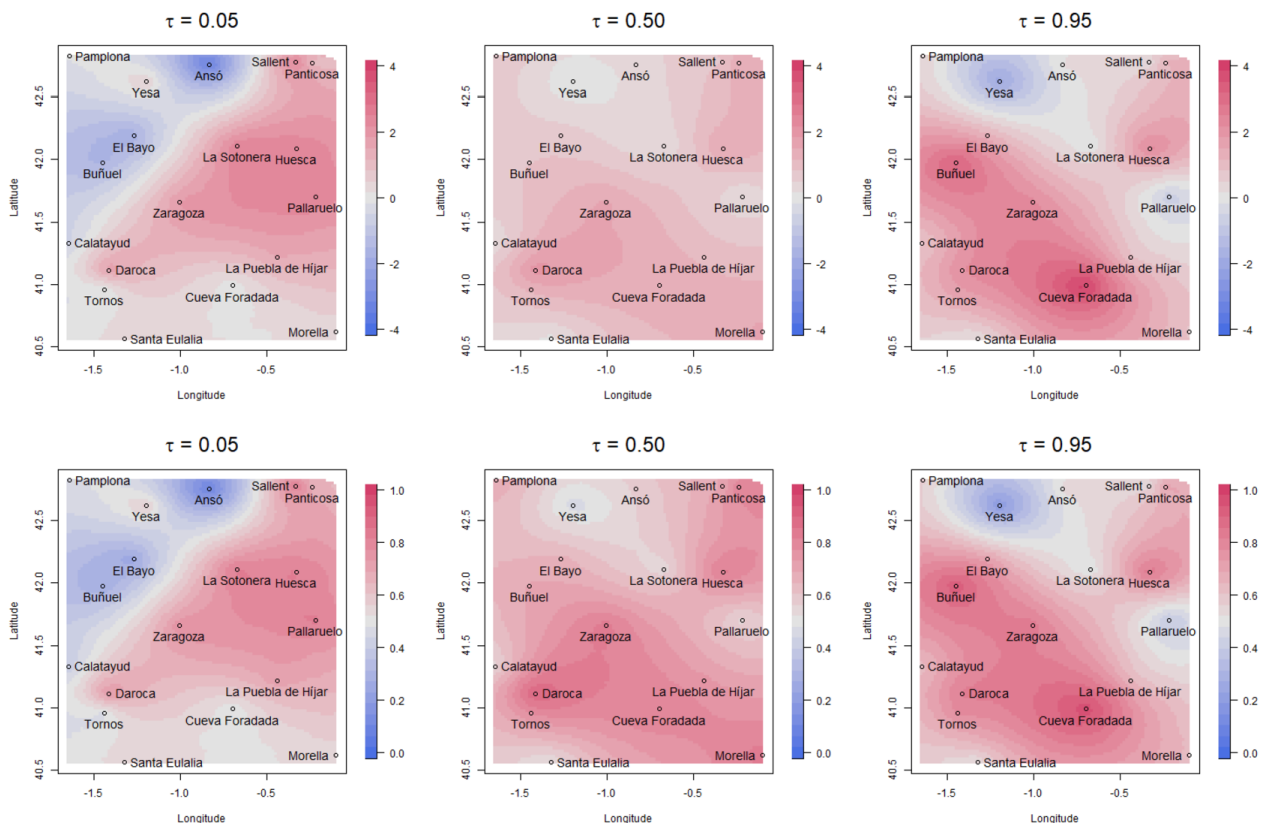

FIG. 9. *Top: Difference (in °C) between the marginal quantiles of the last and the first decades in (3). Bottom: Posterior probability in (4) that a marginal quantile in a day in a year of the first decade is colder than the same day in a year of the last decade. For $\tau = 0.05, 0.50, 0.95$.*
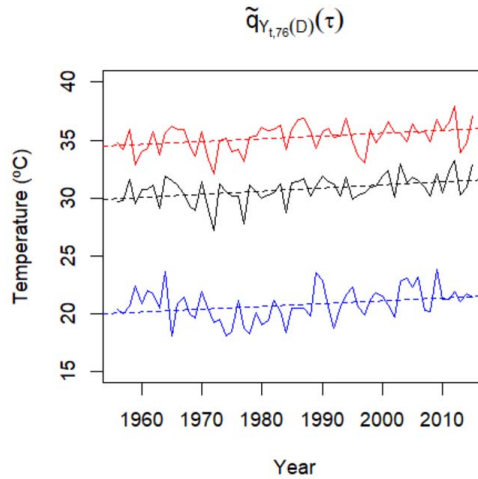
$$\tilde{q}_{Y_{t,76}(\mathcal{D})}(\tau)$$

FIG. 10. *Evolution of the block average $\tilde{q}_{Y_{t,76}(\mathcal{D})}(\tau)$ on July 15 against year for $\tau = 0.05, 0.50, 0.95$.*

This value represents the posterior probability that the marginal quantile of a day in any year of the first decade is colder than the marginal quantile of that same day for any year of the last decade. The pattern is the same as above, but in this case instead of seeing results in °C, we see them in terms of probabilities. To summarize the results, if we also condition on $\mathbf{s} \in \mathcal{D}$ in (4), these posterior probabilities are 0.67 for $\tau = 0.50$, 0.66 for $\tau = 0.95$, and 0.57 for $\tau = 0.05$.

Figure 10 shows the evolution over the years of the block average $\tilde{q}_{Y_{t\ell}(\mathcal{D})}(\tau)$ for July 15 ($\ell = 76$). A warming trend is observed in all three quantiles. Although the baseline may change for other choices of ℓ, the pattern through the years will be common.

4.3. *Marginal quantiles for averages.* Next, we take up averaging over time to obtain say, a weekly τ quantile at site $\mathbf{s}$ in year t. To simplify notation, we suppress the year, the site, and the superscript τ. Suppose we want to average back r days, starting at day $\ell \geq r$. So, we are seeking the marginal τ-quantile of $\bar{Y}_\ell^{(r)} = \frac{1}{r} \sum_{j=\ell-r+1}^{\ell} Y_j$. To be clear, we want the quantile of this average, not the average of the daily quantiles. Going one step further, if we want to average over space (as above) and time, we should first average over time and then average over space. That is, the quantile of the temporal average is not the average of the temporal quantiles but $\tilde{q}_{Y_{t\ell}(\mathcal{B})}(\tau)$, by definition, is an average of quantiles.

From above, we have $Y_\ell = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1}) + \epsilon_\ell$ where the $\epsilon_\ell \sim$ i.i.d. AL$(0, \sigma, \tau)$. So, again, $Q_{Y_\ell}(\tau \mid Y_{\ell-1}) = q_\ell + \rho(Y_{\ell-1} - q_{\ell-1})$ is the τ quantile for the QAR for day ℓ. Again, for convenience, write this model as $W_\ell = \rho W_{\ell-1} + \epsilon_\ell$ with $W_\ell = Y_\ell - q_\ell$.

Then, if $\bar{W}_\ell^{(r)} = \frac{1}{r} \sum_{j=\ell-r+1}^{\ell} W_j$, $\bar{W}_\ell^{(r)} = \bar{Y}_\ell^{(r)} - \bar{q}_\ell^{(r)}$ where $\bar{q}_\ell^{(r)}$ averages the q's accordingly. So, the marginal τ quantile for $\bar{Y}_\ell^{(r)}$ will be $\bar{q}_\ell^{(r)}$ plus an adjustment.

Since $W_j = \rho^j W_0 + \sum_{k=0}^{j-1} \rho^k \epsilon_{j-k}$, we need the τ quantile of

$$\bar{\bar{\epsilon}}_\ell^{(r)} \equiv \frac{1}{r} \sum_{j=\ell-r+1}^{\ell} \sum_{k=0}^{j-1} \rho^k \epsilon_{j-k}.$$

Note that, while the double sum is over $r(\ell - r) + r(r + 1)/2$ terms, it only involves $\epsilon_1, \ldots, \epsilon_\ell$. We can rewrite the sum in terms of these ℓ distinct ϵ's but to no advantage. Rather, we need to generate ℓ associated U's, that is, ℓ i.i.d. Exp(1) random variables. For each ρ, σ, and given these U's,

$$\bar{\bar{\epsilon}}_\ell^{(r)} \sim N\left(\frac{1 - 2\tau}{\sigma\tau(1 - \tau)} \frac{1}{r} \sum_{j=\ell-r+1}^{\ell} \sum_{k=0}^{j-1} \rho^k U_{j-k}, \frac{2}{\sigma^2\tau(1 - \tau)} \frac{1}{r^2} \sum_{j=\ell-r+1}^{\ell} \sum_{k=0}^{j-1} \rho^{2k} U_{j-k} \right).$$
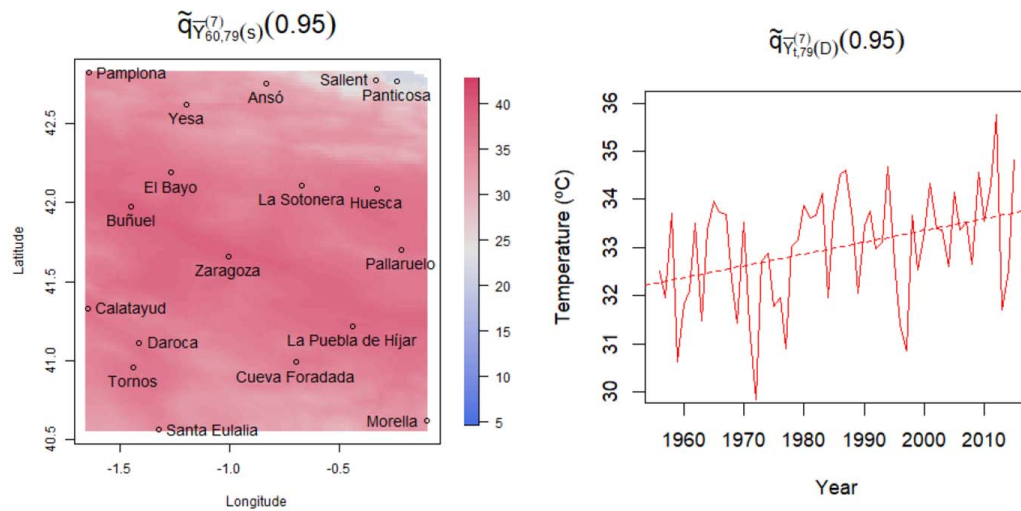
FIG. 11.　*Left*: *Map of the posterior mean of the weekly marginal quantiles* $\tilde{q}_{\bar{Y}^{(7)}_{60,79}(\mathbf{s})}(0.95)$ *centered on July* 15, 2015. *Right*: *Evolution of the block average of the weekly marginal quantiles* $\tilde{q}_{\bar{Y}^{(7)}_{t,79}(\mathcal{D})}(0.95)$ *centered on July* 15 *against year.*

Again, using a simple search, we can find the τ quantile of $\bar{\tilde{\epsilon}}^{(r)}_\ell$ and, therefore, we can find the adjustment to obtain the τ marginal quantile for $\bar{Y}^{(r)}_\ell$. In our modeling setting, we denote this quantile by $\tilde{q}_{\bar{Y}^{(r)}_{t\ell}(\mathbf{s})}(\tau)$. With posterior samples from the model fitting, we can create a posterior distribution for this marginal quantile. We can do this for any location, averaging back from any day within a year, and for any year. Thus, we can create an associated spatial surface. Note that we can reuse the U's except the number that we require depends upon ℓ as above. Finally, if we want this τ marginal quantile averaged over a region say, $\mathcal{B}$, we can implement an analogous Monte Carlo integration as above, replacing $\tilde{q}_{Y_{t\ell}(\mathbf{s})}(\tau)$ with $\tilde{q}_{\bar{Y}^{(r)}_{t\ell}(\mathbf{s})}(\tau)$.

4.3.1. *Weekly marginal quantiles at a given site.*　Note that in climate analysis some extreme events are defined by the integration of the temperature series over several days. For example, Cattiaux and Ribes (2018) study the probability of extremely hot temperatures during a moving-window of $r = 2, 3, \ldots, 15$ days in Paris, and Lee (2021) analyzes trends in extreme "weather whiplash" events defined by daily temperature in a 7-day moving window. Here, as an example of application of the proposed methodology, we follow the setup in Section 4.2.1 to obtain weekly marginal quantiles $\tilde{q}_{\bar{Y}^{(r)}_{t\ell}(\mathbf{s})}(\tau)$ centered at July 15, 2015 ($t = 60$, $\ell = 79$ and $r = 7$), for $\tau = 0.95$.

In particular, the plot on the left of Figure 11 shows the map of the posterior mean of $\tilde{q}_{\bar{Y}^{(r)}_{t\ell}(\mathbf{s})}(\tau)$ centered on July 15, 2015 ($t = 60$, $\ell = 79$ and $r = 7$) for $\tau = 0.95$. The spatial pattern is the same as for the daily marginal quantiles, but the temperature range is 21.2–38.9°C. The plot on the right of Figure 11 shows the weekly block average centered on July 15 across years for $\tau = 0.95$.

Lastly, other values of r can be considered but, if r is taken too large, many values will be averaged and, therefore, the quantiles from this averaging will get closer and closer. This can cause quantile crossing even for quantiles that are far from each other.

5. Summary and future work.　This paper develops a modeling approach to predict a specific quantile in a spatiotemporal framework. We have specified a spatial conditional autoregressive model on a daily scale using the AL distribution for the errors. The considered specification enables spatial autoregression at a daily scale that captures serial correlation and facilitates assessment of persistence. The flexibility of the model is increased by considering

two scales of time—days within summer season and years—as well as seasonal behavior, time trend, and four GPs that represent the spatial dependence of the intercept, the trend, the serial dependence, and the scale of the AL errors. Bayesian model fitting enables full posterior inference for a given quantile. Although the model gives conditional quantiles, we offer an attractive approach to obtain marginal quantiles at daily scale. These marginal quantiles enable interpolation. The approach can also provide marginal quantiles associated with averages of the response variable, both spatially and dynamically. Posterior inference to evaluate changes between marginal quantiles of spatial and time averages can also be implemented.

The suggested QAR modeling approach is shown to be flexible enough to represent the evolution of different quantiles of the distribution of daily maximum temperatures and to capture the effects of climate change during the period 1956–2015 in Aragón, a small region but with a wide variety of climate conditions. The strong serial correlation of daily temperature is adequately captured by the autoregressive structure. The elevation, the only spatial covariate in the model, together with the four considered GPs are able to capture the great variability in climate conditions over the region; in particular, they capture observed features of temperature by allowing mean levels, trends over time, and serial correlation of temperature to vary spatially. The QAR models fitted for $\tau = 0.05, 0.50, 0.95$ show different spatial and temporal patterns, revealing important differences in the behavior of the tails versus the central part of the distribution of daily temperature in summer. More precisely, comparing the increases over time observed in the median, the 0.95 quantile shows higher increases in some areas of the Valley, while for the 0.05 quantile, no increase is observed in the northwest. These differences confirm the importance of modeling the entire distribution of daily temperature, rather than just the mean (as done in many climate studies). A useful climate application of the proposed methodology to estimate quantiles is the computation of thresholds to define extreme indexes or extreme events, taking into account changes over time of temperature.

Future work will consider different spatial regions providing more spatial sites than our sparse Aragón data set. This will enable comparison of temperature trends at larger spatial scales. Additionally, our modeling approach could be useful in other environmental analyses such as pollution exposure or biological experimental data, where the objective is to identify distributional changes over time and compare these changes across different spatial locations. Further, though the proposed modeling analyzes daily series across years, the approach could be applied to other time scales.

SUPPLEMENTARY MATERIAL

Supplement to "Spatial quantile autoregression for season within year daily maximum temperature data" (DOI: 10.1214/22-AOAS1719SUPP; .pdf). This Supplementary Material provides additional details for the descriptive data analysis, a residual analysis, the details of the Gibbs sampler algorithm used to fit the model, and additional figures for the results of model fitting.

REFERENCES

AEMET (2011). Atlas Climático Ibérico—Iberian Climate Atlas. Ministerio de Medio Ambiente, y Medio Rural y Marino; Agencia Estatal de Meteorología; and Instituto de Meteorologia de Portugal. https://doi.org/10.31978/784-11-002-5

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184 https://doi.org/10.1201/b17115

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 https://doi.org/10.2307/1390675

CASTILLO-MATEO, J., LAFUENTE, M., ASÍN, J., CEBRIÁN, A. C., GELFAND, A. E. and ABAURREA, J. (2022). Spatial modeling of day-within-year temperature time series: An examination of daily maximum temperatures in Aragón, Spain. *J. Agric. Biol. Environ. Stat.* **27** 487–505. MR4459077 https://doi.org/10.1007/s13253-022-00493-3

CASTILLO-MATEO, J., ASÍN, J., CEBRIÁN, A. C., GELFAND, A. E. and ABAURREA, J. (2023). Supplement to "Spatial quantile autoregression for season within year daily maximum temperature data." https://doi.org/10.1214/22-AOAS1719SUPP

CATTIAUX, J. and RIBES, A. (2018). Defining single extreme weather events in a climate perspective. *Bull. Am. Meteorol. Soc.* **99** 1557–1568. https://doi.org/10.1175/BAMS-D-17-0281.1

CHEN, X. and TOKDAR, S. T. (2021). Joint quantile regression for spatial data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 826–852. MR4320003 https://doi.org/10.1111/rssb.12467

CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. MR1984779 https://doi.org/10.1016/S0304-4076(03)00100-3

DAS, P. and GHOSAL, S. (2017a). Bayesian quantile regression using random B-spline series prior. *Comput. Statist. Data Anal.* **109** 121–143. MR3603645 https://doi.org/10.1016/j.csda.2016.11.014

DAS, P. and GHOSAL, S. (2017b). Analyzing ozone concentration by Bayesian spatio-temoral quantile regression. *Environmetrics* **28** e2443, 15 pp. MR3660099 https://doi.org/10.1002/env.2443

GAO, M. and FRANZKE, C. L. E. (2017). Quantile regression-based spatiotemporal analysis of extreme temperature change in China. *J. Climate* **30** 9897–9914. https://doi.org/10.1175/JCLI-D-17-0356.1

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82** 479–488. MR1366275 https://doi.org/10.1093/biomet/82.3.479

HALLIN, M., LU, Z. and YU, K. (2009). Local linear spatial quantile regression. *Bernoulli* **15** 659–686. MR2555194 https://doi.org/10.3150/08-BEJ168

HAUGEN, M. A., STEIN, M. L., MOYER, E. J. and SRIVER, R. L. (2018). Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression. *J. Climate* **31** 8573–8588. https://doi.org/10.1175/JCLI-D-17-0782.1

KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657 https://doi.org/10.1017/CBO9780511754098

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644 https://doi.org/10.2307/1913643

KOENKER, R. and MACHADO, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *J. Amer. Statist. Assoc.* **94** 1296–1310. MR1731491 https://doi.org/10.2307/2669943

KOENKER, R. and XIAO, Z. (2006). Quantile autoregression. *J. Amer. Statist. Assoc.* **101** 980–990. MR2324109 https://doi.org/10.1198/016214506000000672

KOTZ, S., KOZUBOWSKI, T. J. and PODGÓRSKI, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance.* Birkhäuser, Inc., Boston, MA. MR1935481 https://doi.org/10.1007/978-1-4612-0173-1

KOZUMI, H. and KOBAYASHI, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **81** 1565–1578. MR2851270 https://doi.org/10.1080/00949655.2010.496117

LEE, C. C. (2021). Weather whiplash: Trends in rapid temperature changes in a warming climate. *Int. J. Climatol.* **42** 4214–4222. https://doi.org/10.1002/joc.7458

LI, G., LI, Y. and TSAI, C.-L. (2015). Quantile correlations and quantile autoregressive modeling. *J. Amer. Statist. Assoc.* **110** 246–261. MR3338500 https://doi.org/10.1080/01621459.2014.892007

LUM, K. and GELFAND, A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Anal.* **7** 235–258. MR2934947 https://doi.org/10.1214/12-BA708

MCKINNON, K. A. and POPPICK, A. (2020). Estimating changes in the observed relationship between humidity and temperature using noncrossing quantile smoothing splines. *J. Agric. Biol. Environ. Stat.* **25** 292–314. MR4132962 https://doi.org/10.1007/s13253-020-00393-4

NAVARRO-SERRANO, F., LÓPEZ-MORENO, J. I., AZORIN-MOLINA, C., ALONSO-GONZÁLEZ, E., TOMÁS-BURGUERA, M., SANMIGUEL-VALLELADO, A., REVUELTO, J. and VICENTE-SERRANO, S. M. (2018). Estimation of near-surface air temperature lapse rates over continental Spain and its mountain areas. *Int. J. Climatol.* **38** 3233–3249. https://doi.org/10.1002/joc.5497

NEELON, B., LI, F., BURGETTE, L. F. and BENJAMIN NEELON, S. E. (2015). A spatiotemporal quantile regression model for emergency department expenditures. *Stat. Med.* **34** 2559–2575. MR3368401 https://doi.org/10.1002/sim.6480

PEÑA-ANGULO, D., GONZALEZ-HIDALGO, J. C., SANDONÍS, L., BEGUERÍA, S., TOMAS-BURGUERA, M., LÓPEZ-BUSTINS, J. A., LEMUS-CANOVAS, M. and MARTIN-VIDE, J. (2021). Seasonal temperature trends on the Spanish mainland: A secular study (1916–2015). *Int. J. Climatol.* **41** 3071–3084. https://doi.org/10.1002/joc.7006

PETERS, G. W. (2018). General quantile time series regressions for applications in population demographics. *Risks* **6** 97. https://doi.org/10.3390/risks6030097

REICH, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 535–553. MR2960737 https://doi.org/10.1111/j.1467-9876.2011.01025.x

REICH, B. J., FUENTES, M. and DUNSON, D. B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. MR2816698 https://doi.org/10.1198/jasa.2010.ap09237

SRIRAM, K., RAMAMOORTHI, R. V. and GHOSH, P. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Anal.* **8** 479–504. MR3066950 https://doi.org/10.1214/13-BA817

TAN, X., GAN, T. Y. and CHEN SHU LIU, B. (2019). Modeling distributional changes in winter precipitation of Canada using Bayesian spatiotemporal quantile regression subjected to different teleconnections. *Clim. Dyn.* **52** 2105–2124. https://doi.org/10.1007/s00382-018-4241-0

TOKDAR, S. T. and KADANE, J. B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Anal.* **7** 51–72. MR2896712 https://doi.org/10.1214/12-BA702

YANG, Y. and HE, X. (2015). Quantile regression for spatially correlated data: An empirical likelihood approach. *Statist. Sinica* **25** 261–274. MR3328814

YANG, C., LI, L. and XU, J. (2018). Changing temperature extremes based on CMIP5 output via semi-parametric quantile regression approach. *Int. J. Climatol.* **38** 3736–3748. https://doi.org/10.1002/joc.5524

YANG, Y. and TOKDAR, S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *J. Amer. Statist. Assoc.* **112** 1107–1120. MR3735363 https://doi.org/10.1080/01621459.2016.1192545

YANG, Y., WANG, H. J. and HE, X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *Int. Stat. Rev.* **84** 327–344. MR3580414 https://doi.org/10.1111/insr.12114

YU, K. and MOYEED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. MR1861390 https://doi.org/10.1016/S0167-7152(01)00124-9

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

Supplement to "Spatial quantile autoregression for season within year daily maximum temperature data"

Jorge Castillo-Mateo, Jesús Asín, Ana C. Cebrián, Alan E. Gelfand and Jesús Abaurrea

S1. Descriptive analysis

S1.1. Quantile autocorrelation

For random variables X and Y, let $Q_Y(\tau)$ be the τ unconditional quantile of Y. For $\tau \in (0,1)$, Li, Li and Tsai (2015) define the quantile covariance as

$$\mathrm{qcov}_\tau(Y, X) \equiv \mathrm{cov}\left[\mathbf{1}(Y - Q_Y(\tau) > 0), X\right] = E\left[\delta_\tau^*(Y - Q_Y(\tau))(X - E(X))\right],$$

where the function $\delta_\tau^*(u) = \tau - \mathbf{1}(u < 0)$. Subsequently, they define the quantile correlation as follows,

$$\mathrm{qcor}_\tau(Y, X) \equiv \frac{\mathrm{qcov}_\tau(Y, X)}{\sqrt{\mathrm{var}\left[\delta_\tau^*(Y - Q_Y(\tau))\right]\mathrm{var}(X)}} = \frac{E\left[\delta_\tau^*(Y - Q_Y(\tau))(X - E(X))\right]}{\sqrt{(\tau - \tau^2)\mathrm{var}(X)}}.$$

We compute the empirical first-order quantile autocorrelation for our data by substituting each element for its sample or empirical estimates in $\mathrm{qcor}_\tau(Y, X)$ and considering Y and X as follows. Let $\{y_t\}$ be the observed daily maximum temperatures for any site from January 1, 1956 to December 31, 2015. Then, we take $\{y_t\}$ as a sample of Y and $\{y_{t-1}\}$ as a sample of X for t in the 60×92 days of the JJA months. Figure S1 (left) shows this autocorrelation as explained in Figure 2 of the Main Manuscript. The correlation is strong in all locations and quantiles with positive values in the range 0.27–0.63. When exploring a second-order quantile autocorrelation, we find no evidence of its existence in the data. Figure S1 (right) shows the quantile (partial) autocorrelation between $\{y_t\}$ and $\{y_{t-2}\}$ conditioned on $\{y_{t-1}\}$. The 0.05, 0.50 and 0.95 quantiles of these correlations are -0.07, -0.03 and 0.02, suggesting that quantile correlation between $\{y_t\}$ and $\{y_{t-2}\}$ given $\{y_{t-1}\}$ is negligible.

Figure S2 shows the effect of the first and second-order quantile autocorrelations for three illustrative sites: Zaragoza, Daroca and Yesa. It shows the empirical 0.05, 0.50 and 0.95 quantiles of y_t conditional on y_{t-1} and y_{t-2}, for days in JJA. To obtain the plots, we separate observations in three sets: y_t with $y_{t-2} \leq q_Y^{emp}(1/3)$ (blue), y_t with $q_Y^{emp}(1/3) < y_{t-2} \leq q_Y^{emp}(2/3)$ (black), and y_t with $q_Y^{emp}(2/3) < y_{t-2}$ (red); where $q_Y^{emp}(1/3)$ and $q_Y^{emp}(2/3)$ are the corresponding empirical quantiles of $\{y_t\}$. In each set, we compute the empirical quantiles of the y_t's in subsets where the y_{t-1}'s take values in an interval of length $2°C$ represented in the horizontal axis (only quantiles calculated with at least 30 data are shown). The increasing linear trends in the empirical quantiles give evidence of an strong effect of y_{t-1} on y_t, but the overlap of the three colors gives little or no evidence of an effect of y_{t-2} on y_t given y_{t-1}.
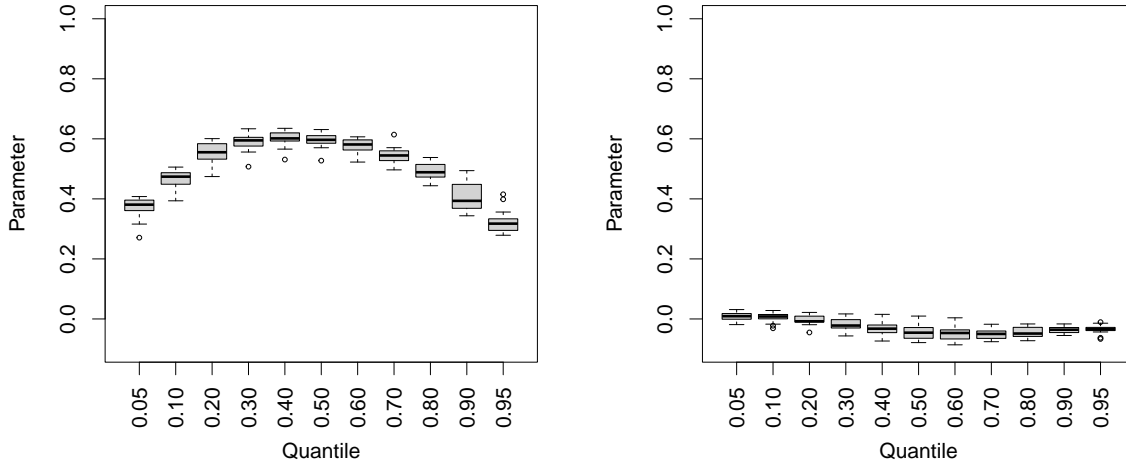
FIG S1. *Boxplots of the 18 sites across quantiles to describe the empirical quantile features: first-order quantile autoregression (right) and second-order quantile (partial) autoregression (left).*
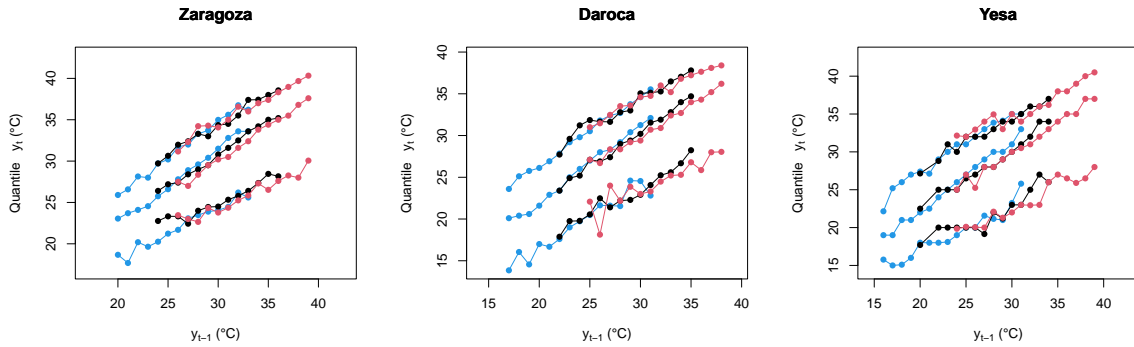

FIG S2. *Empirical 0.05, 0.50 and 0.95 quantiles vs. data from the previous day separated by three sets of data from two days ago. The three sets are defined according to $y_{t-2} \leq q_Y^{emp}(1/3)$ (blue), $q_Y^{emp}(1/3) < y_{t-2} \leq q_Y^{emp}(2/3)$ (black), and $q_Y^{emp}(2/3) < y_{t-2}$ (red). Here, we show Zaragoza, Daroca and Yesa as an illustration of the 18 sites.*

S1.2. Elevation, latitude and temporal patterns

We follow Section 2.1 of the Main Manuscript on the empirical description of quantiles. Figure S3 shows the 0.05, 0.50 and 0.95 quantiles vs. elevation (left) and latitude (right) for each site. These quantiles are empirical for daily data in JJA from 1956 to 2015 at each site, i.e., each quantile (point in the plot) is calculated with 60×92 observations. For each quantile a line of least squares is also drawn. There is a clear negative relation approximately linear between temperature and elevation, and slightly more pronounced for the 0.05 quantile. For latitude, although the lines suggest a slight negative trend between temperature and latitude, this is mainly caused by a set of two higher elevation sites located in the mountains of the north of the region. No pattern is observed in the rest of the sites.

Figure S4 shows the temporal evolution of the quantiles for the three illustrative sites. The first row shows the 0.05, 0.50 and 0.95 quantiles vs. day in JJA. The quantile of a given day is the empirical quantile of the daily data in a 21-day window centered on that day in the first (1956–1985) (dashed) and the last (1986–2015) (solid) 30-year periods. Finally, to eliminate variability, what is shown in the figure is a smoothing using a LOWESS of these quantiles. As a general conclusion, there is a strong potentially common seasonal component between quantiles that has not changed between the two 30-year periods. The change observed between both 30-year periods appears reasonably homogeneous throughout the summer, although it seems to be heterogeneous across quantiles. The second row shows the kernel density estimation for the data in a 21-day window centered on July 15 in the first (dashed blue) and the last (solid red) 30-year periods. In Zaragoza and Daroca the density of the last period is placed to the right of the first, indicating warmer temperatures for all quantiles. In Yesa the behavior is different, while the 0.05 quantile of the first period is a little to the left than that of the last, the median and in particular the 0.95 quantile of the first period are located to the right of those of the first period, indicating a cool-down period.
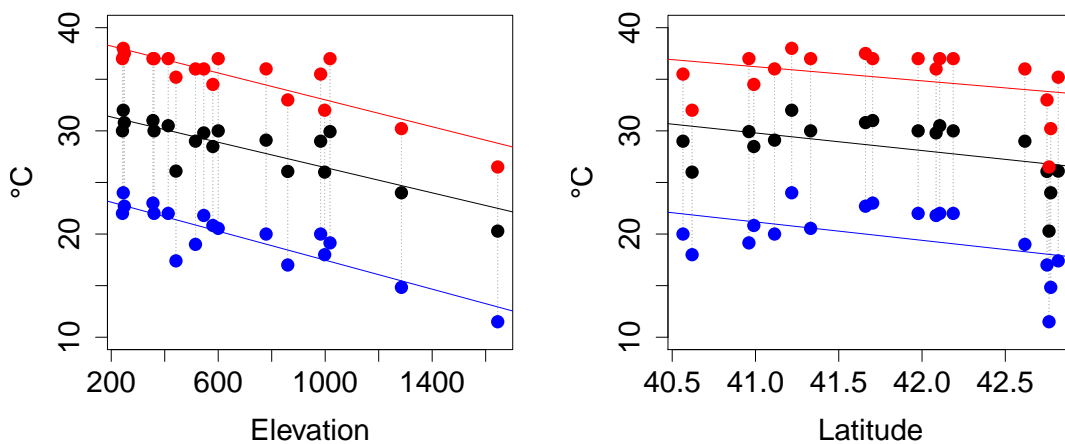

FIG S3. *Empirical 0.05 (blue), 0.50 (black) and 0.95 (red) quantiles for data in JJA months from 1956 to 2015 by site vs. elevation (left) and latitude (right).*
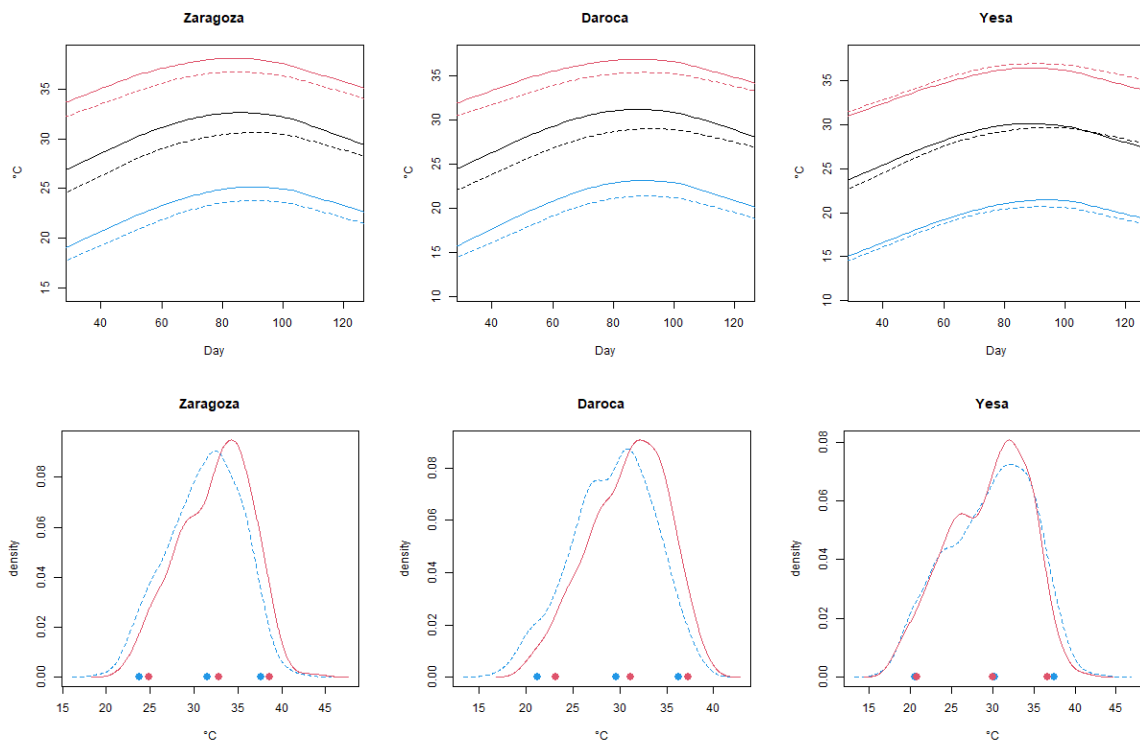
FIG S4. *Top: Smoothed empirical* 0.05 *(blue),* 0.50 *(black) and* 0.95 *(red) quantiles across days in JJA months, 1956–1985 (dashed) and 1986–2015 (solid). Bottom: Kernel density estimation of the data in a 21-day window centered on July 15 in 1956–1985 (dashed blue) and 1986–2015 (solid red). The* 0.05, 0.50 *and* 0.95 *quantiles are denoted by points on the horizontal axis. Here, we show Zaragoza, Daroca and Yesa as an illustration of the* 18 *sites.*

S2. Analysis of residuals

Here, we show model assessment of the seasonal cycle. We check the assumption of a spatially common seasonal cycle through the analysis of residuals by months and locations to detect if any of them show any seasonal pattern not captured by the common seasonal term. Residuals are defined as $Y_{t\ell}(\mathbf{s}) - Q_{Y_{t\ell}(\mathbf{s})}(\tau \mid Y_{t,\ell-1}(\mathbf{s}))$. Figure S5 shows the boxplots of residuals by months in JJA, in the three illustrative locations, and for three different quantiles, $\tau = 0.05, 0.50, 0.95$. The seasonal behavior of the series seems to be well captured by the common seasonal term in all the locations, since no bias is observed.

We check the assumption of a seasonal cycle that does not vary in years through the analysis of residuals by months, decades and locations to detect if any of them show any seasonal pattern not captured by the non-varying seasonal term. Figure S6 show the boxplots of the residuals in Zaragoza in the 6 available decades in June (left), July (center) and August (right) for three different quantiles, $\tau = 0.05, 0.50, 0.95$. The evolution of the boxplots across decades is very similar in June, July and August, and no bias or evidence of a remaining seasonal behavior is observed. Similar behaviors are observed in all locations.

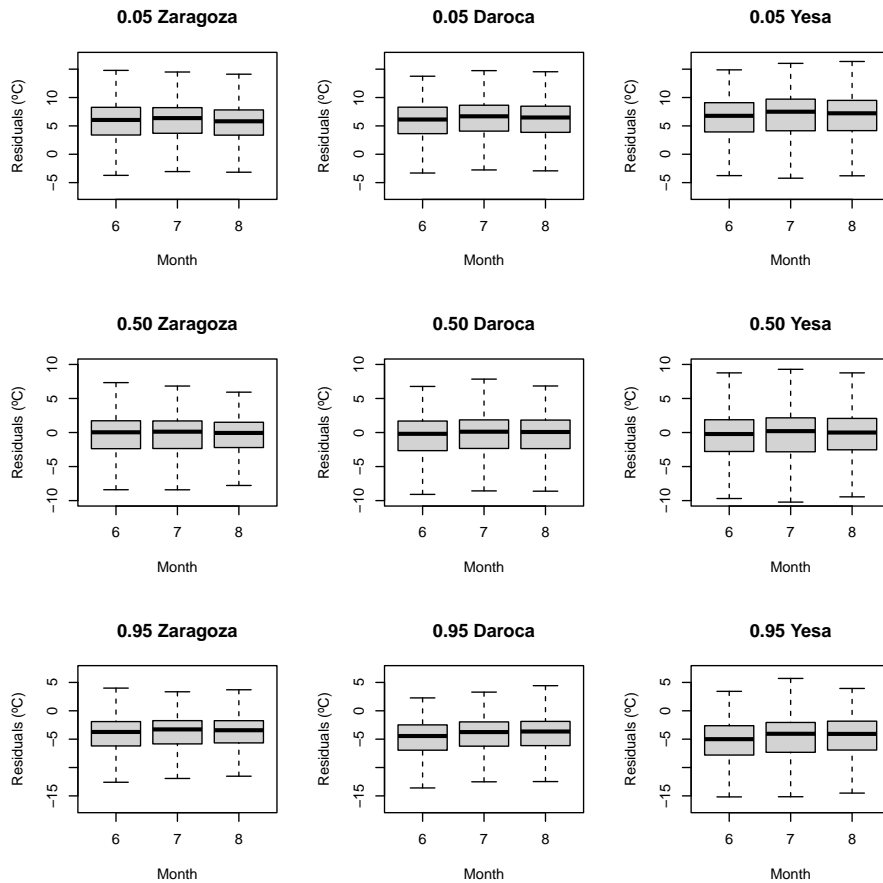FIG S5. *Boxplots of residuals, of the $\tau = 0.05, 0.50, 0.95$ quantile models, by months in JJA. Here, we show Zaragoza, Daroca and Yesa as an illustration of the 18 sites.*

0.05 Zaragoza

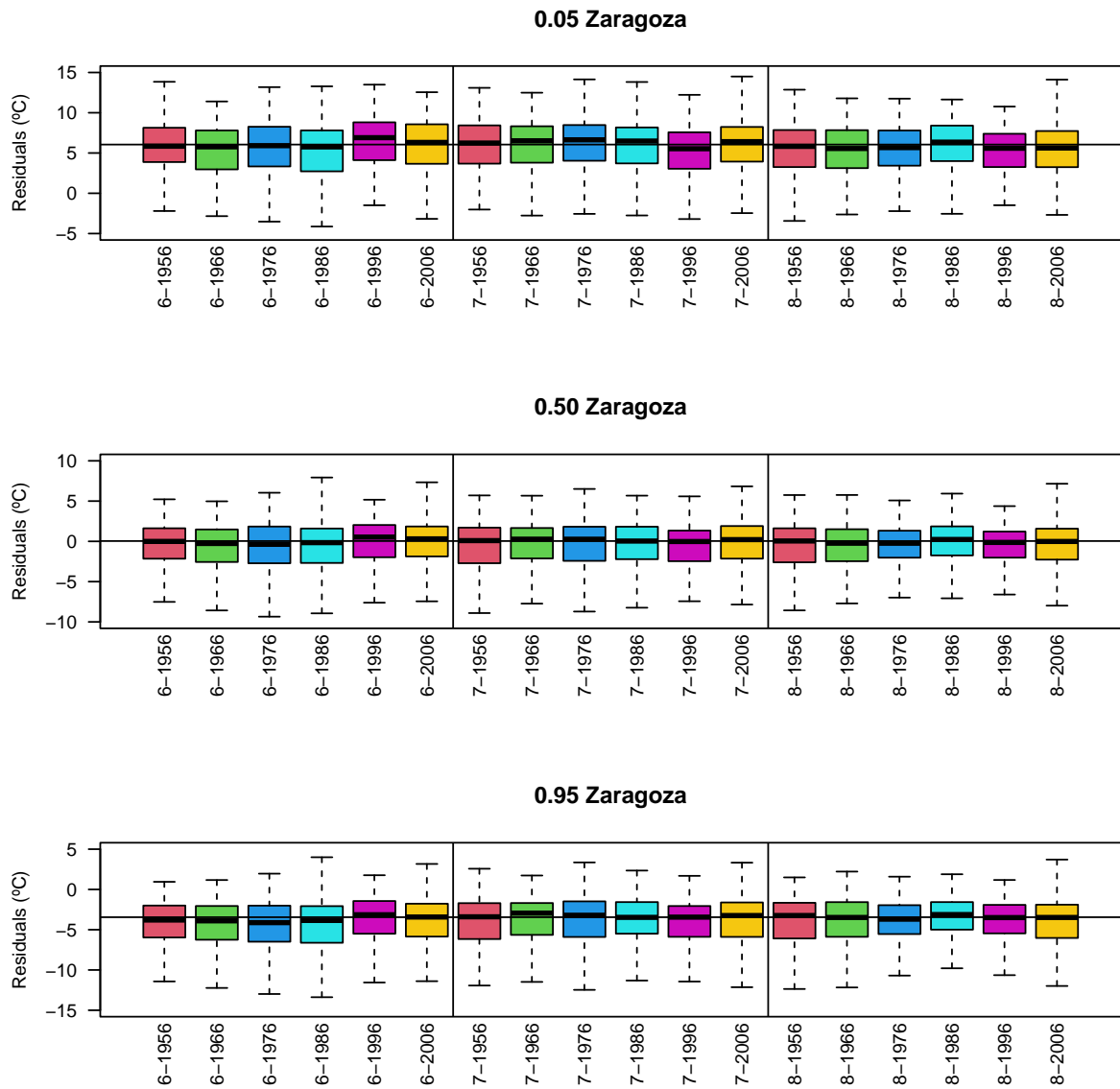**0.50 Zaragoza**

0.95 Zaragoza

FIG S6. *Boxplots of residuals, of the $\tau = 0.05, 0.50, 0.95$ quantile models, by months in JJA and 10-year periods. Here, we show Zaragoza as an illustration of the 18 sites.*

S3. Metropolis-within-Gibbs algorithm

This section describes the algorithm used to fit the models. For simplicity we suppress the superscript τ from the parameters. Also for simplicity, we denote by $Q^*_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i))$ the part of $Q_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i))$ that does not contain the variable associated with the full conditional in each case. We denote $c_1 \equiv c_1^\tau = (1-2\tau)/(\tau(1-\tau))$ and $c_2 \equiv c_2^\tau = 2/(\tau(1-\tau))$. We denote by r_{jk} the elements of the inverse matrix of $R(\phi)$, the correlation matrix derived from the exponential covariance function. We denote $\sin_\ell = \sin(2\pi\ell/365)$ and $\cos_\ell = \cos(2\pi\ell/365)$. Also, we use the usual parameterization with AL errors (Kozumi and Kobayashi, 2011), $\xi_{t\ell}(\mathbf{s}) = U_{t\ell}(\mathbf{s})/\sigma(\mathbf{s})$. It does not help us in our setup, but it is useful for a common scale parameter in space and a gamma prior distribution in it.

The joint distribution for data, processes and parameters is

$$
\prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{\ell=2}^{L} \left[Y_{t\ell}(\mathbf{s}_i) \mid Y_{t,\ell-1}(\mathbf{s}_i), \beta_1, \beta_2, \beta_3, \gamma_t(\mathbf{s}_i), \rho(\mathbf{s}_i), \xi_{t\ell}(\mathbf{s}_i), \sigma(\mathbf{s}_i)\right]
$$
$$
\prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{\ell=2}^{L} \left[\xi_{t\ell}(\mathbf{s}_i) \mid \sigma(\mathbf{s}_i)\right] \prod_{i=1}^{n}\prod_{t=1}^{T} \left[\gamma_t(\mathbf{s}_i) \mid \tilde{\beta}_0(\mathbf{s}_i), \tilde{\alpha}(\mathbf{s}_i), \psi_t, \sigma_\eta^2\right] \prod_{t=2}^{T}[\psi_t \mid \sigma_\psi^2] \quad\quad (S1)
$$
$$
\left[\{\tilde{\beta}_0(\mathbf{s}_i)\} \mid \beta_0, \sigma_{\beta_0}^2, \phi\right]\left[\{\tilde{\alpha}(\mathbf{s}_i)\} \mid \alpha, \sigma_\alpha^2, \phi\right]\left[\{Z_\rho(\mathbf{s}_i)\} \mid Z_\rho, \sigma_\rho^2, \phi\right]\left[\{Z_\sigma(\mathbf{s}_i)\} \mid Z_\sigma, \sigma_\sigma^2, \phi\right]
$$
$$
[\beta_0]\,[\alpha]\,[\beta_1]\,[\beta_2]\,[\beta_3]\,[Z_\rho]\,[Z_\sigma]\,[\sigma_\psi^2]\,[\sigma_\eta^2]\,[\sigma_{\beta_0}^2]\,[\sigma_\alpha^2]\,[\sigma_\rho^2]\,[\sigma_\sigma^2]
$$

provided one starts any year t with the observed $Y_{t1}(\mathbf{s})$, and for a common fixed ϕ.

The Gibbs sampler algorithm for Equation S1 is initialized giving initial values to all the parameters. Then, updating from iteration b to $b+1$ consists of drawing a sample from the following full conditional distributions:

- The full conditional distributions of $\beta_0, \alpha, \beta_1, \beta_2, \beta_3, Z_\rho, Z_\sigma$ are all normal, in particular

$$
[\beta_0 \mid \cdots] \propto N\left(\beta_0 \mid \frac{\sum_{j,k} r_{jk}\tilde{\beta}_0(\mathbf{s}_k)}{\sum_{j,k} r_{jk}}, \frac{\sigma_{\beta_0}^2}{\sum_{j,k} r_{jk}}\right) \times N(\beta_0 \mid a_{\beta_0}, b_{\beta_0}^2)
$$

$$
[\alpha \mid \cdots] \propto N\left(\alpha \mid \frac{\sum_{j,k} r_{jk}\tilde{\alpha}(\mathbf{s}_k)}{\sum_{j,k} r_{jk}}, \frac{\sigma_\alpha^2}{\sum_{j,k} r_{jk}}\right) \times N(\alpha \mid a_\alpha, b_\alpha^2)
$$

$$
[\beta_1 \mid \cdots] \propto \prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{\ell=2}^{L} N\left(\beta_1 \mid \frac{Y_{t\ell}(\mathbf{s}_i) - Q^*_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i)) - c_1\xi_{t\ell}(\mathbf{s}_i)}{(\sin_\ell - \rho(\mathbf{s}_i)\sin_{\ell-1})},\right.
$$
$$
\left. \frac{c_2\xi_{t\ell}(\mathbf{s}_i)}{(\sin_\ell - \rho(\mathbf{s}_i)\sin_{\ell-1})^2\sigma(\mathbf{s}_i)}\right) \times N(\beta_1 \mid a_{\beta_1}, b_{\beta_1}^2)
$$

$$
[\beta_2 \mid \cdots] \propto \prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{\ell=2}^{L} N\left(\beta_2 \mid \frac{Y_{t\ell}(\mathbf{s}_i) - Q^*_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i)) - c_1\xi_{t\ell}(\mathbf{s}_i)}{(\cos_\ell - \rho(\mathbf{s}_i)\cos_{\ell-1})},\right.
$$
$$
\left. \frac{c_2\xi_{t\ell}(\mathbf{s}_i)}{(\cos_\ell - \rho(\mathbf{s}_i)\cos_{\ell-1})^2\sigma(\mathbf{s}_i)}\right) \times N(\beta_2 \mid a_{\beta_2}, b_{\beta_2}^2)
$$

$$[\beta_3 \mid \cdots] \propto \prod_{i=1}^{n}\prod_{t=1}^{T}\prod_{\ell=2}^{L} N\left(\beta_3 \mid \frac{Y_{t\ell}(\mathbf{s}_i) - Q^*_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i)) - c_1\xi_{t\ell}(\mathbf{s}_i)}{(1-\rho(\mathbf{s}_i))\mathrm{elev}(\mathbf{s}_i)}, \right.$$
$$\left. \frac{c_2\xi_{t\ell}(\mathbf{s}_i)}{(1-\rho(\mathbf{s}_i))^2\mathrm{elev}(\mathbf{s}_i)^2\sigma(\mathbf{s}_i)} \right) \times N(\beta_3 \mid a_{\beta_3}, b_{\beta_3}^2)$$

$$[Z_\rho \mid \cdots] \propto N\left(Z_\rho \mid \frac{\sum_{j,k} r_{jk} Z_\rho(\mathbf{s}_k)}{\sum_{j,k} r_{jk}}, \frac{\sigma_\rho^2}{\sum_{j,k} r_{jk}}\right) \times N(Z_\rho \mid a_\rho, b_\rho^2)$$

$$[Z_\sigma \mid \cdots] \propto N\left(Z_\sigma \mid \frac{\sum_{j,k} r_{jk} Z_\sigma(\mathbf{s}_k)}{\sum_{j,k} r_{jk}}, \frac{\sigma_\sigma^2}{\sum_{j,k} r_{jk}}\right) \times N(Z_\sigma \mid a_\sigma, b_\sigma^2)$$

- The full conditional distributions for $\sigma_\psi^2, \sigma_\eta^2, \sigma_{\beta_0}^2, \sigma_\alpha^2, \sigma_\rho^2, \sigma_\sigma^2$ are all inverse gamma as follows,

$$1/\sigma_\psi^2 \mid \cdots \sim G\left(\frac{T-1}{2} + a_{\sigma_\psi}, \frac{1}{2}\sum_{t=2}^{T}\psi_t^2 + b_{\sigma_\psi}\right)$$

$$1/\sigma_\eta^2 \mid \cdots \sim G\left(\frac{nT}{2} + a_{\sigma_\eta}, \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\gamma_t(\mathbf{s}_i) - (\tilde{\beta}_0(\mathbf{s}_i) + \tilde{\alpha}(\mathbf{s}_i)t + \psi_t)\right)^2 + b_{\sigma_\eta}\right)$$

$$1/\sigma_{\beta_0}^2 \mid \cdots \sim G\left(\frac{n}{2} + a_{\sigma_{\beta_0}}, \frac{1}{2}(\{\tilde{\beta}_0(\mathbf{s}_i)\} - \beta_0\mathbf{1})^\top R(\phi)^{-1}(\{\tilde{\beta}_0(\mathbf{s}_i)\} - \beta_0\mathbf{1}) + b_{\sigma_{\beta_0}}\right)$$

$$1/\sigma_\alpha^2 \mid \cdots \sim G\left(\frac{n}{2} + a_{\sigma_\alpha}, \frac{1}{2}(\{\tilde{\alpha}(\mathbf{s}_i)\} - \alpha\mathbf{1})^\top R(\phi)^{-1}(\{\tilde{\alpha}(\mathbf{s}_i)\} - \alpha\mathbf{1}) + b_{\sigma_\alpha}\right)$$

$$1/\sigma_\rho^2 \mid \cdots \sim G\left(\frac{n}{2} + a_{\sigma_\rho}, \frac{1}{2}(\{Z_\rho(\mathbf{s}_i)\} - Z_\rho\mathbf{1})^\top R(\phi)^{-1}(\{Z_\rho(\mathbf{s}_i)\} - Z_\rho\mathbf{1}) + b_{\sigma_\rho}\right)$$

$$1/\sigma_\sigma^2 \mid \cdots \sim G\left(\frac{n}{2} + a_{\sigma_\sigma}, \frac{1}{2}(\{Z_\sigma(\mathbf{s}_i)\} - Z_\sigma\mathbf{1})^\top R(\phi)^{-1}(\{Z_\sigma(\mathbf{s}_i)\} - Z_\sigma\mathbf{1}) + b_{\sigma_\sigma}\right)$$

- The full conditionals for the $\tilde{\beta}_0(\mathbf{s}_i)$'s and $\tilde{\alpha}(\mathbf{s}_i)$'s are normal. Note that we consider the hierarchical centering of these random effects to improve convergence behavior. For $i = 1, \ldots, n$, the full conditionals are

$$[\tilde{\beta}_0(\mathbf{s}_i) \mid \cdots] \propto \prod_{t=1}^{T} N\left(\tilde{\beta}_0(\mathbf{s}_i) \mid \gamma_t(\mathbf{s}_i) - \tilde{\alpha}(\mathbf{s}_i)t - \psi_t, \sigma_\eta^2\right)$$
$$\times N\left(\tilde{\beta}_0(\mathbf{s}_i) \mid \beta_0 + \frac{\sum_{k\neq i} r_{ik}(\beta_0 - \tilde{\beta}_0(\mathbf{s}_k))}{r_{ii}}, \frac{\sigma_{\beta_0}^2}{r_{ii}}\right)$$

$$[\tilde{\alpha}(\mathbf{s}_i) \mid \cdots] \propto \prod_{t=1}^{T} N\left(\tilde{\alpha}(\mathbf{s}_i) \mid (\gamma_t(\mathbf{s}_i) - \tilde{\beta}_0(\mathbf{s}_i) - \psi_t)/t, \sigma_\eta^2/t^2\right)$$
$$\times N\left(\tilde{\alpha}(\mathbf{s}_i) \mid \alpha + \frac{\sum_{k\neq i} r_{ik}(\alpha - \tilde{\alpha}(\mathbf{s}_k))}{r_{ii}}, \frac{\sigma_\alpha^2}{r_{ii}}\right)$$

- The full conditional distributions of the $Z_\rho(\mathbf{s}_i)$'s and $Z_\sigma(\mathbf{s}_i)$'s are non-standard. To draw samples from them, we suggest a random walk Metropolis-Hastings algorithm with normal distribution proposals with the mean at the current parameter value. The variance of the proposals was tuned until the acceptance rate was about 40%. For $i = 1, \ldots, n$, the full conditionals are proportional to

$$
[Z_\rho(\mathbf{s}_i) \mid \cdots]
$$
$$
\propto \prod_{t=1}^{T} \prod_{\ell=2}^{L} N\left(\frac{e^{Z_\rho(\mathbf{s}_i)} - 1}{e^{Z_\rho(\mathbf{s}_i)} + 1} \,\Bigg|\, \frac{Y_{t\ell}(\mathbf{s}_i) - q_{t\ell}(\mathbf{s}_i) - c_1 \xi_{t\ell}(\mathbf{s}_i)}{Y_{t,\ell-1}(\mathbf{s}_i) - q_{t,\ell-1}(\mathbf{s}_i)}, \frac{c_2 \xi_{t\ell}(\mathbf{s}_i)}{(Y_{t,\ell-1}(\mathbf{s}_i) - q_{t,\ell-1}(\mathbf{s}_i))^2 \sigma(\mathbf{s}_i)} \right)
$$
$$
\times N\left(Z_\rho(\mathbf{s}_i) \,\Bigg|\, Z_\rho + \frac{\sum_{k \neq i} r_{ik}(Z_\rho - Z_\rho(\mathbf{s}_k))}{r_{ii}}, \frac{\sigma_\rho^2}{r_{ii}} \right)
$$

$$
[Z_\sigma(\mathbf{s}_i) \mid \cdots]
$$
$$
\propto G\left(\exp\{Z_\sigma(\mathbf{s}_i)\} \,\Bigg|\, \frac{3}{2} T(L-1) + 1, \sum_{t=1}^{T} \sum_{\ell=2}^{L} \left[\frac{(Y_{t\ell}(\mathbf{s}_i) - Q_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i)) - c_1 \xi_{t\ell}(\mathbf{s}_i))^2}{2 c_2 \xi_{t\ell}(\mathbf{s}_i)} + \xi_{t\ell}(\mathbf{s}_i) \right] \right)
$$
$$
\times N\left(Z_\sigma(\mathbf{s}_i) \,\Bigg|\, Z_\sigma + \frac{\sum_{k \neq i} r_{ik}(Z_\sigma - Z_\sigma(\mathbf{s}_k))}{r_{ii}}, \frac{\sigma_\sigma^2}{r_{ii}} \right)
$$

- We obtain the normal full conditionals for the ψ_t's as follows. For identifiability, ψ_1 is fixed to zero. Then, for $t = 2, \ldots, T$, we have

$$
[\psi_t \mid \cdots] \propto \prod_{i=1}^{n} N\left(\psi_t \mid \gamma_t(\mathbf{s}_i) - \tilde{\beta}_0(\mathbf{s}_i) - \tilde{\alpha}(\mathbf{s}_i)t, \sigma_\eta^2 \right) \times N\left(\psi_t \mid 0, \sigma_\psi^2 \right)
$$

- The full conditionals for the $\gamma_t(\mathbf{s}_i)$'s are all normal. For $i = 1, \ldots, n$, and $t = 1, \ldots, T$,

$$
[\gamma_t(\mathbf{s}_i) \mid \cdots] \propto \prod_{\ell=2}^{L} N\left(\gamma_t(\mathbf{s}_i) \,\Bigg|\, \frac{Y_{t\ell}(\mathbf{s}_i) - Q^*_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i)) - c_1 \xi_{t\ell}(\mathbf{s}_i)}{(1 - \rho(\mathbf{s}_i))}, \frac{c_2 \xi_{t\ell}(\mathbf{s}_i)}{(1 - \rho(\mathbf{s}_i))^2 \sigma(\mathbf{s}_i)} \right)
$$
$$
\times N\left(\gamma_t(\mathbf{s}_i) \mid \tilde{\beta}_0(\mathbf{s}_i) + \tilde{\alpha}(\mathbf{s}_i)t + \psi_t, \sigma_\eta^2 \right)
$$

- Finally, we sample each one of the $n \times T \times (L-1)$ parameters $\xi_{t\ell}(\mathbf{s}_i)$ using a Metropolis-Hastings step. This is the main computational bottleneck, since the number of parameters is equal to the number of data. Following Lum and Gelfand (2012), the proposal distribution for each $\xi_{t\ell}(\mathbf{s}_i)$ is an exponential distribution with rate $\sigma(\mathbf{s}_i)$, as is specified in the hierarchy.

$$
[\xi_{t\ell}(\mathbf{s}_i) \mid \cdots]
$$
$$
\propto \xi_{t\ell}(\mathbf{s}_i)^{-1/2} \exp\left\{ \frac{-\sigma(\mathbf{s}_i)}{2 c_2 \xi_{t\ell}(\mathbf{s}_i)} \left[Y_{t\ell}(\mathbf{s}_i) - Q_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i)) - c_1 \xi_{t\ell}(\mathbf{s}_i) \right]^2 - \sigma(\mathbf{s}_i) \xi_{t\ell}(\mathbf{s}_i) \right\}
$$

Note in the expressions above that the product of normal densities is proportional to a normal density with parameters as follows

$$
\prod_{i=1}^{n} N\left(x \mid \mu_i, \sigma_i^2 \right) \propto N\left(x \,\Bigg|\, \sum_{i=1}^{n} \frac{\mu_i}{\sigma_i^2} \Big/ \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, 1 \Big/ \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right).
$$

Note that an alternative procedure to generate the $\xi_{t\ell}(\mathbf{s})$'s would characterize their full conditional as generalized inverse Gaussian (GIG) (Kozumi and Kobayashi, 2011), from which samples could be generated using the algorithm proposed by Dagpunar (1989). Let $X \sim GIG(a, b, \nu)$, then its pdf is

$$f(x \mid a, b, \nu) \propto x^{\nu-1} \exp\left\{-(ax + b/x)/2\right\}.$$

With this parameterization the full conditional for $\xi_{t\ell}(\mathbf{s}_i)$ follows

$$\xi_{t\ell}(\mathbf{s}_i) \mid \cdots \sim GIG\left(\sigma(\mathbf{s}_i)\left(2 + \frac{c_1^2}{c_2}\right), \sigma(\mathbf{s}_i)\frac{\left(Y_{t\ell}(\mathbf{s}_i) - Q_{Y_{t\ell}(\mathbf{s}_i)}(\tau \mid Y_{t,\ell-1}(\mathbf{s}_i))\right)^2}{c_2}, \frac{1}{2}\right).$$

S4. Results of model fitting

Here, we follow Section 3.3 of the Main Manuscript with additional figures.

Figure S7 shows the parameters not shown in Figure 3 of the Main Manuscript, some general conclusions are as follows. We observe that the parameters β_1^τ and β_2^τ are very different across quantiles, however, the seasonal pattern they generate is relatively similar as we can see in Figure 3. The parameter σ_ψ^τ suggests greater variability across years in quantiles close to the extremes of the left tail than in the other quantiles that show similar variability. The parameter σ_η^τ suggests that the difference across yearly shifts across locations is small for the median and increases for the extreme quantiles.

Figure S8 shows boxplots of the four spatial processes at the 18 observed sites for $\tau = 0.05, 0.50, 0.95$. This figure is complementary to Figures 4 and 5 of the Main Manuscript.

Figure S9 shows the posterior mean and 90% credible intervals of the spatial random effects, $\tilde{\beta}_0^\tau(\mathbf{s}), \tilde{\alpha}^\tau(\mathbf{s}), \rho^\tau(\mathbf{s})$ and $\sigma^\tau(\mathbf{s})$ at Zaragoza across quantiles. This figure supports Figure 4 of the Main Manuscript to give an idea of the uncertainties of the estimates, in particular, only Zaragoza is shown but the width of the credible intervals are very similar for all locations.

Figure S10 shows whether an 80% credible interval of the spatially varying linear trend $\tilde{\alpha}(\mathbf{s})$ contains zero for $\tau = 0.05, 0.50, 0.95$. There is strong evidence of region-wide warming in the center of the distribution, however there is not as much evidence at the extremes.

Figure S11 shows boxplots of the ψ's across years for $\tau = 0.05, 0.50, 0.95$. We note the similarity of the distribution of $\psi_t^{0.50}$ and $\psi_t^{0.95}$, while $\psi_t^{0.05}$ is slightly different. It is observed that the effects may add or subtract in a given year up to roughly 3°C.
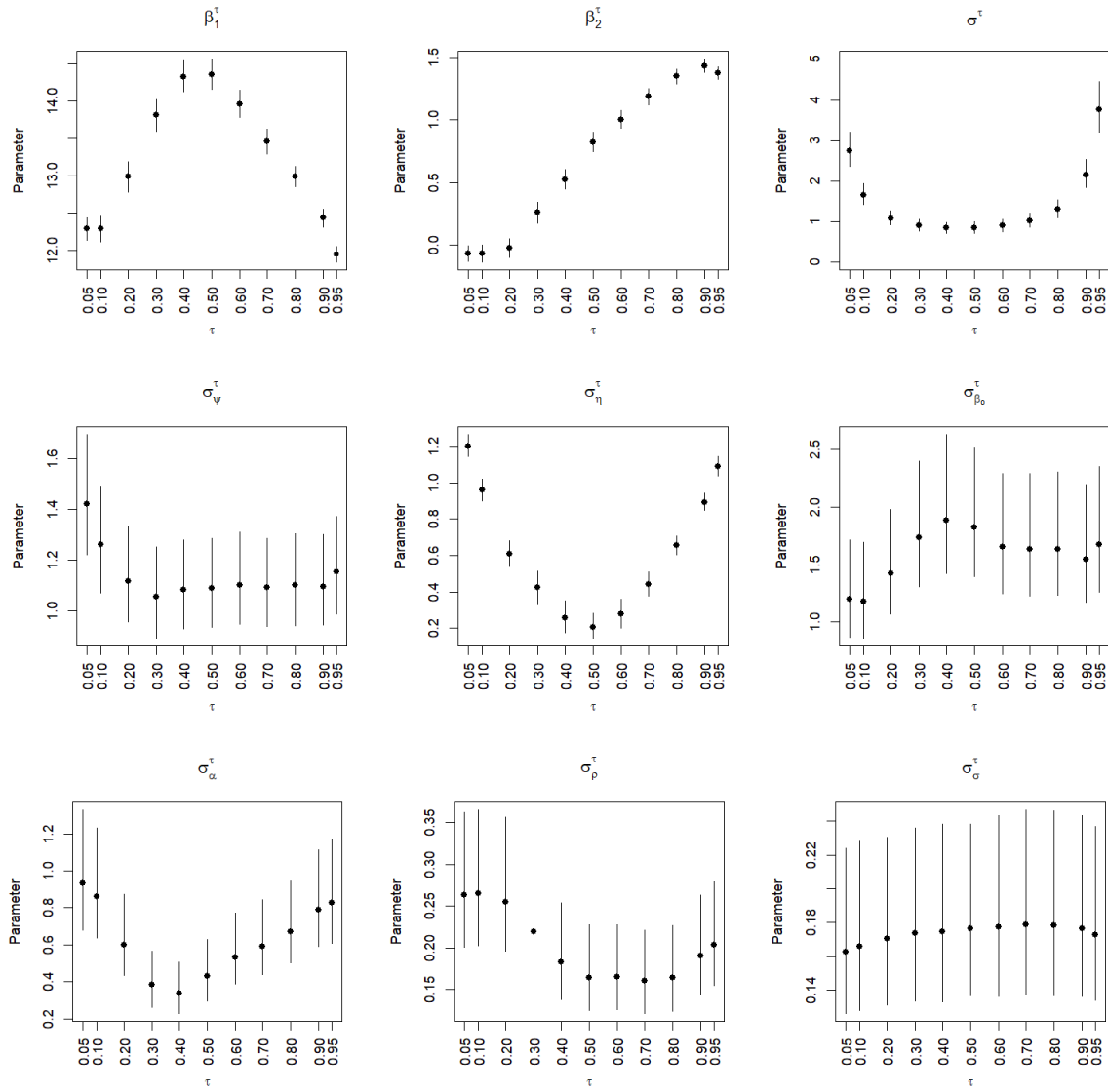
FIG S7. *Posterior median and* 90% *credible interval of the parameters not shown in Figure 3 of the Main Manuscript across quantiles.*
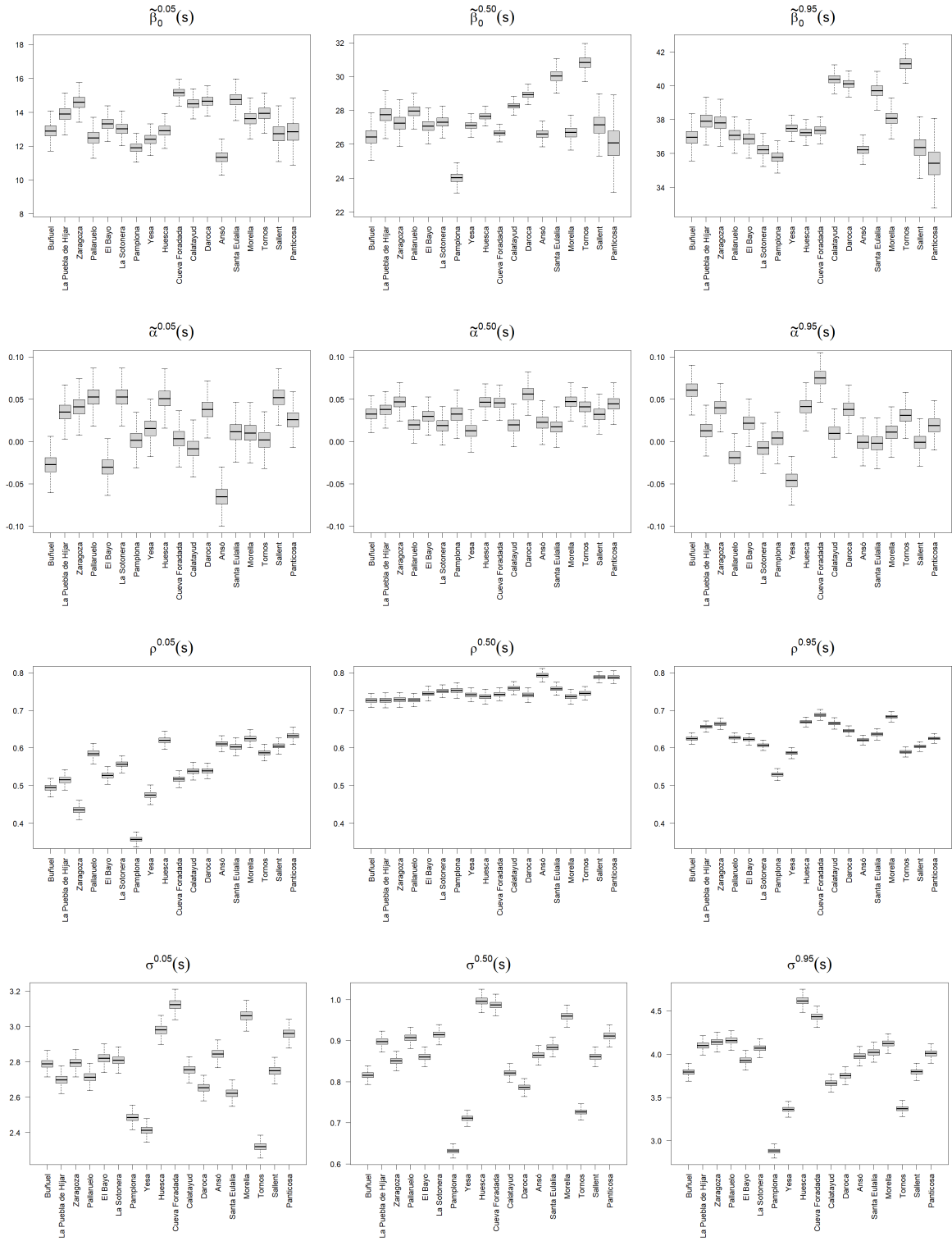
Fig S8. *Boxplots of the posterior distribution of $\tilde{\beta}_0^\tau(s), \tilde{\alpha}^\tau(s), \rho^\tau(s), \sigma^\tau(s)$ for $\tau = 0.05, 0.50, 0.95$. Locations are sorted by elevation, from lowest to highest.*
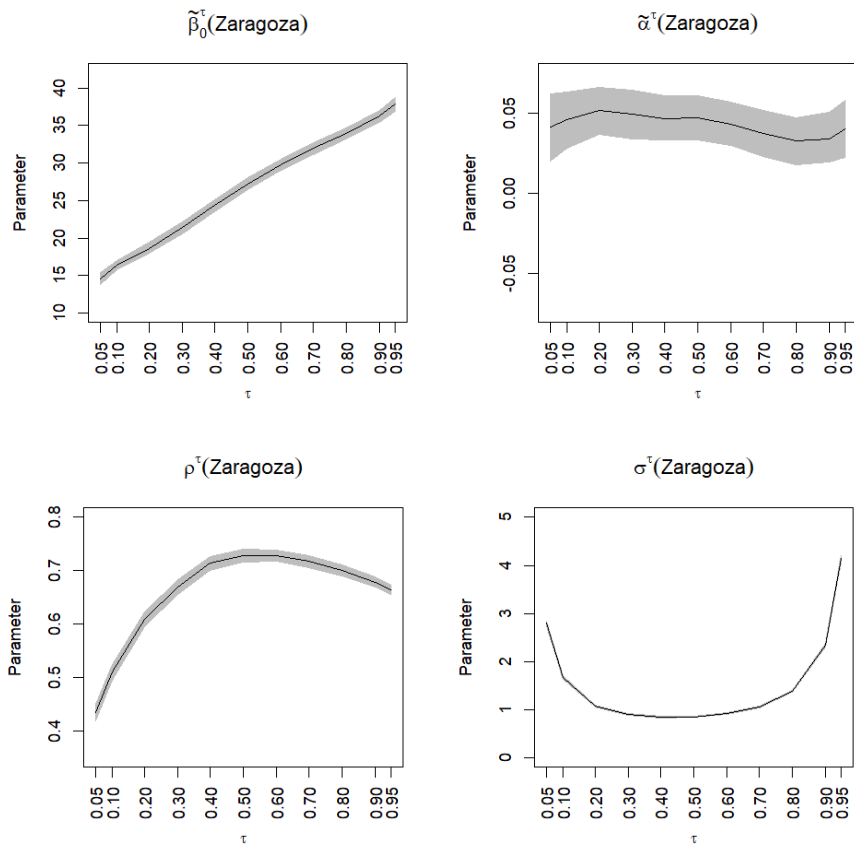
FIG S9. *Posterior mean and 90% credible intervals of the four spatial random effects in Zaragoza across quantiles.*
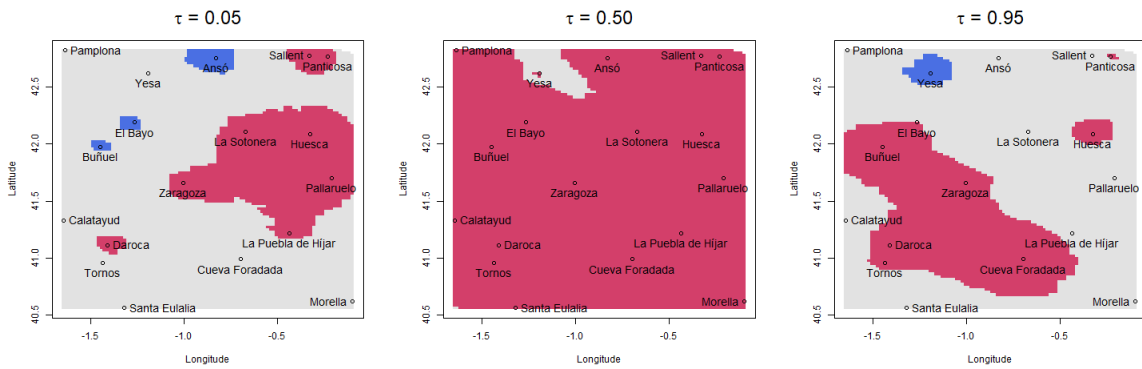
FIG S10. *Summary map of the 80% credible intervals of $\tilde{\alpha}^\tau(s)$ for $\tau = 0.05, 0.50, 0.95$. Blue means that the 80% credible interval does not contain zero and is negative, red means that the 80% credible interval does not contain zero and is positive, and gray means that the 80% credible interval contains zero.*
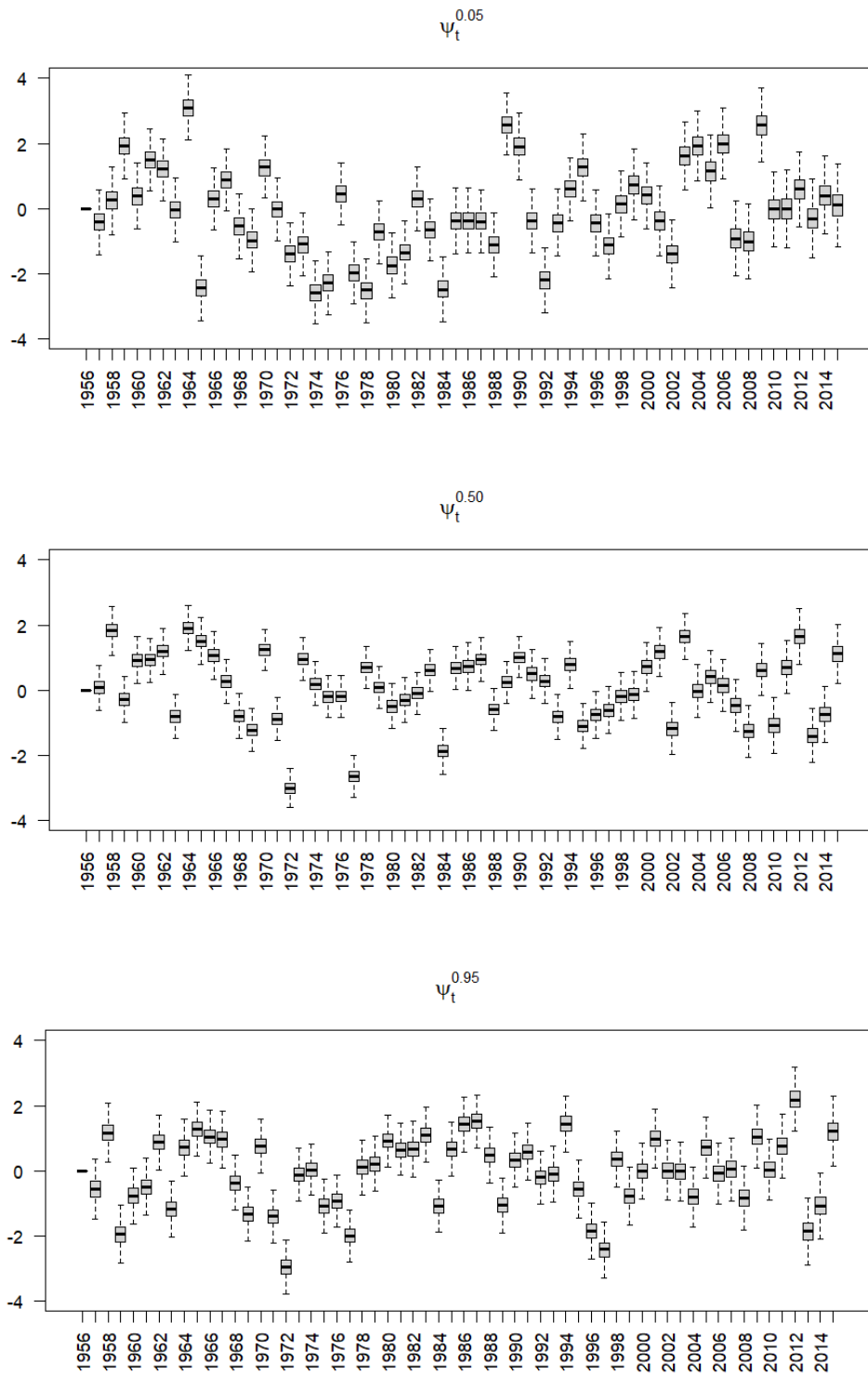
FIG S11. *Boxplots of the posterior distributions of the annual random effects* ψ_t^τ *with* $\tau = 0.05, 0.50, 0.95$.

References

DAGPUNAR, J. S. (1989). An easily implemented generalised inverse Gaussian generator. *Communications in Statistics - Simulation and Computation* **18** 703–710.

KOZUMI, H. and KOBAYASHI, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation* **81** 1565–1578.

LI, G., LI, Y. and TSAI, C.-L. (2015). Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association* **110** 246–261.

LUM, K. and GELFAND, A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis* **7** 235–258.

3.4 Bayesian joint quantile autoregression

This manuscript was published in:

Castillo-Mateo, J., Gelfand, A. E., Asín, J., Cebrián, A. C., & Abaurrea, J. (*in press*). Bayesian joint quantile autoregression. *TEST*. [arXiv:2305.19080]

And it was disseminated (speaker emphasized) in:

- ***Castillo-Mateo, J.***, Gelfand, A. E., Asín, J., & Cebrián, A. C. (2023, July 18–21). *Joint quantile autoregression for space-time data* [Contributed talk]. Spatial Statistics 2023 Conference, Boulder, CO, USA.

- ***Castillo-Mateo, J.***, Gelfand, A. E., Asín, J., & Cebrián, A. C. (2023, June 27–30). *Joint quantile autoregressive modeling for univariate and spatial time-series data in a Bayesian framework* [Contributed talk]. XIX Conferencia Española y VIII Encuentro Iberoamericano de Biometría, Vigo, Spain. *Award for the best work developed by a young researcher.*

- ***Castillo-Mateo, J.***, Gelfand, A. E., Asín, J., & Cebrián, A. C. (2023, May 22–24). *Quantile autoregression* [Contributed talk]. 13th Bayesian Inference for Stochastic Processes, Madrid, Spain.

"Statisticians, like artists, have the bad habit of falling in love with their models."

George E. P. Box

Bayesian joint quantile autoregression

Jorge Castillo-Mateo[1*], Alan E. Gelfand[2], Jesús Asín[1],
Ana C. Cebrián[1], Jesús Abaurrea[1]

[1]Department of Statistical Methods, University of Zaragoza, Pedro
Cerbuna 12, Zaragoza, 50009, Spain.
[2]Department of Statistical Science, Duke University, 415 Chapel Dr,
Durham, 27705, NC, USA.

*Corresponding author(s). E-mail(s): jorgecm@unizar.es;
Contributing authors: alan@stat.duke.edu; jasin@unizar.es;
acebrian@unizar.es; abaurrea@unizar.es;

Abstract

Quantile regression continues to increase in usage, providing a useful alternative to customary mean regression. Primary implementation takes the form of so-called *multiple* quantile regression, creating a separate regression for each quantile of interest. However, recently, advances have been made in *joint* quantile regression, supplying a quantile function which avoids crossing of the regression across quantiles. Here, we turn to quantile autoregression (QAR), offering a fully Bayesian version. We extend the initial quantile regression work of Koenker and Xiao (2006) in the spirit of Tokdar and Kadane (2012). We offer a directly interpretable parametric model specification for QAR. Further, we offer a p-th order QAR(p) version, a multivariate QAR(1) version, and a spatial QAR(1) version. We illustrate with simulation as well as a temperature dataset collected in Aragón, Spain.

Keywords: copula model, Gaussian process, joint quantile model, Markov chain Monte Carlo, spatial quantile autoregression

MSC Classification: 62F15 , 62G08 , 62H05 , 62M10 , 62M30

1

1 Introduction

For time series data, autoregressive (AR) modeling is perhaps the most common approach. A lag one, AR(1), takes the form $Y_t = \mu + \rho(Y_{t-1} - \mu) + \epsilon_t$, with ϵ_t following a suitable zero-mean error distribution; a conditional mean is provided. By analogy, quantile autoregression (QAR) considers conditional quantiles.

An issue with quantile regression (QR) is the so-called quantile crossing problem. Modeling quantiles individually enables rich modeling for a given quantile but allows for crossing of quantiles across quantile level τ. For arbitrary values of the regressors, $\mathbf{X}$, we can not ensure that the resulting modeled quantiles will increase in τ. Such modeling is referred to as *multiple* QR. Inference typically proceeds by minimizing a check loss function or, more formally, assuming an asymmetric Laplace (AL) error term. Examples of multiple QR with AL errors appear in Yu and Moyeed (2001), and Kozumi and Kobayashi (2011) present a Gibbs sampler model fitting implementation. Following those ideas, Peng et al. (2023) deals with variable selection in the context of QAR models with AL errors. Lum and Gelfand (2012) work in the context of spatially referenced data and extend the AL model to a spatial process. Castillo-Mateo et al. (2023) propose a very flexible spatial AL mixed effects QAR model.

Recent effort has focused on a *joint* QR modeling to avoid quantile crossing. Adopting restricted support for the regressors, $\mathbf{X}$, the τ-quantile will increase monotonically over $\tau \in (0, 1)$. Bondell et al. (2010) offer a non-crossing approach for a fixed set of quantiles of interest. Foundational work appears in Tokdar and Kadane (2012) using Gaussian process's (GP's) with follow on work in Das and Ghosal (2017) using splines. Reich et al. (2011) developed a spatial joint QR model through spatially varying regression coefficients using Bernstein polynomials. Yang and Tokdar (2017) propose a novel parameterization that characterizes any collection of non-crossing quantile planes over arbitrarily shaped convex predictor domains. This parameterization was extended to spatial data by Chen and Tokdar (2021) through a copula process but a non-spatially varying quantile function results. Joint modeling imposes strong restrictions on the class of permissible specifications; models outside of this class may be preferred.

Motivation for joint or non-crossing quantile modeling appears in, e.g., Bondell et al. (2010) who highlight a problem that appears when modeling a wind speed dataset given climatological regressors; the upper quantiles cross not far from the mean. As a further example in this regard, accurate quantile predictions across quantile levels are essential in forecasting of wind power generation (Cui et al., 2023). Formal joint modeling is necessary in applications where coherent estimates of several quantiles or a generative model are of interest; otherwise, quantile crossing leads to an invalid distribution for the response. Also, multiple modeling fails to do justice to the full potential of the model. Joint modeling helps to avoid the lack of data which emerges when attempting to fit individual quantile curves (Tokdar and Kadane, 2012). Joint modeling in the context of QAR is relevant in risk management for estimating value-at-risk, as well as in demand forecasting, where understanding the complete demand distribution is crucial for effective production planning and supply chain management. Further, these models for daily temperatures could improve operational prediction accuracy with forecast intervals, and high quantile simulated series behavior (Thrasher et al., 2012).

2

Koenker and Xiao (2006) offered an initial version of a joint QAR(p) model. Illustrating with $p = 1$, they consider the generative model

$$Y_t = \theta_0(U_t) + \theta_1(U_t)Y_{t-1}, \tag{1}$$

where U_t is a sequence of independent and identically distributed (i.i.d.) standard uniform random variables. The θ functions, from $[0,1] \to \mathbb{R}$, need to be estimated. Provided that the right side of expression (1) is monotone increasing in U_t, the τ conditional quantile function of Y_t given y_{t-1} increases in τ and is

$$Q_{Y_t}(\tau \mid y_{t-1}) = \theta_0(\tau) + \theta_1(\tau)y_{t-1}. \tag{2}$$

Koenker and Xiao (2006) required both θ_0 and θ_1 to be strictly increasing functions (referred to as co-monotonicity). Their suggested choices were $\theta_0(\tau) = \sigma\Phi^{-1}(\tau)$ with Φ the cumulative distribution function (cdf) of a standard normal distribution and $\theta_1(\tau) = \min\{\gamma_0 + \gamma_1\tau, 1\}$ for $\gamma_0 \in (0,1)$ and $\gamma_1 > 0$. If $y_{t-1} \geq 0$, co-monotonicity ensures that $Q_{Y_t}(\tau \mid y_{t-1})$ will not cross as τ increases but under the restrictive assumption that the autoregression coefficient strictly increases in τ.

Our contribution is to reconsider the work of Koenker and Xiao (2006) in the context of Tokdar and Kadane (2012), providing flexible joint QAR modeling in a Bayesian framework. We characterize non-crossing QAR(1) also using two monotone curves, through a convenient class of cdf's. We note extension to the QAR(p) model. We consider bivariate QAR, capturing dependence through a copula process. Then, for spatially referenced time series, we introduce spatial dependence in the realizations and obtain spatially varying QAR's through spatially varying coefficients.

QAR models arise when time series are observed to display asymmetric dynamics; such data often appears in economic applications. Koenker and Xiao (2006) show empirical applications of the QAR model to the USA unemployment rate and gasoline prices. Further examples in the literature consider dynamic additive quantile models, QR with cointegrated time series, and conditional quantiles with GARCH models. Applications include stock returns, house price returns, and gold prices. See, e.g., Li et al. (2015) or Yang et al. (2023) and references therein. QR is also popular for climate data (see Gao and Franzke, 2017, for an extensive review). Yang et al. (2018) propose a semiparametric QAR model including lagged data to develop quantile-based temperature extreme indices. Zhang et al. (2022) use QR models conditional on the state of the previous observation time to predict short-term wind speed or velocity. Castillo-Mateo et al. (2023) use a rich QAR model to compare the effects of climate change in daily maximum temperature.

The outline of the paper is as follows. Section 2 provides a model characterization for the QAR(1) case. Further, it offers explicit parametric model specifications, the resulting likelihood for Bayesian model fitting, some criteria for model assessment, and a simulation study. Section 3 looks at the QAR(p) case. Section 4 considers the bivariate QAR(1) setting. Section 5 develops a fully spatial version through the use of a Gaussian copula. Section 6 employs time series of daily temperature data from 18 spatial locations to illustrate the previous four sections. Finally, Section 7 presents a brief summary and possibilities for future work.

2 The QAR(1) case

2.1 The support of the data

For a non-crossing *linear* QAR specification we need to restrict the support of the time series data, $\{y_t^* : t = 1, \ldots, T\}$, to a bounded interval on the real line.[1] We take this interval to be $[0, 1]$ and implement this by making a transformation of the data,

$$y_t = \frac{y_t^* - m}{M - m}, \tag{3}$$

where $m < \min y_t^*$ and $M > \max y_t^*$. In fact, m and M are chosen such that $\min y_t$ is close to but above 0 and $\max y_t$ is close to but below 1. This enables the most flexibility for the quantile function under our proposed QAR modeling and we offer an automatic selection approach below.

Two points are important to note. First, we can not take $m = \min y_t^*$ and $M = \max y_t^*$. The data must be in the interior of the unit interval in order to enable distinct quantiles as τ varies across $(0, 1)$. Second, choosing m and M is merely a device for working on the unit interval. There is no connection between these values and the potential practical support of the y_t^*'s. Imposing bounding on the support is unavoidable for a valid linear specification of $Q_{Y_t}(\tau \mid y_{t-1})$ of the form $\theta_0(\tau) + \theta_1(\tau)y_{t-1}$ because the only non-intersecting lines under unbounded support are parallel lines.

A convenient "automatic" strategy for selecting m and M is as follows. The idea recalls basic results from the theory of order statistics. If we have T independent observations from a uniform distribution on (m, M), $\{y_t^* : t = 1, \ldots, T\}$, then $[E(Y_{(1)}^*) - m]/(M - m) = 1/(T + 1)$ and $[E(Y_{(T)}^*) - m]/(M - m) = T/(T + 1)$. So we can say $y_{(1)}^* \approx (mT + M)/(T + 1)$ and $y_{(T)}^* \approx (m + TM)/(T + 1)$. This gives two equations in two unknowns to solve for m and M. We obtain

$$m = \frac{Ty_{(1)}^* - y_{(T)}^*}{T - 1} \qquad \text{and} \qquad M = \frac{Ty_{(T)}^* - y_{(1)}^*}{T - 1}. \tag{4}$$

Of course the Y_t^*'s are not independent, they do not come from a distribution on a bounded interval, and marginally, we would not expect them to follow a uniform distribution on (m, M). We only implement a simple automatic bounding strategy.

2.2 The model

A straightforward characterization of the required monotonicity of the QAR lines is offered by the following result, inspired from Tokdar and Kadane (2012).

Theorem 1. *An autoregressive specification of the form of* (2) *with* $\theta_1(\tau) \in [-1, 1]$ *for* $\tau \in [0, 1]$*, is monotonically increasing in* τ *for* Y_t *taking values in* $[0, 1]$ *and* $y_{t-1} \in [0, 1]$ *if and only if*

$$Q_{Y_t}(\tau \mid y_{t-1}) = y_{t-1}\eta_1(\tau) + (1 - y_{t-1})\eta_2(\tau) \tag{5}$$

where $\eta_1, \eta_2 : [0, 1] \longrightarrow [0, 1]$ *are monotonically increasing.*

---

[1]This is the analogue of the restriction over the predictor domain in Yang and Tokdar (2017).

Proof. Any monotonicity obeying $Q_{Y_t}(\tau \mid y_{t-1})$ given by (2) can be expressed as (5) by taking $\eta_1(\tau) = \theta_0(\tau) + \theta_1(\tau) = Q_{Y_t}(\tau \mid 1)$ and $\eta_2(\tau) = \theta_0(\tau) = Q_{Y_t}(\tau \mid 0)$. For the converse, if $Q_{Y_t}(\tau \mid y_{t-1})$ is given by (5) then it must be monotonically increasing in τ for every $y_{t-1} \in [0,1]$ for which both y_{t-1} and $1 - y_{t-1}$ are non-negative. One can express such a $Q_{Y_t}(\tau \mid y_{t-1})$ by defining $\theta_0(\tau) = \eta_2(\tau)$ and $\theta_1(\tau) = \eta_1(\tau) - \eta_2(\tau) \in [-1, 1]$. $\qquad\square$

If we focus on (5), a model for functions η_1 and η_2, each from $[0,1] \to [0,1]$, induces a QAR(1) model over all valid QAR(1) specifications of $Q_{Y_t}(\tau \mid y_{t-1})$, provided the boundary conditions $Q_{Y_t}(0 \mid y_{t-1}) = 0$ and $Q_{Y_t}(1 \mid y_{t-1}) = 1$ for all $y_{t-1} \in [0,1]$ are satisfied. The above condition can be rewritten as $\eta_j(0) = 0$ and $\eta_j(1) = 1$ for $j = 1, 2$. Next we show how to specify these two monotone functions.

2.2.1 Specification for the two monotone curves

Specifically, both $\eta_1(\cdot)$ and $\eta_2(\cdot)$ again must be strictly monotone from $[0,1] \to [0,1]$. A convenient class to work with are cdf's for continuous random variables with support $[0,1]$. In fact, a rich class would arise as probabilistic mixtures of such cdf's, leading to the general form

$$\eta(\tau) = \sum_{k=1}^{K} \lambda_k F(\tau \mid \mathbf{\Omega}_k) \qquad (6)$$

such that $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$ and $F : [0,1] \to [0,1]$ is strictly increasing for any parameters $\mathbf{\Omega}_k$.

A convenient class of F's to work with are the cdf's of the two parameter Kumaraswamy (1980) distribution (also known as the minimax distribution, Jones, 2009). Specifically, the probability density function (pdf) and cdf are

$$f(x \mid a, b) = abx^{a-1}(1 - x^a)^{b-1} \quad \text{and} \quad F(x \mid a, b) = 1 - (1 - x^a)^b, \qquad (7)$$

where $x \in [0,1]$ and $a, b > 0$. The Kumaraswamy distributions are a family with behavior similar to the beta distribution. However, for our purposes, they are much simpler to use especially in the context of simulation since the cdf and quantile function can be expressed in closed form, i.e., $Q(\tau \mid a, b) = (1 - (1 - \tau)^{1/b})^{1/a}$ where $\tau \in [0,1]$. The flexibility of the Kumaraswamy distributions is shown in Section S1 of the Supplementary Information (SI) employing different combinations of parameters (a, b).

To work with the mixture form for $\eta(\tau)$, we investigated two mixture strategies. The first lets K be small but assumes the a's and b's are unknowns. The second lets K be larger but adopts a fixed set of a's and b's, in the spirit of basis function forms. Specifically, we consider K Kumaraswamy distributions with medians $k/(K + 1)$, respectively. In the former, with $K = 2$ we have a total of five parameters (two a's, two b's, and a λ) while in the latter, with $K = 6$ again we have five parameters (five λ's). Increasing the number of "basis" components in the specification of the η's need not provide better model performance. From considerable simulation experience, model performance is very sensitive to the choice of parameters in the mixture components. So, in the sequel, we work with $K = 1$ or 2 (QAR1K1 and QAR1K2, hereafter) and fit the a's and b's. As for priors, with $K = 1$, we consider $\log a_1, \log b_1 \sim N(0, \sigma_{ab}^2)$

5

with $\sigma_{ab} = 3$, which gives a weak prior on the log-scale. With $K = 2$, we consider $\lambda_1 \sim U(0, 1/2)$ and $\log a_1, \log a_2, \log b_1, \log b_2 \sim N(0, \sigma_{ab}^2)$ with $\sigma_{ab} = 1.5$. Restricting λ_1 to $(0, 1/2)$ avoids identification issues, while σ_{ab} is taken smaller than in the $K = 1$ case to penalize values of a's and b's too small or large. Values of a's and b's that are close to zero or very large can cause negligible numerical errors in the root-finder to generate a numerical overflow in the likelihood (see Equations 8 and 9 below) and thus degeneracy.

2.3 Likelihood evaluation and model fitting

An important feature of a valid joint specification of $Q_{Y_t}(\tau \mid y_{t-1})$ for all $\tau \in (0, 1)$, following Tokdar and Kadane (2012), is that it uniquely defines the conditional response density given $y_{t-1} \in [0, 1]$. Specifically, this density is given by

$$f_{Y_t}(y_t \mid y_{t-1}) = \left. \frac{1}{\frac{d}{d\tau} Q_{Y_t}(\tau \mid y_{t-1})} \right|_{\tau = \tau_{y_{t-1}}(y_t)}, \tag{8}$$

where $\tau_{y_{t-1}}(y_t)$ solves $y_t = y_{t-1}\eta_1(\tau) + (1 - y_{t-1})\eta_2(\tau)$ in τ and is numerically approximated to arbitrary precision via a one-dimensional root-finder. We implement the hybrid root-finding algorithm combining the bisection method, the secant method, and inverse quadratic interpolation, so-called Brent's method (Brent, 1973). Consequently, given the data at $t = 1$, y_1, we can write a valid log-likelihood score

$$\begin{aligned} \ell(\mathbf{\Omega} \mid \mathbf{y}) &= \sum_{t=2}^{T} \log f_{Y_t}(y_t \mid y_{t-1}) \\ &= -\sum_{t=2}^{T} \log \left\{ y_{t-1}\dot{\eta}_1(u_t) + (1 - y_{t-1})\dot{\eta}_2(u_t) \right\}, \end{aligned} \tag{9}$$

where $u_t = \tau_{y_{t-1}}(y_t)$, $\mathbf{y}^\top = (y_1, \ldots, y_T)$ are all of the observed data, $\mathbf{\Omega}$ are the model parameters, and the $\dot{\eta}$'s are the derivatives of the η's.

We implement a block-Metropolis sampler algorithm with an adaptive period (Haario et al., 2001) during warm-up to obtain Markov chain Monte Carlo (MCMC) samples from the posterior distribution of the parameters and to summarize the posterior distribution of the conditional quantile function. Furthermore, with a posterior realization of the model parameters and a given value of y_{t-1}, we can use (8) with discretization, to obtain a posterior realization of the density function that is driving the joint quantiles. Averaging over these realizations provides the posterior mean of the density.

Expression (8) reveals an important difference between our QAR approach and other nonlinear joint modeling versions in the literature. For example, the nonlinear QAR model in Chen et al. (2009) specifies a joint dist for (Y_t, Y_{t-1}) using a copula. It yields a conditional distribution for $Y_t \mid Y_{t-1}$ which has a nonlinear quantile function that is monotone in τ. What we do is the reverse. We specify a non-crossing quantile function and obtain the induced conditional distribution for $Y_t \mid Y_{t-1}$. Our quantile

function is also nonlinear as a function of Y_{t-1}. Their quantile function depends upon the choice of copula and the copula parameters. Our quantile function depends on the Kumaraswamy distribution and the associated parameters. Their Gaussian version has a conditional quantile function which is linear in Y_{t-1}, which may be restrictive. Their t-copula version yields a quantile function which has a perhaps unattractive form as the square root of a function of the square of Y_{t-1}. If the goal is to model the quantile function directly as nonlinear and flexible, rather than seeing what is induced by a copula, our approach yields a simple form and may be more attractive.

2.4 Model comparison and simulation study

Working within our parametric Bayesian framework, for any τ, posterior samples of the model parameters, $\{\mathbf{\Omega}_b^* : b = 1, \ldots, B\}$, produce posterior samples of the conditional quantile function for Y_t, $Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega}_b^*)$. Essentially, for each Y_t (with associated y_{t-1}) and any τ, we obtain the posterior distribution of $Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega})$. We use these posterior distributions along with the dataset, $\mathbf{y}$, to offer model assessment.

We propose two novel approaches. First, for any y, consider $\mathbf{1}(y < Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega}))$ where $\mathbf{1}$ denotes the indicator function. Then, let $p_t(\tau) \equiv E[\mathbf{1}(y_t < Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega})) \mid \mathbf{y}]$, i.e., the posterior probability that $Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega})$ exceeds y_t. Suppose we compute $p(\tau) \equiv \sum_{t=2}^T p_t(\tau)/(T-1)$. We note that for any τ and Y regardless of its distribution, $E[\mathbf{1}(Y < Q_Y(\tau)] = \tau$ and $Var[\mathbf{1}(Y < Q_Y(\tau)] = \tau(1-\tau)$. If we let $v \geq 1$ be a real number, then

$$\tilde{p}_v \equiv \sqrt[v]{\int_0^1 \left| \frac{p(\tau) - \tau}{\sqrt{\tau(1-\tau)/(T-1)}} \right|^v d\tau} \tag{10}$$

provides a standardized deviation form as a dimensionless measure of how well the quantile function under the model is capturing conditional quantiles for the given time series. We propose this as a (*global*) measure of model accuracy. With a minimum value of zero, a smaller $\tilde{p}_v$ indicates better accuracy of the model. We would approximate the integral by discretizing τ, in particular, we consider $\tau \in \{0.01, 0.02, \ldots, 0.99\}$.

As a second measure, we turn to the check loss function, usually employed as an optimality function to obtain the τ empirical quantile (Koenker and Bassett, 1978). Here, we adopt $\delta_\tau(u) = u(\tau - \mathbf{1}(u < 0))$, the check loss function associated with the AL distribution. Again, from the posterior distribution of $Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega})$, for any Y_t (with associated y_{t-1}) and τ, we can obtain $\Delta_t(\tau) \equiv \delta_\tau(y_t - E[Q_{Y_t}(\tau \mid y_{t-1}; \mathbf{\Omega}) \mid \mathbf{y}])$. As above suppose we compute $\Delta(\tau) \equiv \sum_{t=2}^T \Delta_t(\tau)/(T-1)$. Then, for a given τ, $\Delta(\tau)$ provides an average discrepancy for the τ quantile function. The smaller the value, the better the model performance. Then, we propose to weight $\Delta(\tau)$,

$$\tilde{\Delta} \equiv \int_0^1 \omega(\tau) \Delta(\tau) \, d\tau \tag{11}$$

to provide a global measure of model performance. We propose this as a *relative* measure of model performance in making model comparison. The weighting function, $\omega(\tau)$, compensates for the variation in mean of $\Delta(\tau)$ across τ. Again, we would approximate the integral by discretizing τ.

For the weight function, we consider

$$\omega(\tau \mid \mathbf{y}) = \frac{1}{\sum_{t=2}^{T} \delta_\tau(y_t - Q_Y^{emp}(\tau))/(T-1)}. \tag{12}$$

This choice leads to a measure that is closely related to the $R^1(\tau)$ metric by Koenker and Machado (1999). The $R^1(\tau)$ measure is essentially, $1 - \omega(\tau \mid \mathbf{y})\Delta(\tau)$. This measure is viewed as an analogue of R^2 for the classical residual sum of squares, i.e., the check loss function for quantiles replaces the least-squares loss function and the τ empirical marginal quantile $Q_Y^{emp}(\tau)$ replaces the sample mean. With a maximum value of 1, the best model performance is reached at this maximum. Then,

$$\bar{R}^1 \equiv \int_0^1 R^1(\tau)\, d\tau = 1 - \int_0^1 \omega(\tau \mid \mathbf{y})\Delta(\tau)\, d\tau = 1 - \tilde{\Delta}, \tag{13}$$

provides a dimensionless global measure of model performance which can be used for model comparison.

In Section S2 of the SI we present the results of a brief simulation study where the goals were (i) to illustrate parameter recovery under fitting for several models, (ii) to investigate model flexibility, i.e., performance when the sampling model is not the same as the fitting model, and (iii) to consider the effect of sample size with regard to (i) and (ii).

3 The QAR(p) case

We provide a straightforward extension of our joint QAR(1) model to the lag p case. It is not a characterization of the QAR(p) function of Y_t but offers a flexible specification. In this regard, we obtain a form with some restrictions on the autoregressive coefficients but no constraints on the y_t's beyond the bounded interval support. By interpreting $\eta_1(\tau)$ and $\eta_2(\tau)$ in (5) as the conditional quantiles of Y_t at $y_{t-1} \in \{0,1\}$, we build a similar construction for an autoregressive process of order p as follows. Define

$$Q_{Y_t}(\tau \mid y_{t-1}, \ldots, y_{t-p})$$
$$= (\eta_1(\tau), \ldots, \eta_{p+1}(\tau)) \begin{pmatrix} 0 & \pi_1 & 0 & \cdots & 0 \\ 0 & 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \pi_p \\ 1 & -\pi_1 & -\pi_2 & \cdots & -\pi_p \end{pmatrix} \begin{pmatrix} 1 \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix}, \tag{14}$$

where the functions $\eta_1, \ldots, \eta_{p+1} : [0,1] \to [0,1]$ are monotonically increasing and the weights $\pi_1, \ldots, \pi_p$ are such that $\pi_j \geq 0$ and $\sum_j \pi_j = 1$. It is easy to see that such $Q_{Y_t}(\tau \mid y_{t-1}, \ldots, y_{t-p})$ is monotonically increasing in $\tau \in [0,1]$ for every $y_{t-1}, \ldots, y_{t-p} \in [0,1]$.

In particular, for QAR(2), let $\tau, \pi \in [0, 1]$. Then, define

$$Q_{Y_t}(\tau \mid y_{t-1}, y_{t-2}) = \pi y_{t-1} \eta_1(\tau) + (1-\pi) y_{t-2} \eta_2(\tau) + (1 - \pi y_{t-1} - (1-\pi) y_{t-2}) \eta_3(\tau) \tag{15}$$

where the three η functions are all strictly increasing, using forms as above. Rewriting the expression as

$$Q_{Y_t}(\tau \mid y_{t-1}, y_{t-2}) = \eta_3(\tau) + \pi(\eta_1(\tau) - \eta_3(\tau)) y_{t-1} + (1-\pi)(\eta_2(\tau) - \eta_3(\tau)) y_{t-2}, \tag{16}$$

both autoregressive coefficients belong to $[-1, 1]$ and need not be increasing in τ. We fit this QAR(2) model to our real data in Section 6.3. In fact, we only attempt this with $K = 1$ mixture components (seven parameters) to keep the model simple. Further, the second autoregressive term results are not influential for our data. Also, we choose $\log a$'s and $\log b$'s to follow a $N(0, 1.5^2)$ prior and $\pi \sim U(0, 1)$ as a non-informative prior for π.

4 Multivariate QAR(1)

Often a collection of dependent times series is gathered over a common time window. For instance, our illustration below considers the dependent pairs $\{(y_t^{\max}, y_t^{\min}) : t = 1, \ldots, T\}$, the daily maximum and minimum temperature for day t at a site. In fact, the collection of time series might be spatially referenced (leading to a spatial copula model construction, as developed in the next section). What we have is the quantile analogue of usual multivariate AR for time series. Implementation using the class of joint QAR(1) models we have proposed has not appeared in the literature. Our interest is in the quantile function for each time series. We are asking about the amount of dependence between quantile levels regarding the marginal quantile functions.

Here, we illustrate with the bivariate case where we have two models each defined as in (1), introducing dependence in the two time series by making the associated U_t's dependent through $T - 1$ i.i.d. 2-dimensional Gaussian copulas. This specification captures the acknowledged dependence between the pair of time series. We postpone to Section 5 the details of modeling using copulas; in particular, that section develops the form of the general n-dimensional joint density. The only detail that we advance here is that the correlation matrix associated with the copulas contains 1's on the diagonal and ρ on the off-diagonal, where $\rho \sim U(-1, 1)$ measures the correlation between series.

Apart from introducing dependence through $U_t^{\max}$ and $U_t^{\min}$, we could introduce dependence in the η's. For instance, using Kumaraswamy cdf's, under the $K = 1$ case, we consider the pairs $\log a_j^{\max}$ and $\log a_j^{\min}$ and the pairs $\log b_j^{\max}$ and $\log b_j^{\min}$ ($j = 1, 2$) to be bivariate normal. In our data we found little or no correlation between the parameters of the two time series, so in subsequent analyzes we will consider them independent. We do not pursue this case further here except to note the analogy with dependent responses in linear regression models. Introducing dependence through the U_t's is analogous to introducing dependence through the errors in the linear regression while introducing dependence through the η's is analogous to introducing dependence in the mean structure through shared parameters.

An example is presented in Section 6.4. Again, with $K = 1$, this yields four η's, i.e., four independent $\log a$'s and four independent $\log b$'s, each following a weak, say $N(0, 3^2)$ prior, as well as the copula parameter. As a by-product, we show the induced bivariate conditional pdf (arising from the bivariate analogue of Equation 8) for $(Y_t^{\max}, Y_t^{\min})$ with some choices for the y_{t-1}'s.

5 Spatial QAR(1)

In the spatial setting, we consider spatial point-referenced time series data. Here, $Y_t(\mathbf{s})$ denotes the observation for time $t = 1, \ldots, T$ at location $\mathbf{s} \in D$, where $D \subset \mathbb{R}^2$ is the study region. We have a time series at each of the locations, $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, say, the locations of the monitoring stations. The joint spatial QAR model is given by

$$Y_t(\mathbf{s}) = \theta_0(U_t(\mathbf{s}); \mathbf{s}) + \theta_1(U_t(\mathbf{s}); \mathbf{s}) Y_{t-1}(\mathbf{s}), \tag{17}$$

where the θ functions are quantile and spatially varying. Chen and Tokdar (2021) propose to model the spatial dependence of the realizations in a QR model using a spatial copula process. Generalizing it to our model, the vectors $(U_t(\mathbf{s}_1), \ldots, U_t(\mathbf{s}_n))^\top$ follow an independent copula distribution for every t. Supplementing Chen and Tokdar (2021), in (17) we introduce spatially varying coefficients rather than global coefficients. As a consequence, we have dependence in the time series realizations as well as spatially varying quantile functions.

5.1 Modeling spatial dependence

Spatial dependence is captured through spatially varying quantiles which are analogous to introducing spatially varying coefficients in spatial linear regression, and dependent quantile levels which are analogous to introducing dependence through the errors in the linear regression.

5.1.1 Spatially varying quantiles

For the spatially varying coefficients, we consider only one Kumaraswamy cdf for each $\eta(\tau; \mathbf{s})$. In fact, at location $\mathbf{s}$, let assume $\eta_j(\tau; \mathbf{s}) = 1 - (1 - \tau^{a_j(\mathbf{s})})^{b_j(\mathbf{s})}$ with parameters $a_j(\mathbf{s})$ and $b_j(\mathbf{s})$ for $j = 1, 2$. Then, we introduce four independent GP's for the a's and b's on the log-scale. In particular, we model $\log a_j(\mathbf{s}) \sim GP(a_j, \sigma_{a_j}^2 \rho(\mathbf{s}, \mathbf{s}'; \phi_{a_j}))$ and $\log b_j(\mathbf{s}) \sim GP(b_j, \sigma_{b_j}^2 \rho(\mathbf{s}, \mathbf{s}'; \phi_{b_j}))$ where the $\rho(\mathbf{s}, \mathbf{s}'; \phi)$'s are exponential correlation functions with ϕ's as corresponding decay parameters.

We take the ϕ's to be fixed values, according to the spatial scale, because it is usually difficult to estimate them from the data and typically interest focuses on the σ^2's, the spatial uncertainties (Banerjee et al., 2014). Specifically, we fix $\phi = 3/d_{\max}$, with $d_{\max}$ the maximum distance between any pair of spatial locations. Thus, the spatial GP's are only indexed by a mean and a variance parameter. We choose the priors $a_j, b_j, \log \sigma_{a_j}^2, \log \sigma_{b_j}^2 \sim N(0, 3^2)$ $(j = 1, 2)$.

5.1.2 The spatial copula process

A copula is a multivariate cdf for which the marginal distribution of each variable is $U(0,1)$. Copulas are used to model the dependence between random variables. Particularly, Sklar's theorem (Sklar, 1959) states that any multivariate joint pdf can be written in terms of univariate marginal pdf's and a copula which describes the dependence structure between the variables.

Gaussian spatial copulas enable computational advantages, e.g., ease of parameter estimation and scalability with sample size. For a given correlation matrix R, the n-dimensional Gaussian copula function with parameter matrix R becomes

$$C_\Phi(\mathbf{u} \mid R) = \Phi_R(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)), \tag{18}$$

where $\mathbf{u}^\top = (u_1, \ldots, u_n) \in [0,1]^n$, Φ_R is the joint cdf of a multivariate normal distribution with zero-mean vector and covariance matrix R. According to Xue-Kun Song (2000), the associated copula density is

$$c_\Phi(\mathbf{u} \mid R) = |R|^{-1/2} \exp\left\{ \frac{1}{2} \mathbf{q}^\top (\mathbf{I}_n - R^{-1}) \mathbf{q} \right\}, \tag{19}$$

with $\mathbf{q}^\top = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n))$.

With regard to the copula model for (17), we take the processes $U_t(\mathbf{s})$'s to follow a Gaussian copula for each t, induced by a spatial GP. In the spirit of Chen and Tokdar (2021), we define

$$
\begin{aligned}
U_t(\mathbf{s}) = \Phi(Z_t(\mathbf{s})), \quad Z_t(\mathbf{s}) = W_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \\
W_t(\mathbf{s}) \sim GP(0, \gamma\rho(\mathbf{s}, \mathbf{s}'; \phi)), \quad \epsilon_t(\mathbf{s}) \sim \text{i.i.d. } N(0, 1 - \gamma).
\end{aligned}
\tag{20}
$$

The process $W_t(\mathbf{s})$ captures spatial dependence while $\epsilon_t(\mathbf{s})$ is independent pure error. The parameter $\gamma \in [0,1]$ determines the proportion of spatial and independent variation. When $\gamma = 1$, the specification for $Z_t(\mathbf{s})$ is purely spatial. When $\gamma = 0$, we have an independent noise model. With this approach, the Gaussian copula density has correlation matrix $R \equiv \gamma R(\phi) + (1 - \gamma)\mathbf{I}_n$ where $R(\phi)$ is the $n \times n$ correlation matrix induced by $\rho(\mathbf{s}, \mathbf{s}'; \phi)$. To address the final copula piece of our model, we fix ϕ as above, and adopt $\gamma \sim U(0,1)$ as a non-informative prior for γ.

5.2 Likelihood evaluation

We are interested in the likelihood under model (17) using (19) and (20). It is convenient to first obtain the joint distribution for $\mathbf{Y}^\top = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_T^\top)$ where $\mathbf{Y}_t^\top = (Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_n))$, $t = 1, \ldots, T$. That is, each $\mathbf{Y}_t$ is $n \times 1$ and $\mathbf{Y}$ is $Tn \times 1$. By Sklar's theorem, the joint conditional density of responses, $\mathbf{Y}$, given the data at the

initial time, $\mathbf{y}_1$, can be partitioned into a marginal part and a copula part,

$$
f_{\mathbf{Y}}(\mathbf{y} \mid \mathbf{y}_1) = \prod_{t=2}^{T} \left[\prod_{i=1}^{n} f_{Y_t(\mathbf{s}_i)}\big(y_t(\mathbf{s}_i) \mid y_{t-1}(\mathbf{s}_i)\big) \right.
$$
$$
\left. \times\, c_{\Phi}\big(F_{Y_t(\mathbf{s}_1)}(y_t(\mathbf{s}_1) \mid y_{t-1}(\mathbf{s}_1)), \ldots, F_{Y_t(\mathbf{s}_n)}(y_t(\mathbf{s}_n) \mid y_{t-1}(\mathbf{s}_n)) \big) \right],
$$
$$
(21)
$$

where the cdf $F_{Y_t(\mathbf{s}_i)}$ corresponds to the pdf $f_{Y_t(\mathbf{s}_i)}$ and c_{Φ} is the Gaussian copula density in (19). As in Section 2.3, we evaluate $f_{Y_t(\mathbf{s}_i)}$ and $F_{Y_t(\mathbf{s}_i)}$ using:

$$
f_{Y_t(\mathbf{s}_i)}(y_t(\mathbf{s}_i) \mid y_{t-1}(\mathbf{s}_i)) = \left. \frac{1}{\frac{d}{d\tau} Q_{Y_t(\mathbf{s}_i)}(\tau \mid y_{t-1}(\mathbf{s}_i))} \right|_{\tau = \tau_{y_{t-1}(\mathbf{s}_i)}(y_t(\mathbf{s}_i))}, \qquad (22)
$$
$$
F_{Y_t(\mathbf{s}_i)}(y_t(\mathbf{s}_i) \mid y_{t-1}(\mathbf{s}_i)) = \tau_{y_{t-1}(\mathbf{s}_i)}(y_t(\mathbf{s}_i)),
$$

where $\tau_{y_{t-1}(\mathbf{s}_i)}(y_t(\mathbf{s}_i))$ solves $y_t(\mathbf{s}_i) = y_{t-1}(\mathbf{s}_i)\eta_1(\tau; \mathbf{s}_i) + (1 - y_{t-1}(\mathbf{s}_i))\eta_2(\tau; \mathbf{s}_i)$ in τ. Then, the log-likelihood score of the model parameters $\mathbf{\Omega}$ can be expressed by

$$
\ell(\mathbf{\Omega} \mid \mathbf{y}) = \sum_{t=2}^{T} \left[-\sum_{i=1}^{n} \log \left\{ y_{t-1}(\mathbf{s}_i)\dot{\eta}_1(u_t(\mathbf{s}_i); \mathbf{s}_i) + (1 - y_{t-1}(\mathbf{s}_i))\dot{\eta}_2(u_t(\mathbf{s}_i); \mathbf{s}_i) \right\} \right.
$$
$$
\left. + \log c_{\Phi}\left(u_t(\mathbf{s}_1), \ldots, u_t(\mathbf{s}_n) \mid R \right) \right], \qquad (23)
$$

with $u_t(\mathbf{s}_i) = \tau_{y_{t-1}(\mathbf{s}_i)}(y_t(\mathbf{s}_i))$. Finally, note that, for the calculation of the log-likelihood, the value of the $u_t(\mathbf{s}_i)$'s must be solved for, so the number of root-finders needed at each iteration of the MCMC is $n(T-1)$. As a result, likelihood evaluation is expensive, leading to long MCMC run times.

5.3 Spatial interpolation

The quantile $Q_{Y_t(\mathbf{s})}(\tau \mid y_{t-1}(\mathbf{s}))$ is a function of process realizations. Posterior samples for the hyperparameters are available from the model fitting. Posterior samples for the GP's are available, using posterior samples of the hyperparameters, through usual Bayesian kriging (Banerjee et al., 2014). This yields prediction of $a_j(\mathbf{s}_0)$ and $b_j(\mathbf{s}_0)$ ($j = 1, 2$) at a new $\mathbf{s}_0 \in D$, enabling spatially varying quantile functions. Therefore, we can interpolate conditional quantiles to any desired location in the study region given any proposed or reference value for the previous day's temperature at that location. If we do this over a sufficiently spatially resolved grid, we can obtain the posterior mean at each point and show the posterior τ conditional quantile surface for the given day.
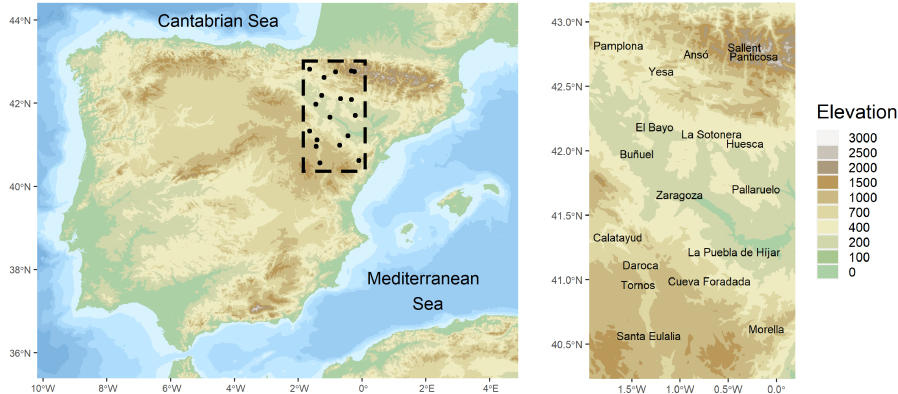
Fig. 1 Location of the 18 sites around Aragón, northeastern Spain.

6 Application to temperature data

6.1 The data

We illustrate the proposed modeling methods with analyses of persistence in point-referenced daily maximum temperatures (°C) from $n = 18$ locations in Aragón, Spain. We bring in daily minimum temperatures for the bivariate QAR(1) case. The data is provided by the State Meteorological Agency (AEMET, in its Spanish acronym). Castillo-Mateo et al. (2022) provide exploratory analysis and spatial hierarchical modeling for this dataset. We analyze responses for 2015, an interesting year because the summer was especially hot in Europe (Dong et al., 2016). There were numerous locations with record-breaking temperatures in July 2015 and the heat was maintained over time. The monthly average value of temperatures was a record in July 2015 for 6 of the 18 locations and in the entire region it was among the 10 hottest monthly averages. We restrict analysis to observations from May, June, July, August, and September (denoted as MJJAS), i.e., the hottest months of the year, resulting in $T = 153$ days. The location of the 18 observatories is shown in Figure 1 and their time series in Figure S4 of the SI.

We begin with a model comparison using QAR(1) and QAR(2) models for all locations. Then, we analyze two illustrative locations within the region, Pamplona and Zaragoza, separately. Subsequently, we implement the bivariate QAR(1) model to the daily maximum and minimum temperature series in Zaragoza. Finally, we implement the general model for spatial QAR(1) with all the locations. Before model fitting, we scale each of the temperature time series to $(0, 1)$ using the transformation in (3) with m and M in (4). We adopt site-level values for m and M.

6.2 The QAR(1) case

Table 1 shows, averaged across locations, the metrics of model adequacy $\tilde{p}_2$ and model comparison $\bar{R}^1$ defined in Section 2.4 for the QAR1K1 and QAR1K2 models, and the model from Koenker and Xiao (2006) fitted under our Bayesian framework using the

13

Table 1 Adequacy and comparison metrics $\tilde{p}_2$ in (10) and $\bar{R}^1$ in (13) averaged across locations for QAR1K1, QAR1K2, QAR2K1, and KX2006 models.

Model	Description	$\tilde{p}_2$	$\bar{R}^1$
QAR1K1	QAR(1) with $K = 1$ in (6)	0.633	0.365
QAR1K2	QAR(1) with $K = 2$ in (6)	0.402	0.365
QAR2K1	QAR(2) with $K = 1$ in (6)	0.542	0.365
KX2006	Koenker and Xiao (2006)	0.683	0.339

density in (8). Table S6 in the SI shows the metrics for each location. For this latter model, denoted as KX2006, we also consider a location parameter μ in the intercept, i.e., $\theta_0(\tau) = \mu + \sigma \Phi^{-1}(\tau)$ and $\theta_1(\tau) = \min\{\gamma_0 + \gamma_1 \tau, 1\}$ for $\gamma_0 \in (0, 1)$ and $\gamma_1 > 0$. With KX2006 we work on the original scale of the data since they are all positive. We choose the priors $\mu \sim N(0, 10^2)$, $\log \sigma, \log \gamma_1 \sim N(0, 3^2)$, and $\gamma_0 \sim U(0, 1)$. The $\bar{R}^1$ does not discriminate much between the proposed models, i.e., the autoregressive term explains much more variability than the difference in specification between the models. However, our proposed models have a slightly higher performance, 0.365, than the KX2006 model, around 0.34. Also, the $\tilde{p}_2$ directly measures how well the quantiles are captured, and its discriminative capacity is much higher. While QAR1K1 obtains a value of 0.633, adding a second component to the mixing improves this measure to 0.402. For its part, the KX2006 model obtains the worst value, 0.683, indicating an overall poorer fitting of the quantiles.

Figure 2 shows the posterior mean of the functions θ_0 and θ_1 in Pamplona and Zaragoza for the models QAR1K1 (dashed) and QAR1K2 (solid). Note that we could recover the intercepts on the original scale as $\theta_0^*(\tau) = m(1 - \eta_1(\tau)) + M\eta_2(\tau)$ and the autoregressive coefficients remain invariant. Further, θ_1 is not monotonic; this aspect of temperature dependence with respect to the previous day's temperature was also observed by Castillo-Mateo et al. (2023). It cannot be reproduced by KX2006. In Pamplona, the QAR1K2 model (the best) estimates a lower autoregressive coefficient than the QAR1K1 for $\tau \in (0.1, 0.7)$. In Zaragoza, similar curves appear for the two values of K, as shown by $\tilde{p}_2$ and $\bar{R}^1$.

Figure 3 shows the posterior mean of the conditional quantile functions $Q_{Y_t}(\tau \mid y)$ for three situations where y is the empirical τ marginal quantile for $\tau = 0.1, 0.5, 0.9$; the legend shows the values that are conditioned on both the original scale and the $(0, 1)$ scale. The smallest values of θ_1 are in extreme τ's, this means that the previous day's temperature is less influential for high quantiles. In fact, the conditional quantiles in Figure 3 overlap for τ's near 0 or near 1.

Figure 4 shows the posterior mean of the conditional density function in (8) under the same conditions as Figure 3. Pamplona presents different shapes in $f_{Y_t}(y_t \mid y)$ for different values of y. The distribution is asymmetrical with positive skewness if we condition on a small value for the previous day's temperature, and negative skewness if we condition on a big value. A general pattern is common in the region, the conditional distribution conditional on the 0.9 marginal quantile is more concentrated than those conditional on the 0.1 quantile. Figures S5, S6, S7, and S8 in the SI present the plots for the 18 locations.
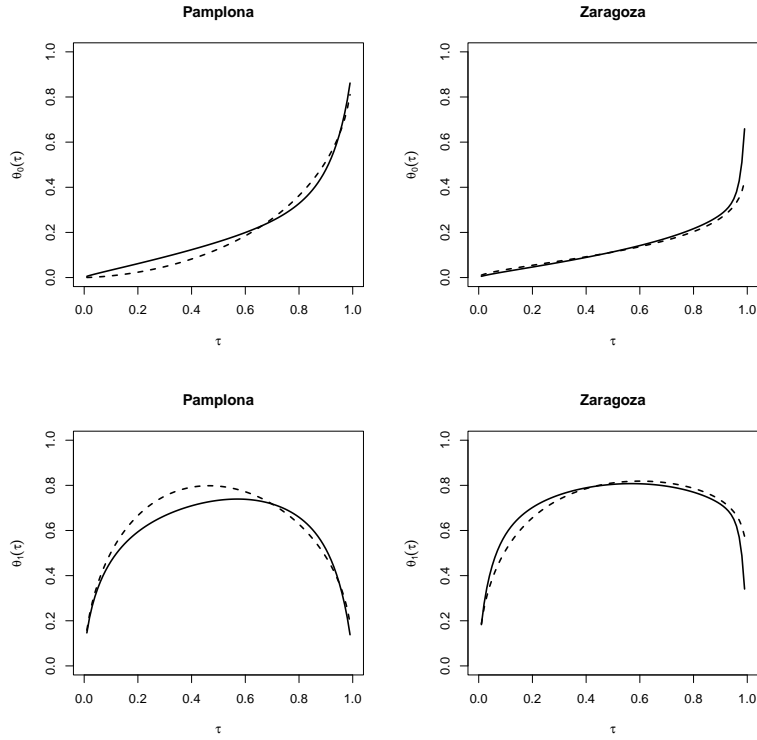
Fig. 2 Posterior mean of $\theta_0(\tau)$ (above) and $\theta_1(\tau)$ (below) vs. τ for QAR1K1 (dashed) and QAR1K2 (solid). Pamplona (left) and Zaragoza (right), MJJAS, 2015.
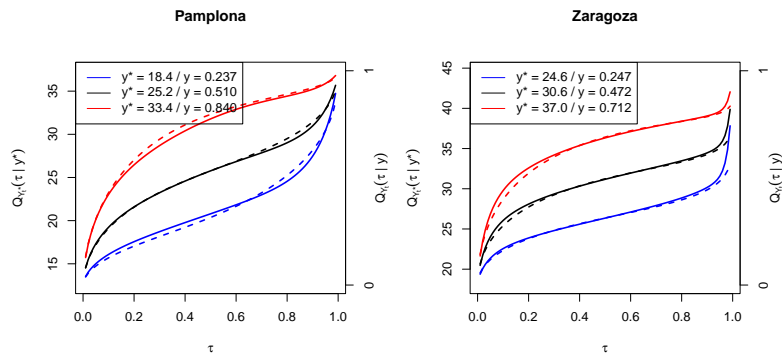

Fig. 3 Posterior mean of the quantile function $Q_{Y_t}(\tau \mid y)$ vs. τ for QAR1K1 (dashed) and QAR1K2 (solid). Here, y is the empirical τ marginal quantile for $\tau = 0.1$ (blue), 0.5 (black), 0.9 (red). Pamplona (left) and Zaragoza (right), MJJAS, 2015.
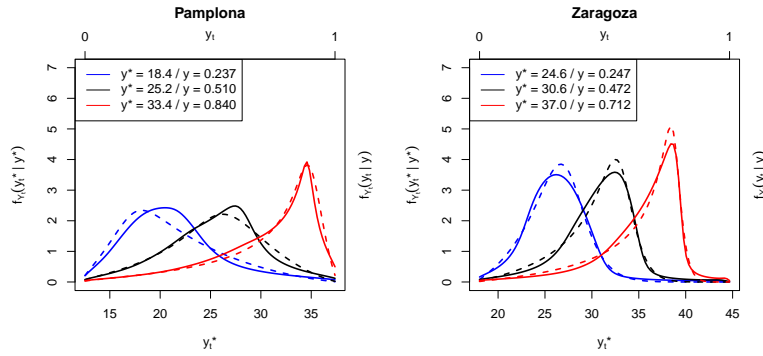
15

Fig. 4 Posterior mean of the density function $f_{Y_t}(x \mid y)$ for QAR1K1 (dashed) and QAR1K2 (solid). Here, y is the empirical τ marginal quantile for $\tau = 0.1$ (blue), 0.5 (black), 0.9 (red). Pamplona (left) and Zaragoza (right), MJJAS, 2015.

6.3 The QAR(2) case

Table 1 uses the criteria $\tilde{p}_2$ and $\bar{R}^1$ for the QAR(2) model with $K = 1$ (QAR2K1). The previous subsection showed that including a first lag improved the performance of the model with respect to an empirical null model. However, including a second lag does not increase the value of $\bar{R}^1$ with respect to a QAR(1) model. On the other hand, the measure of $\tilde{p}_2$ is somewhat better for QAR2K1 than for QAR1K1 but it is still inferior to the QAR1K2 case. Since QAR2K1 does not improve performance, and, as we will see below, there is no evidence that the term $\theta_2(\tau)$ is different from zero for any τ across most locations, there seems to be no value in exploring a QAR(2) model with $K = 2$.

Figure 5 shows the θ functions in the QAR2K1 model for Pamplona and Zaragoza (see Figure S9 in the SI for all locations). The θ_0 and θ_1 functions have a shape very similar to the QAR1K1 case. The θ_2 functions have values that are essentially centered at zero in most locations, giving more evidence that it is not necessary to introduce a lag of order 2 in the model. However, there are four locations with a coefficient slightly away from zero; Buñuel and La Sotonera have a value of $\theta_2(\tau)$ close to 0.2 for non-extreme τ's while Huesca and La Puebla de Híjar have similar behavior with values around 0.1.

6.4 Multivariate QAR(1)

Here, we fit the multivariate QAR(1) model (MQAR1K1) to the daily maximum and minimum temperature series at Zaragoza, $\{(y_t^{\max}, y_t^{\min}) : t = 1, \ldots, T\}$. The same analyses were developed for Pamplona and Daroca, but with different conclusions. Figure 6 shows the θ functions for the $y_t^{\max}$ (red) and $y_t^{\min}$ (blue) series. We see different patterns for $\theta_1^{\max}$ and $\theta_1^{\min}$; $y_t^{\max}$ shows high autocorrelation for high quantiles while $y_t^{\min}$ has less persistence for those quantiles.

For Zaragoza, the posterior mean of ρ is 0.32 with 95% credible interval $(0.17, 0.45)$, indicating the need to include dependence in the quantile levels of both series. For
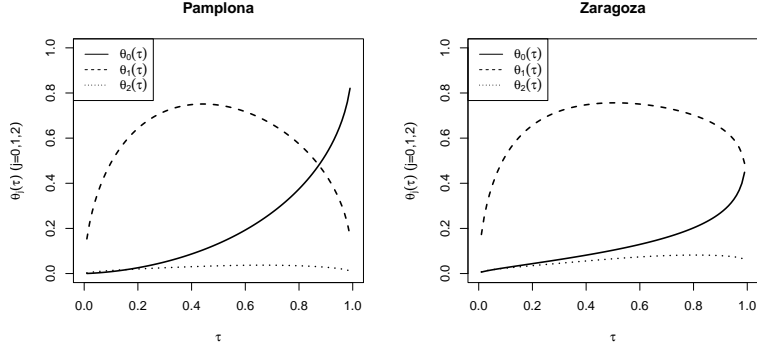
16

Fig. 5 Posterior mean of $\theta_0(\tau)$ (solid), $\theta_1(\tau)$ (dashed) and $\theta_2(\tau)$ (dotted) vs. τ for QAR2K1. Pamplona (left) and Zaragoza (right), MJJAS, 2015.
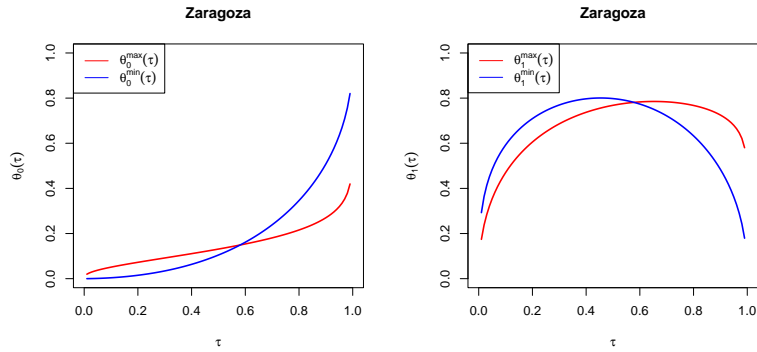

Fig. 6 Posterior mean of $\theta_0^{\max}(\tau)$ and $\theta_0^{\min}(\tau)$ (left), and $\theta_1^{\max}(\tau)$ and $\theta_1^{\min}(\tau)$ (right), vs. τ for MQAR1K1. Zaragoza, MJJAS, 2015.

Pamplona, the posterior mean of ρ is 0.06 with 95% credible interval $(-0.11, 0.23)$. Here, independent models for the conditional quantiles could be adopted. A reasonable explanation is the frequent appearance of fresh wind from the northwest during the night in Pamplona, resulting from proximity to the Cantabrian Sea.

Figure 7 shows level curves of the posterior conditional joint density of the vector $(Y_t^{\max}, Y_t^{\min})$ given the previous day's maximum and minimum temperatures, in Zaragoza (see Figures S10 and S11 in the SI for Pamplona and Daroca). The conditioning values are empirical marginal quantiles of $Y_t^{\max}$ and $Y_t^{\min}$. The first row conditions on the quantile $\tau = 0.5$ (30.6°C) and the second row on the quantile $\tau = 0.9$ (37.0°C) of $Y_t^{\max}$, and the same quantiles of $Y_t^{\min}$, for the first (17.2°C) and second (21.8°C) columns. The different patterns observed in the plots reveal a different relation between $Y_t^{\max}$ and $Y_t^{\min}$ depending on the previous day's temperatures. The conditional posterior distribution is not symmetric, with a different mean vector depending on the conditioning temperatures; the variability of the distribution is smaller when it is conditioned on high quantiles.
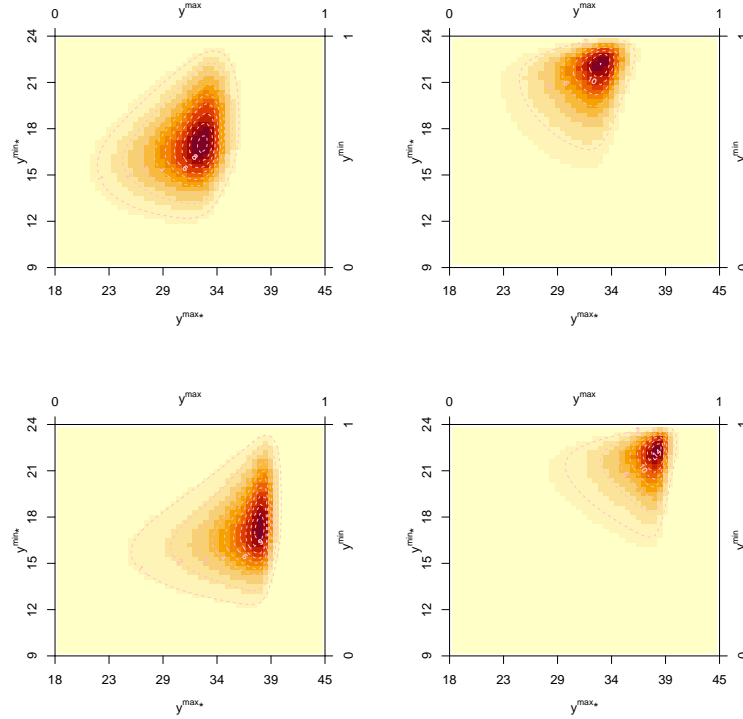
17

Fig. 7 Posterior mean of the density function of $(Y_t^{\max}, Y_t^{\min})$ conditioned on $(y^{\max}, y^{\min})$ for MQAR1K1. Here, $(y^{\max}, y^{\min})$ is equal to the respective empirical marginal quantiles for $\tau = 0.5$ (above), 0.9 (below) for the maximum; and $\tau = 0.5$ (left), 0.9 (right) for the minimum. Zaragoza, MJJAS, 2015.

6.5 Spatial QAR(1)

The spatial QAR model is fitted to the series of MJJAS in 18 locations in Aragón for the year 2015. The posterior mean of γ, the proportion of spatial dependence in (20), is 0.96 with 95% credible interval $(0.94, 0.98)$ indicating very strong spatial dependence in the quantile levels of the temperature series. Figure S12 in the SI provides maps of the posterior mean surface of the model GP's. We notice that $b_1(\mathbf{s})$ and $b_2(\mathbf{s})$ show approximately opposite spatial behavior since $b_1(\mathbf{s})$ has the highest values where $b_2(\mathbf{s})$ has the lowest, in the central and southeastern areas. Figure S13 of the SI shows boxplots of the posterior distribution of the GP's at each observed location; locations are sorted by elevation. The results suggest that the GP of $a_2(\mathbf{s})$ might be not necessary since the boxplots in the 18 locations have very similar ranges. The spatial variability of $a_1(\mathbf{s})$ is higher and, although it is not related to the elevation, it could be related to the distance to the coast.

The posterior distribution of $\theta_1(\tau; \mathbf{s})$, which captures the autoregressive structure, is summarized using the same type of plots. Figure S14 of the SI shows boxplots presenting the posterior distribution of $\theta_1(\tau; \mathbf{s})$ at the observed locations while Figure 8
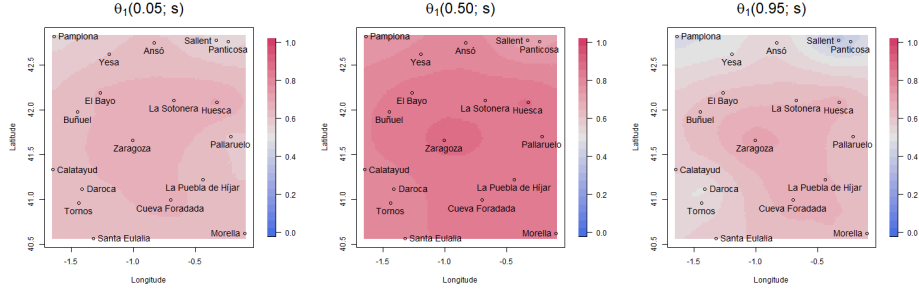
Fig. 8 Maps of the posterior mean of $\theta_1(\tau; \mathbf{s})$ for $\tau = 0.05, 0.50, 0.95$.

the maps of the posterior mean surface of $\theta_1(\tau; \mathbf{s})$, both for $\tau = 0.05, 0.50, 0.95$. The spatial GP's in the parameters of the Kumaraswamy distribution allow the model to fit different spatial patterns in each τ. The results show that the posterior mean of $\theta_1(\tau; \mathbf{s})$ is higher in the central quantiles. The spatial pattern of $\theta_1(\tau; \mathbf{s})$ is not symmetric around $\tau = 0.5$ and, e.g., values of $\theta_1(0.95; \mathbf{s})$ in the Pyrenees and northwestern areas are smaller than $\theta_1(0.05; \mathbf{s})$ in the same areas. Although $\theta_1(\tau; \mathbf{s})$ tends to be lower in locations with higher elevation, its spatial pattern cannot be explained by elevation alone. Consequently, the spatial GP's cannot be replaced with an elevation fixed effect.

The spatial joint model can also be used to estimate parameters related to the conditional distribution, e.g., conditional quantiles at unobserved locations. As a brief example, Figure S15 of the SI shows this through maps of the posterior mean of $Q_{Y_t(\mathbf{s})}(\tau \mid y)$ for $\tau = 0.05, 0.50, 0.95$, and $y = 0.05, 0.50, 0.95$. If it were desired to obtain the quantiles on the original scale of the data rather than the scale $(0, 1)$, we could consider a kriging of $m(\mathbf{s})$ and $M(\mathbf{s})$. With the same kriging procedure we could condition on values $y(\mathbf{s})$'s relative to a certain empirical marginal quantile for each location.

The spatial modeling here is primarily illustrative. For instance, the assumption of asymptotic tail independence, imposed by the Gaussian copula, may not be suitable. Examination of alternative copulas is beyond the scope of this work.

7 Summary and future work

We have presented consequentially expanded modeling for joint (non-quantile crossing) QAR. In particular, we have characterized the QAR(1) setting in a way that allows for a more flexible autocorrelation structure than the one in the seminal paper by Koenker and Xiao (2006). We have extended this to the QAR(p) case. We have offered a novel multiple time series version using a Gaussian copula. We have elaborated a spatial version, using a GP copula based upon a GP in conjunction with four additional GP's. This model enables spatially varying quantile functions. Our modeling is entirely parametric through the use of the Kumaraswamy distributions. A software implementation of our methods is available as the R-package "QAR" through GitHub: https://github.com/JorgeCastilloMateo/QAR.

We have illustrated the above contributions through time series of daily temperatures from sites in Aragón, Spain. The joint QAR model, with greater flexibility in

19

the modeling of the θ functions, allows us to capture autoregression structure in daily temperature data, which is not strictly increasing in τ, but decreasing in both tails.

A critical challenge in employing this work is model fitting. We can make specifications as rich as needed through the use of probabilistic mixtures of Kumaraswamy cdf's. However, it is well-known that model fitting employing MCMC with mixture specifications is often poorly identified. This issue is compounded in our case by the fact that calculation of the likelihood requires constant use of a one-dimensional root-finder. Ongoing work is attempting to address these computational difficulties.

It is important to note that we have not introduced any regressors into our modeling. This adds substantial complication to the joint approach. In order to consider coherent implementation of regressors, conditions have to be imposed on the support for the regressors, seeking to bridge our modeling with the work of Yang and Tokdar (2017). However, we briefly note a simple approximation strategy to incorporate regressors, e.g., seasonality, into our joint QAR approach. Suppose we introduce a regression structure, μ_t into the QAR(1) and estimate by $\hat{\mu}_t$, creating residuals $r_t = Y_t - \hat{\mu}_t$. Then, we could apply the above methodology to obtain the QAR(1) for r_t. Our strategy for selecting m and M can be applied to residuals. More precisely, let $r_t = \theta_{(0)}(U_t) + \theta_{(1)}(U_t)r_{t-1}$. This would yield the conditional quantile function, $Q_{r_t}(\tau \mid r_{t-1}) = \theta_{(0)}(\tau) + \theta_{(1)}(\tau)r_{t-1}$. Solving for the quantile function for Y_t we obtain

$$Q_{Y_t}(\tau \mid Y_{t-1}) = \hat{\mu}_t + \theta_{(0)}(\tau) + \theta_{(1)}(\tau)(Y_{t-1} - \hat{\mu}_{t-1}). \tag{24}$$

We acknowledge that this approximation can be criticized for two reasons: (i) we are creating $\hat{\mu}_t$ as if we were fitting a usual AR(1), and (ii) the resulting quantiles are not coherent since $\hat{\mu}_t$ is a function of $\{Y_t : t = 1, \ldots, T\}$. The QAR(1) is not defined until the end of the observation window.

Sections 4 and 5 could be combined to build a bivariate spatial QAR model for daily maximum and daily minimum temperature. Another challenge for the multivariate and spatial modeling would be to consider alternative copula choices, e.g., t-copulas in order to allow tail dependence for high quantiles.

Supplementary Information

SI for "Bayesian joint quantile autoregression" contains details on the Kumaraswamy distribution. Details on the simulation study. More results on the application with temperature series.

Acknowledgments

authors also thank Surya T. Tokdar from Duke University for fruitful discussions on joint QR modeling.

Declarations

Conflict of interest. The authors declare that they have no conflict of interest.

References

Banerjee, S., B.P. Carlin, and A.E. Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data* (2 ed.). New York, NY: Chapman and Hall/CRC.

Bondell, H.D., B.J. Reich, and H. Wang. 2010. Noncrossing quantile regression curve estimation. *Biometrika 97*(4): 825–838. https://doi.org/10.1093/biomet/asq048 .

Brent, R.P. 1973. *Algorithms for Minimization without Derivatives* (1 ed.). Englewood Cliffs, NJ: Prentice-Hall.

Castillo-Mateo, J., J. Asín, A.C. Cebrián, A.E. Gelfand, and J. Abaurrea. 2023. Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics 17*(3): 2305–2325. https://doi.org/10.1214/22-AOAS1719 .

Castillo-Mateo, J., M. Lafuente, J. Asín, A.C. Cebrián, A.E. Gelfand, and J. Abaurrea. 2022. Spatial modeling of day-within-year temperature time series: An examination of daily maximum temperatures in Aragón, Spain. *Journal of Agricultural, Biological and Environmental Statistics 27*(3): 487–505. https://doi.org/10.1007/s13253-022-00493-3 .

Chen, X., R. Koenker, and Z. Xiao. 2009. Copula-based nonlinear quantile autoregression. *The Econometrics Journal 12*(s1): S50–S67. https://doi.org/10.1111/j.1368-423X.2008.00274.x .

Chen, X. and S.T. Tokdar. 2021. Joint quantile regression for spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83*(4): 826–852. https://doi.org/10.1111/rssb.12467 .

Cui, W., C. Wan, and Y. Song. 2023. Ensemble deep learning-based non-crossing quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Transactions on Power Systems 38*: 3163–3178. https://doi.org/10.1109/TPWRS.2022.3202236 .

Das, P. and S. Ghosal. 2017. Bayesian quantile regression using random B-spline series prior. *Computational Statistics & Data Analysis 109*: 121–143. https://doi.org/10.1016/j.csda.2016.11.014 .

Dong, B., R. Sutton, L. Shaffrey, and L. Wilcox. 2016. The 2015 European heat wave. *Bulletin of the American Meteorological Society 97*(12): 57–62. https://doi.org/10.1175/BAMS-D-16-0140.1 .

Gao, M. and C.L.E. Franzke. 2017. Quantile regression-based spatiotemporal analysis of extreme temperature change in China. *Journal of Climate 30*(24): 9897–9914. https://doi.org/10.1175/JCLI-D-17-0356.1 .

Haario, H., E. Saksman, and J. Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli 7*(2): 223–242. https://doi.org/bj/1080222083 .

Jones, M.C. 2009. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology 6*(1): 70–81. https://doi.org/10.1016/j.stamet.2008.04.001 .

Koenker, R. and G. Bassett. 1978. Regression quantiles. *Econometrica 46*(1): 33–50. https://doi.org/10.2307/1913643 .

Koenker, R. and J.A.F. Machado. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association 94*(448): 1296–1310. https://doi.org/10.1080/01621459.1999.10473882 .

Koenker, R. and Z. Xiao. 2006. Quantile autoregression. *Journal of the American Statistical Association 101*(475): 980–990. https://doi.org/10.1198/016214506000000672 .

Kozumi, H. and G. Kobayashi. 2011. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation 81*(11): 1565–1578. https://doi.org/10.1080/00949655.2010.496117 .

Kumaraswamy, P. 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology 46*(1): 79–88. https://doi.org/10.1016/0022-1694(80)90036-0 .

Li, G., Y. Li, and C.L. Tsai. 2015. Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association 110*(509): 246–261. https://doi.org/10.1080/01621459.2014.892007 .

Lum, K. and A.E. Gelfand. 2012. Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis 7*(2): 235–258. https://doi.org/10.1214/12-BA708 .

Peng, B., K. Yang, and X. Dong. 2023. Variable selection for quantile autoregressive model: Bayesian methods versus classical methods. *Journal of Applied Statistics 0*(0): 1–33. https://doi.org/10.1080/02664763.2023.2178642 .

Reich, B.J., M. Fuentes, and D.B. Dunson. 2011. Bayesian spatial quantile regression. *Journal of the American Statistical Association 106*(493): 6–20. https://doi.org/10.1198/jasa.2010.ap09237 .

Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris* 8: 229–231 .

Thrasher, B., E.P. Maurer, C. McKellar, and P.B. Duffy. 2012. Bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrology and Earth System Sciences 16*(9): 3309–3314. https://doi.org/10.5194/hess-16-3309-2012 .

Tokdar, S.T. and J.B. Kadane. 2012. Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis 7*(1): 51–72. https://doi.org/10.1214/12-BA702 .

Xue-Kun Song, P. 2000. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics 27*(2): 305–320. https://doi.org/10.1111/1467-9469.00191 .

Yang, C., L. Li, and J. Xu. 2018. Changing temperature extremes based on CMIP5 output via semi-parametric quantile regression approach. *International Journal of Climatology 38*(9): 3736–3748. https://doi.org/10.1002/joc.5524 .

Yang, K., B. Peng, and X. Dong. 2023. Bayesian inference for quantile autoregressive model with explanatory variables. *Communications in Statistics-Theory and Methods 52*(9): 2946–2965. https://doi.org/10.1080/03610926.2021.1964529 .

Yang, Y. and S.T. Tokdar. 2017. Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association 112*(519): 1107–1120. https://doi.org/10.1080/01621459.2016.1192545 .

Yu, K. and R.A. Moyeed. 2001. Bayesian quantile regression. *Statistics & Probability Letters 54*(4): 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9 .

Zhang, C., C. Ji, L. Hua, H. Ma, M.S. Nazir, and T. Peng. 2022. Evolutionary quantile regression gated recurrent unit network based on variational mode decomposition, improved whale optimization algorithm for probabilistic short-term wind speed prediction. *Renewable Energy* 197: 668–682. https://doi.org/10.1016/j.renene.2022.07.123 .

Supplementary Information for "Bayesian joint quantile autoregression"

Jorge Castillo-Mateo, Alan E. Gelfand, Jesús Asín,
Ana C. Cebrián, and Jesús Abaurrea

S1 Kumaraswamy distribution

Figures S1 and S2 show the pdf and the cdf of the Kumaraswamy distribution for different combinations of parameters (a, b).
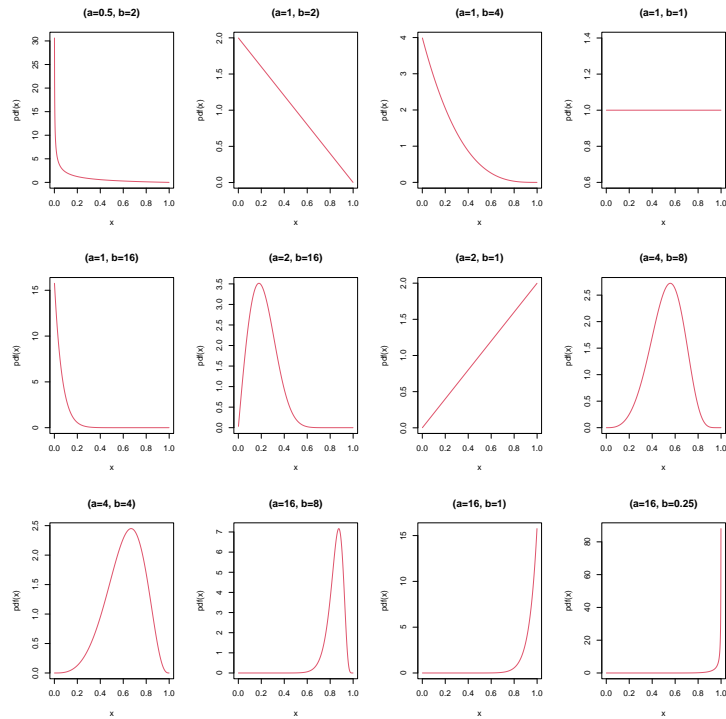

Figure S1: Probability density function of the Kumaraswamy distribution for different parameters.
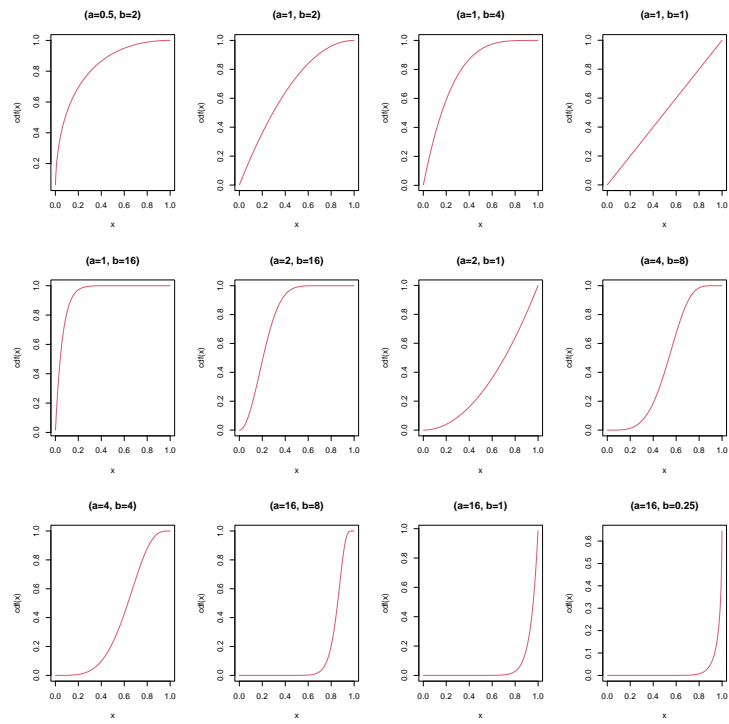
Figure S2: Cumulative distribution function of the Kumaraswamy distribution for different parameters.

S2   Simulation study

To illustrate parameter recovery under the modeling in Section 2 of the Main Manuscript (MM), we perform a simulation study where we consider QAR(1) models with $K = 1$ and $K = 2$ (QAR1K1 and QAR1K2) Kumaraswamy cdf's in the probabilistic mixture from Section 2.2.1, and scenarios with different combinations of parameter values. For each scenario we simulate $B = 100$ datasets of length $T = 150$ and $T = 500$ time points, using the model in (5) from the MM with the η's again specified as probabilistic mixtures. For each simulated dataset, a QAR(1) model with K equal to the value used in the data generating process is fitted. First, 90% credible intervals are computed for $\theta_0(\tau)$ and $\theta_1(\tau)$ in a grid of values of τ. Then, the coverage (CVG) of each value of each function is computed as the proportion of the B computed credible intervals that contain the corresponding true values.

Table S1 shows $\overline{CVG}(\theta_0)$ and $\overline{CVG}(\theta_1)$, the average of the CVG's for the values of the functions in $\tau \in \{0.01, 0.02, \ldots, 0.99\}$ based on the QAR1K1 model fitted to data generated with the same model, under four scenarios. The individual CVG's of $\theta_0(\tau)$ and $\theta_1(\tau)$ for $\tau = 0.01, 0.50, 0.99$ are shown in Table S2. Scenario SC1 considers independent series of data with $a_1 = 0.5$, $b_1 = 2$, $a_2 = 0.5$ and $b_2 = 2$ (here, the subscript indicates the function η_j to which the parameter corresponds). The other scenarios correspond to series with different correlation structures. Parameters in Scenario SC2 ($a_1 = 4$, $b_1 = 4$, $a_2 = 1$, $b_2 = 2$) give values of $\theta_1(\tau)$ moving from -0.5 to 0, with the maximum negative correlation for the central values of τ. Parameters in Scenario SC3 ($a_1 = 0.5$, $b_1 = 2$, $a_2 = 2$, $b_2 = 1$) give $\theta_1(\tau)$ values from 0.2 to 0.7 for $\tau < 0.75$. Parameters in Scenario SC4 ($a_1 = 0.3$, $b_1 = 6$, $a_2 = 12$, $b_2 = 8$) give values of $\theta_1(\tau)$ from 0.8 to 1 for $\tau < 0.75$. The θ functions corresponding to the generated data are plotted in Figure S3. The same study is repeated for series of data generated with a QAR1K2 model under three scenarios: SC5 gives, again, uncorrelated series; SC6 positive $\theta_1(\tau)$ between 0.5 and 0.91 for $\tau < 0.8$; and SC7 gives $\theta_1(\tau)$ moving from -0.5 to 0, with the maximum negative values for the central values of τ and decreasing slowly towards 0 in the extremes. Table S3 contains the parameters used in each scenario. The average of the CVG's, $\overline{CVG}(\theta_0)$ and $\overline{CVG}(\theta_1)$, when fitting the QAR1K2 model under these scenarios are also shown in Table S1, and plots of $\theta_0(\tau)$ and $\theta_1(\tau)$ are shown in Figure S3.

The results are satisfactory, with the average of the CVG's close to the nominal level of 0.90, varying between 0.86 and 0.92 in all scenarios for both $T = 150$ and 500. Not only the averages but also the individual CVG's for each τ, even for extreme quantiles ($\tau = 0.01, 0.99$), are close to 0.90 (see Tables S2 and S4). It is noteworthy that Scenarios SC3, SC4 and SC6 could correspond to a usual correlation structure in climate and environmental data, with a strong positive dependence in the central values of τ that weakens at the extremes.

To study the flexibility of the QAR1K1 model, Table S5 summarizes the metrics described in Section 2.4 of the MM, obtained from fitting the QAR1K1 and QAR1K2 models to the previous data series generated with QAR1K2 models. The values of $\bar{R}^1$ and $\tilde{p}_2$ are the metrics averaged across the $B = 100$ simulations. QAR1K1 models are quite flexible and their metrics are only slightly poorer than QAR1K2 models. Note that Scenario SC5 imposes independence and $\bar{R}^1$ is expected to be zero. We emphasize that comparison with Koenker and Xiao (2006) do not make sense under scenarios with negative trends and non-monotonic coefficients. Oppositely, simulating data under the Koenker and Xiao model, QAR1K1 obtains results not too far from QAR1K2 that becomes competitive against the truth.

Function	T	SC1	SC2	SC3	SC4	SC5	SC6	SC7
θ_0	150	0.87	0.86	0.91	0.87	0.92	0.88	0.88
	500	0.89	0.89	0.88	0.89	0.91	0.89	0.91
θ_1	150	0.88	0.90	0.88	0.86	0.90	0.87	0.87
	500	0.89	0.89	0.90	0.87	0.91	0.90	0.90

Table S1: The 90% $\overline{CVG}$ of $\theta_0(\tau)$ and $\theta_1(\tau)$ across $\tau \in \{0.01, 0.02, \ldots, 0.99\}$ obtained from fitting QAR1K1 models in Scenarios SC1–SC4 and QAR1K2 models in Scenarios SC5–SC7.

Scenario	T	$\theta_0(0.01)$	$\theta_0(0.99)$	$\theta_0(0.50)$	$\theta_1(0.01)$	$\theta_1(0.99)$	$\theta_1(0.50)$
SC1	150	0.88	0.86	0.85	0.87	0.91	0.86
	500	0.91	0.89	0.88	0.89	0.90	0.90
SC2	150	0.87	0.90	0.92	0.87	0.89	0.90
	500	0.87	0.89	0.88	0.87	0.88	0.91
SC3	150	0.91	0.83	0.89	0.90	0.83	0.89
	500	0.90	0.87	0.88	0.83	0.89	0.95
SC4	150	0.88	0.88	0.86	0.88	0.88	0.86
	500	0.86	0.86	0.80	0.93	0.86	0.81

Table S2: The 90% CVG of $\theta_0(\tau)$ and $\theta_1(\tau)$ ($\tau = 0.01, 0.50, 0.99$) in QAR1K1 models fitted to Scenarios SC1–SC4.

Scenario	$a_{1,1}$	$b_{1,1}$	$a_{1,2}$	$b_{1,2}$	$a_{2,1}$	$b_{2,1}$	$a_{2,2}$	$b_{2,2}$	λ_1	λ_2
SC5	0.5	2.0	4.0	8.0	0.5	2.0	4.0	8.0	0.3	0.3
SC6	0.5	2.0	0.3	6.0	1.0	1.0	12.0	8.0	0.4	0.1
SC7	3.0	0.5	2.0	1.0	1.0	2.0	0.5	1.0	0.2	0.4

Table S3: Values of the parameters of the QAR1K2 models used to generate data under Scenarios SC5–SC7. (First subscript indicates the mixture function η_j and second subscript indicates the k-th component of that mixture to which the parameter corresponds.)

Scenario	T	$\theta_0(0.01)$	$\theta_0(0.99)$	$\theta_0(0.50)$	$\theta_1(0.01)$	$\theta_1(0.99)$	$\theta_1(0.50)$
SC5	150	0.97	0.91	0.97	0.86	0.91	0.97
	500	0.90	0.80	0.93	0.86	0.80	0.93
SC6	150	0.95	0.85	0.90	0.87	0.83	0.91
	500	0.90	0.93	0.89	0.80	0.93	0.89
SC7	150	0.95	0.85	0.90	0.87	0.83	0.91
	500	0.94	0.88	0.89	0.94	0.86	0.89

Table S4: The 90% CVG of $\theta_0(\tau)$ and $\theta_1(\tau)$ ($\tau = 0.01, 0.50, 0.99$) in QAR1K2 models fitted to Scenarios SC5–SC7.
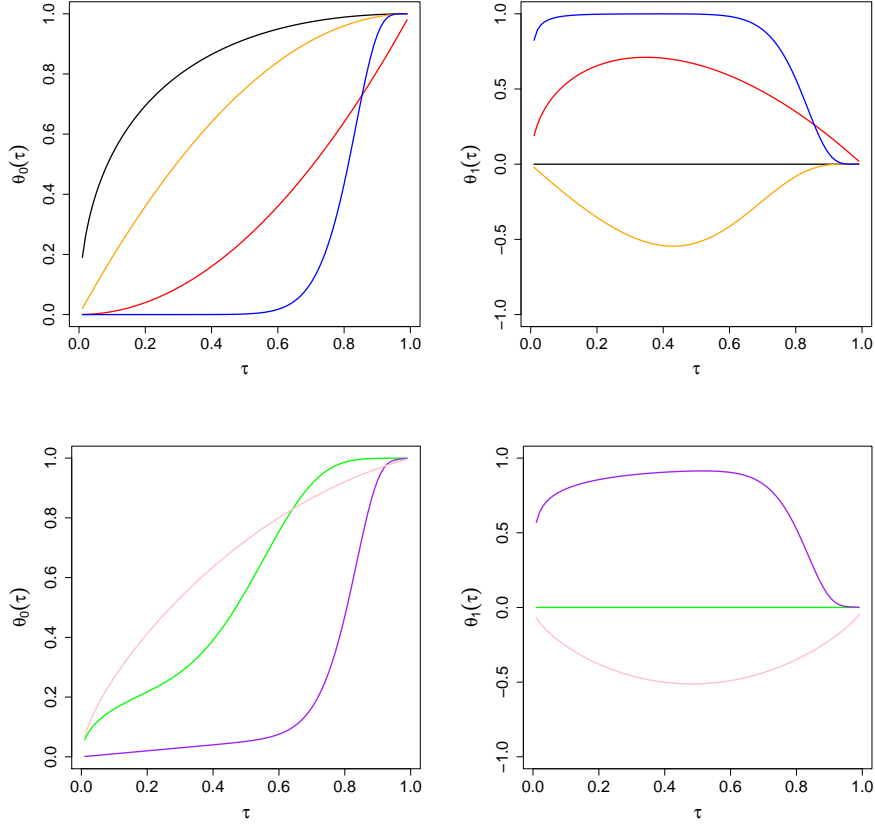
4

Figure S3: Functions $\theta_0(\tau)$ (left) and $\theta_1(\tau)$ (right) vs. τ. Under Scenarios SC1, (black), SC2 (orange), SC3 (red), and SC4 (blue) used to generate QAR1K1 series (above); and under Scenarios SC5 (green), SC6 (purple), and SC7 (pink) used to generate QAR1K2 series (below).

Scenario	T	QAR1K1		QAR1K2	
		$\bar{R}^1$	$\tilde{p}_2$	$\bar{R}^1$	$\tilde{p}_2$
SC5	150	-0.0387	0.0019	0.0021	0.0002
	500	-0.0439	0.0019	0.0002	0.0001
SC6	150	0.3082	0.0003	0.3091	0.0001
	500	0.3155	0.0002	0.3171	0.0000
SC7	150	0.0781	0.0003	0.0776	0.0001
	500	0.0737	0.0001	0.0737	0.0001

Table S5: Adequacy and comparison metrics for QAR1K1 and QAR1K2 models in Scenarios SC5–SC7. Values are averaged across $B = 100$ simulations.

5

S3   Application to temperature data
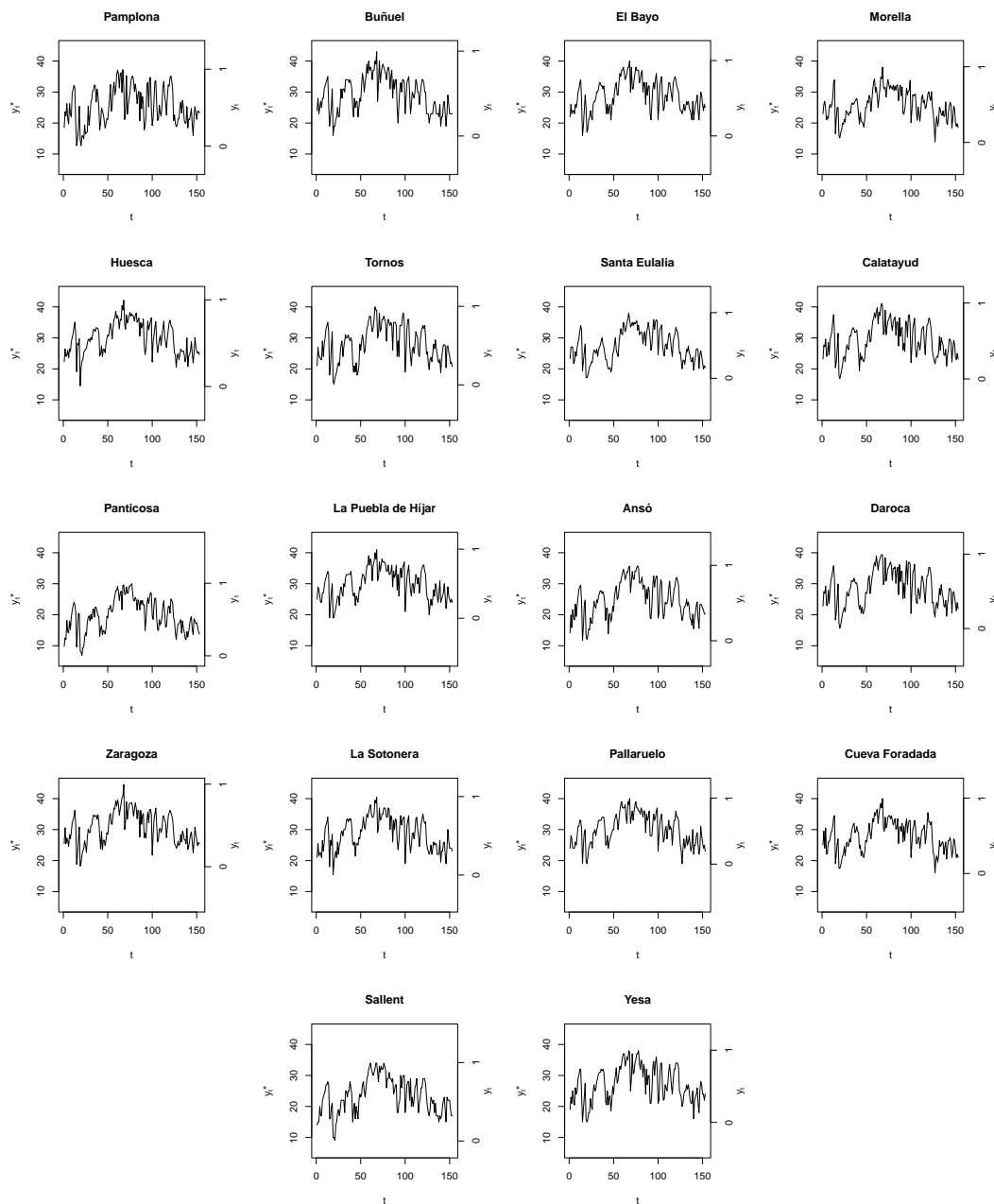
S3.1   Time series data

Figure S4: Daily maximum temperature time series of the 18 locations in MJJAS, 2015.

S3.2   Model comparison

Location	QAR1K1		QAR1K2		QAR2K1		KX2006	
	$\tilde{p}_2$	$\bar{R}^1$	$\tilde{p}_2$	$\bar{R}^1$	$\tilde{p}_2$	$\bar{R}^1$	$\tilde{p}_2$	$\bar{R}^1$
Pamplona	0.75	0.267	0.46	0.270	0.65	0.265	0.65	0.274
Buñuel	0.56	0.358	0.35	0.356	0.52	0.357	0.82	0.337
El Bayo	0.67	0.362	0.39	0.364	0.61	0.361	0.44	0.331
Morella	0.59	0.352	0.35	0.351	0.45	0.351	0.73	0.318
Huesca	0.69	0.381	0.50	0.382	0.39	0.390	0.83	0.342
Tornos	0.61	0.352	0.39	0.352	0.55	0.352	0.54	0.306
Santa Eulalia	0.60	0.454	0.31	0.455	0.52	0.453	0.66	0.445
Calatayud	0.57	0.335	0.44	0.335	0.50	0.336	1.01	0.328
Panticosa	0.60	0.447	0.36	0.448	0.51	0.443	0.40	0.431
La Puebla de Híjar	0.80	0.333	0.37	0.336	0.64	0.339	0.69	0.301
Ansó	0.42	0.411	0.34	0.410	0.37	0.410	0.51	0.384
Daroca	0.58	0.334	0.44	0.335	0.56	0.337	0.37	0.297
Zaragoza	0.62	0.351	0.45	0.351	0.50	0.356	0.74	0.304
La Sotonera	0.67	0.341	0.40	0.344	0.58	0.345	0.51	0.300
Pallaruelo	0.80	0.362	0.50	0.362	0.63	0.361	0.80	0.335
Cueva Foradada	0.46	0.359	0.47	0.353	0.40	0.365	0.76	0.334
Sallent	0.71	0.418	0.30	0.419	0.73	0.413	0.53	0.401
Yesa	0.72	0.343	0.42	0.344	0.64	0.340	1.31	0.344
$\sum/18$	**0.633**	**0.365**	**0.402**	**0.365**	**0.542**	**0.365**	**0.683**	**0.339**

Table S6: Adequacy and comparison metrics for QAR1K1, QAR1K2, QAR2K1, and KX2006 models for the 18 locations and averaged across locations.
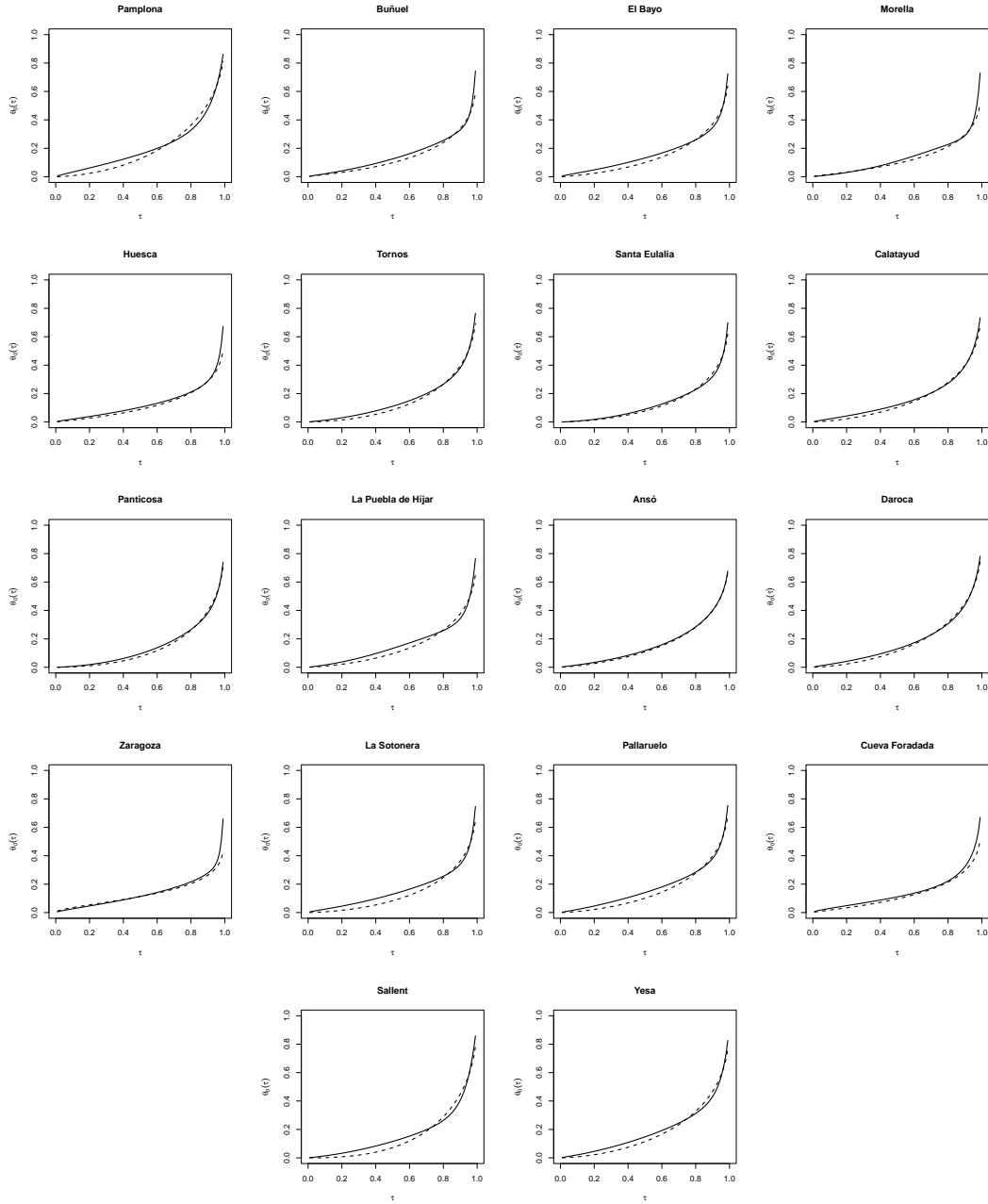
S3.3 The QAR(1) case

Figure S5: Posterior mean of $\theta_0(\tau)$ vs. τ for QAR1K1 (dashed) and QAR1K2 (solid). All locations, MJJAS, 2015.
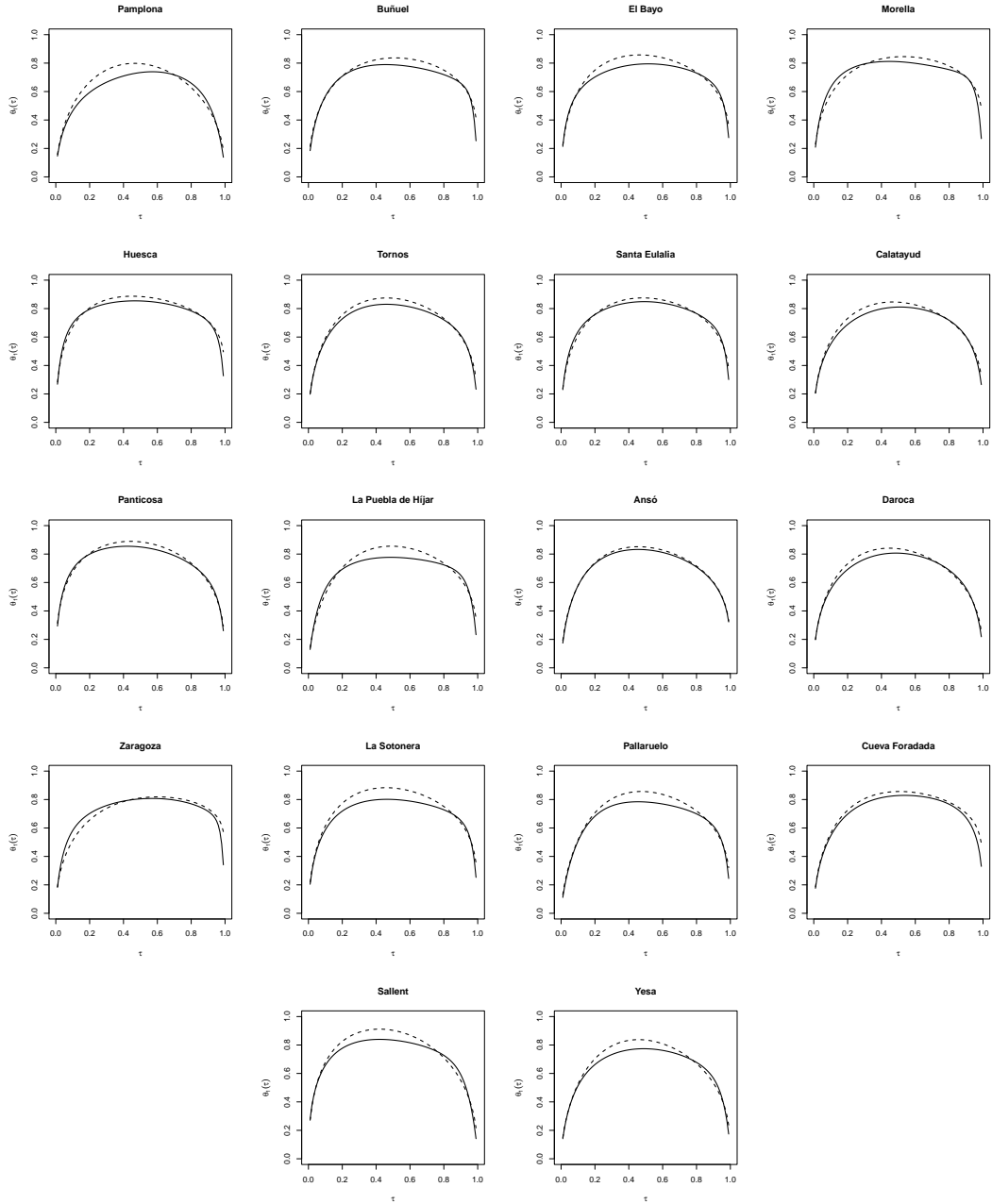
Figure S6: Posterior mean of $\theta_1(\tau)$ vs. τ for QAR1K1 (dashed) and QAR1K2 (solid). All locations, MJJAS, 2015.
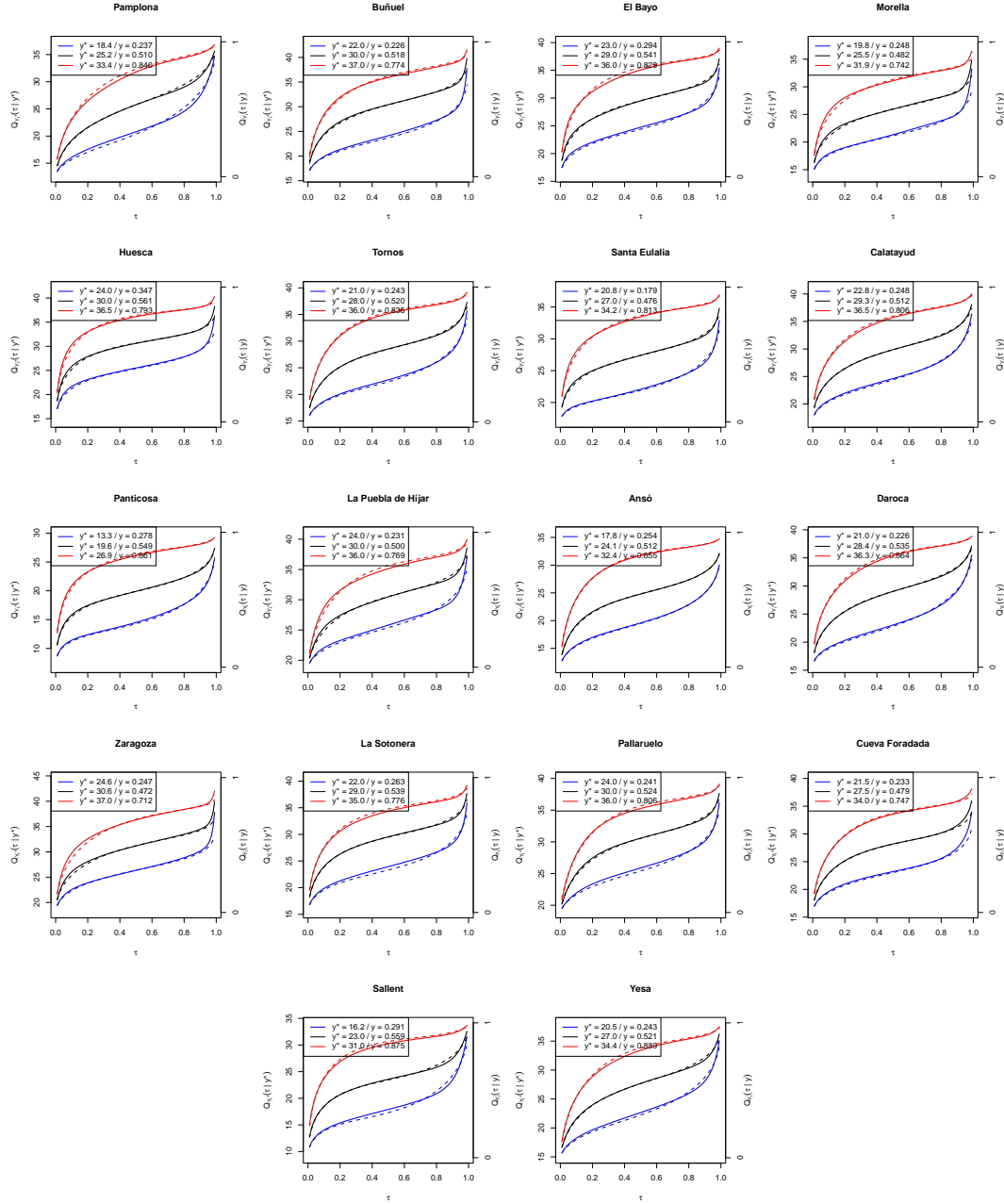
9

Figure S7: Posterior mean of the quantile function $Q_{Y_t}(\tau \mid y)$ vs. τ for QAR1K1 (dashed) and QAR1K2 (solid). Here, y is the empirical marginal quantile for $\tau = 0.1$ (blue), 0.5 (black), 0.9 (red). All locations, MJJAS, 2015.

10

Figure S8: Posterior mean of the density function $f_{Y_t}(y_t \mid y)$ for QAR1K1 (dashed) and QAR1K2 (solid). Here, y is the empirical marginal quantile for $\tau = 0.1$ (blue), 0.5 (black), 0.9 (red). All locations, MJJAS, 2015.
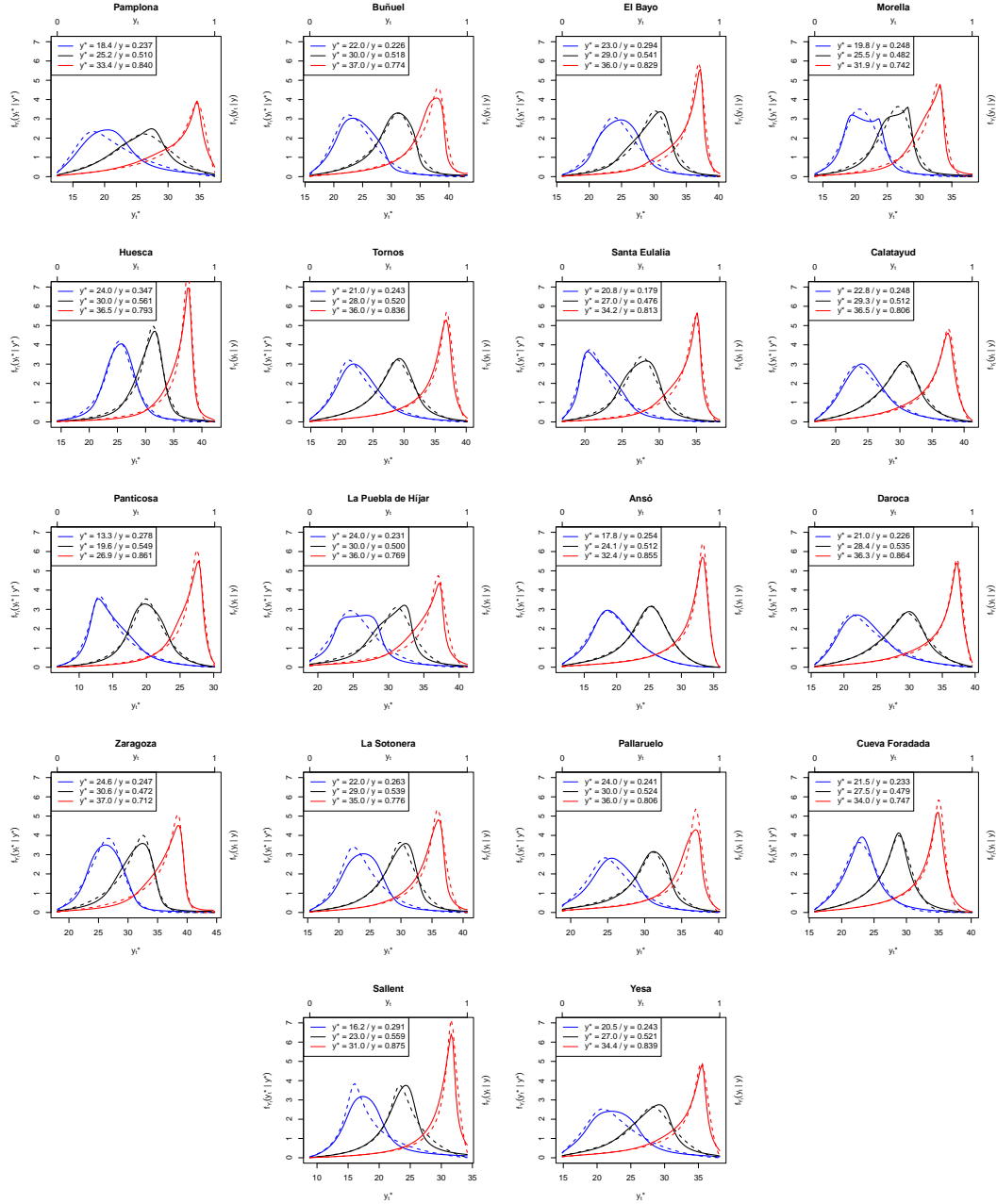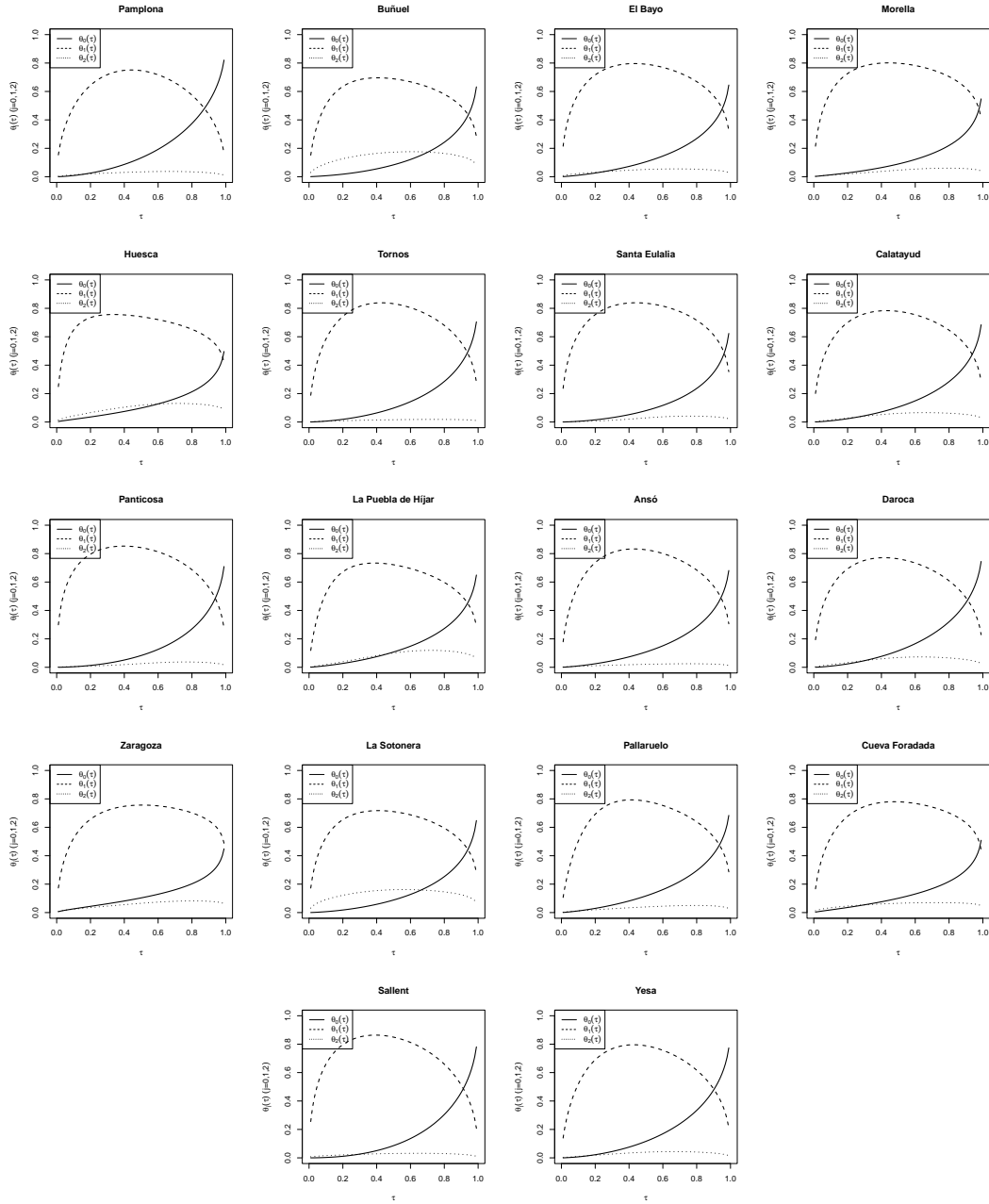
S3.4 The QAR(2) Case

Figure S9: Posterior mean of $\theta_0(\tau)$ (solid), $\theta_1(\tau)$ (dashed) and $\theta_2(\tau)$ (dotted) vs. τ for QAR2K1. All locations, MJJAS, 2015

S3.5 Multivariate QAR(1)

Figure S10: Posterior mean of the density function of $(Y_t^{\max}, Y_t^{\min})$ conditioned on $(y^{\max}, y^{\min})$ for MQAR1K1. Here, $(y^{\max}, y^{\min})$ is equal to the respective empirical marginal quantiles for $\tau = 0.5$ (above), 0.9 (below) for the maximum; and $\tau = 0.5$ (left), 0.9 (right) for the minimum. Pamplona, MJJAS, 2015.
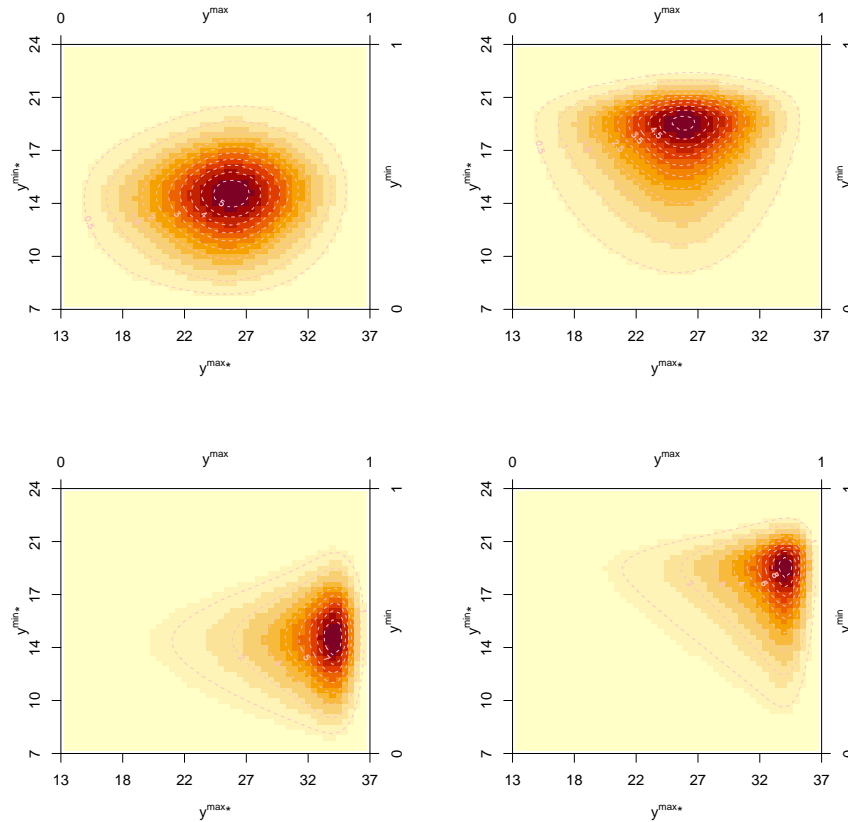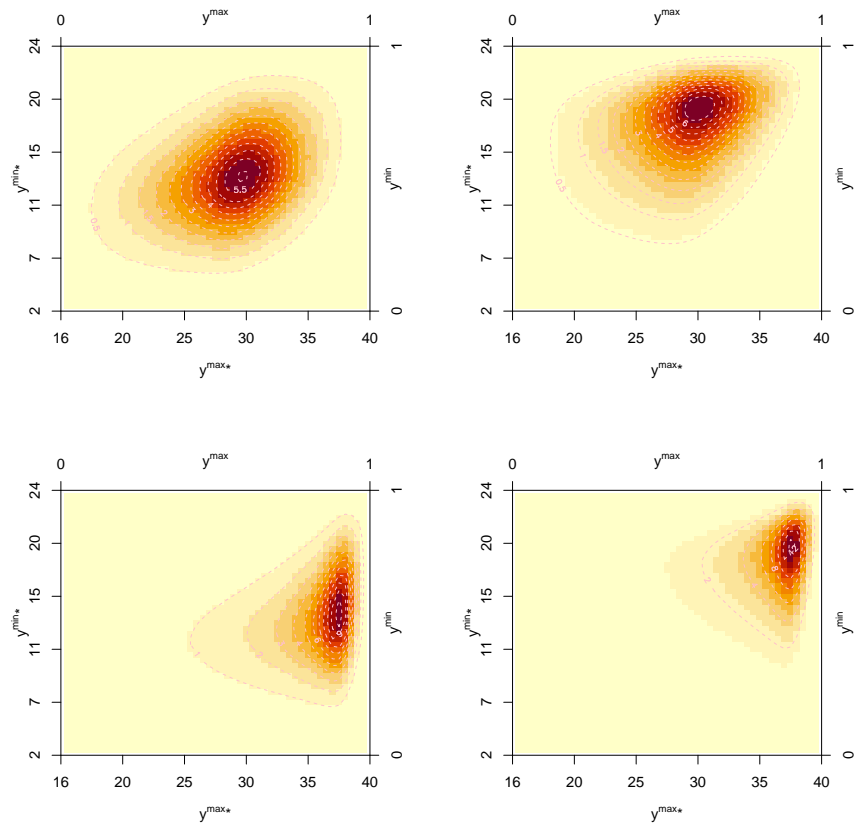
Figure S11: Posterior mean of the density function of $(Y_t^{\max}, Y_t^{\min})$ conditioned on $(y^{\max}, y^{\min})$ for MQAR1K1. Here, $(y^{\max}, y^{\min})$ is equal to the respective empirical marginal quantiles for $\tau = 0.5$ (above), 0.9 (below) for the maximum; and $\tau = 0.5$ (left), 0.9 (right) for the minimum. Daroca, MJJAS, 2015.

S3.6 Spatial QAR(1)
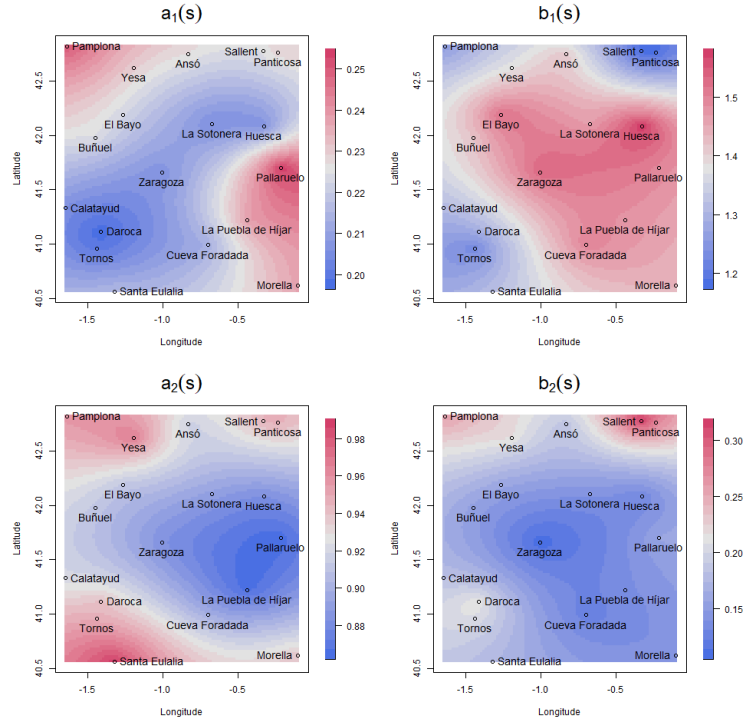
Figure S12: Maps of the posterior mean of $a_1(\mathbf{s})$, $b_1(\mathbf{s})$, $a_2(\mathbf{s})$, and $b_2(\mathbf{s})$.
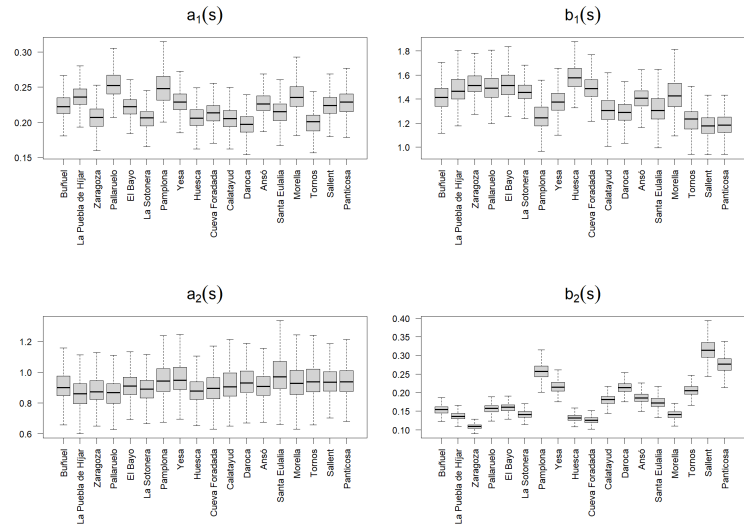
Figure S13: Boxplots summarizing the posterior distribution of $a_1(\mathbf{s})$, $b_1(\mathbf{s})$, $a_2(\mathbf{s})$, and $b_2(\mathbf{s})$ at the observed locations sorted by elevation.
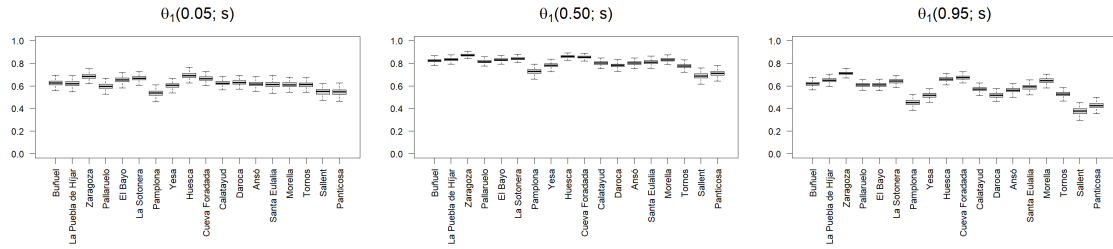
Figure S14: Boxplots summarizing the posterior distribution of $\theta_1(\tau; \mathbf{s})$ for $\tau = 0.05, 0.50, 0.95$ at the observed locations sorted by elevation.
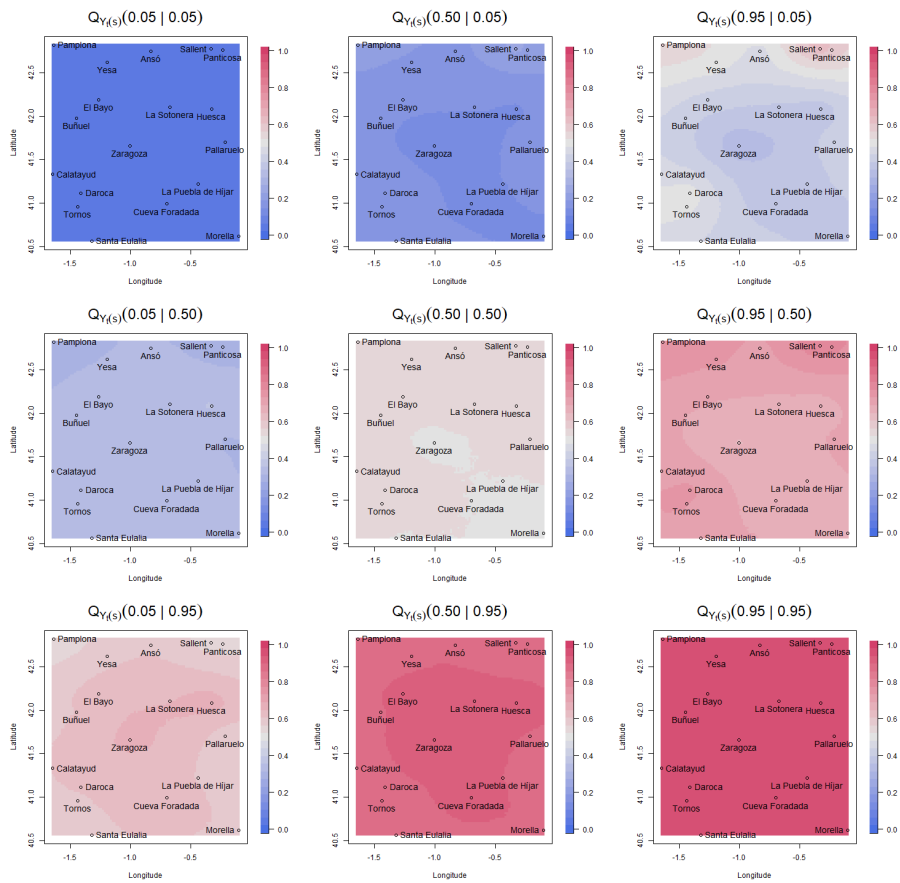
Figure S15: Maps of the posterior mean of $Q_{Y_t(\mathbf{s})}(\tau \mid y)$ for $\tau = 0.05, 0.50, 0.95$ and $y = 0.05, 0.50, 0.95$.

16

3.5 Record tests to detect non stationarity in the tails with an application to climate change

This manuscript was published in:

And it was disseminated (speaker emphasized) in:

- *Cebrián, A. C.*, **Castillo-Mateo, J.**, & Asín, J. (2022, December 17–19). *Nonparametric tests based on records to detect trends in the upper tail of climate series* [Invited talk]. CFE-CMStatistics 2022, London, UK.

- ***Castillo-Mateo, J.***, Cebrián, A. C., & Asín, J. (2022, September 14–16). *Records tests and applications to climate change* [Contributed talk]. VI Jornadas Científicas de Estudiantes de la SEB, Valencia, Spain. (See also in Section 3.6.)

- *Cebrián, A. C.*, **Castillo-Mateo, J.**, & Asín, J. (2022, June 20–24). *Nonparametric tests to detect trends based on theory of records. An application to the analysis of climate change* [Contributed talk]. International Symposium on Nonparametric Statistics, Paphos, Cyprus.

- ***Castillo-Mateo, J.*** (2020, November 24–December 9). *Tests for non-stationarity based on theory of records: An application to climate change* [Contributed talk]. 1st International Workshop on Stochastic Processes and their Applications, Online.

"The situation is this: ¡Oh! ¡No tengo datos!"

Jesús Asín, in a group meeting

ORIGINAL PAPER

Record tests to detect non-stationarity in the tails with an application to climate change

Ana C. Cebrián[1] · Jorge Castillo-Mateo[1] · Jesús Asín[2]

Abstract

The analysis of trends and other non-stationary behaviours at the extremes of a series is an important problem in global warming. This work proposes and compares several statistical tools to analyse that behaviour, using the properties of the occurrence of records in i.i.d. series. The main difficulty of this problem is the scarcity of information in the tails, so it is important to obtain all the possible evidence from the available data. First, different statistics based on upper records are proposed, and the most powerful is selected. Then, using that statistic, several approaches to join the information of four types of records, upper and lower records of forward and backward series, are suggested. It is found that these joint tests are clearly more powerful. The suggested tests are specifically useful in analysing the effect of global warming in the extremes, for example, of daily temperature. They have a high power to detect weak trends and can be widely applied since they are non-parametric. The proposed statistics join the information of M independent series, which is useful given the necessary split of the series to arrange the data. This arrangement solves the usual problems of climate series (seasonality and serial correlation) and provides more series to find evidence. These tools are used to analyse the effect of global warming on the extremes of daily temperature in Madrid.

Keywords Non-parametric tests · Records · Extremes · Non-stationary tails · Trend · Record-breaking temperatures · Climate change

1 Introduction

Clear evidence of global warming has been found in many areas of the planet. Concerning temperature, there is no question that Earth's average temperature is increasing (Sánchez-Lugo et al. 2019), and there is a general consensus to conclude the existence of an increasing trend in its mean evolution. Most of the works on climate change focus on the analysis of mean values; however, other relevant aspects are changes in variability and the tails of the distributions. Many works show the interest of analysing whether the occurrence of extremes is affected by climate change (Benestad 2004; Xu and Wu 2019; Saddique et al.

2020). Moreover, consequences of global warming on human health, agriculture, and other fields are often related to the occurrence of increasingly intense extremes (Coumou and Rahmstorf 2012).

Although of great interest, the analysis of non-stationary behaviour in extremes is acknowledged to be difficult due to the scarcity of data. It is also difficult to link it to the mean evolution of temperature since, given the small magnitude of this trend in terms of the variability of the daily temperatures, its effect on the extremes is not evident. Even assuming that global warming affects the occurrence of extremes, there may be considerable climate variability in different areas of the planet, and more research on this topic is needed. Conducting this type of study would be eased by the existence of simple statistical tools, in the same way that studies about the mean temperature have been favoured by the availability of simple non-parametric tests, such as the Mann–Kendall (MK) test (Kendall and Gibbons 1990).

Climate models that do not adequately represent non-stationary behaviours in the extremes can yield important

✉ Ana C. Cebrián
acebrian@unizar.es

1 Departamento de Métodos Estadísticos. Matemáticas, University of Zaragoza, C. Pedro Cerbuna, 12, Zaragoza, Spain

2 Departamento de Métodos Estadísticos. EINA, University of Zaragoza, C. María de Luna, 3, Zaragoza, Spain

biases in the results, especially in extreme value statistics such as return values or return periods. For example, ensembles of climate-model simulations are often useless since the frequency of extremes is too low to be well sampled by the ensemble (Durran 2020). Statistical models that represent the whole distribution of temperature also tend to badly fit the tails of the distributions. The validation of climate models must include analysing their capability to properly reproduce the most extreme values, but specific statistical tests to this end are not available.

In this context, it is of great interest to develop statistical tools to analyse non-stationary behaviours in the extremes. The annual maxima or the excesses over the threshold are the traditional approaches to study extremes in environmental sciences, and they are still commonly used (Prosdocimi and Kjeldsen 2021). However, a different approach based on the analysis of records is proposed in this work. This approach has some important advantages due to the probabilistic properties of records. In particular, the fact that the distribution of the occurrence of records in an independent identically distributed (i.i.d.) series (X_t) does not depend on the distribution of X_t. This property makes it easier to develop distribution-free statistics, use Monte Carlo methods to implement inference tools, and join the information of the records of different series. Another advantage of using records is that they do not require the information of the whole series. This is common, for example, in sports or in old climate series. Coumou et al. (2013) underlines the interest of this type of analysis and the importance of quantifying how the number of temperature records is changing worldwide and establishing its relationship with global warming.

Different approaches have been used to study the occurrence of temperature records. Redner and Petersen (2007) compared the observed values of records with the expected values under a stationary climate in a descriptive way and using simulations with given distributions. Benestad (2004) compared the observed and expected numbers of records under stationarity in a more formal way using a χ^2 test and graphical tools. Another common approach is to assume a Gaussian distribution of the temperature. For example, Newman et al. (2010) used simulations to determine the influence of trends and long-term correlations on record-breaking statistics. Franke et al. (2010) investigated the asymptotic behaviour of the probability of record at a given time, and characterised it under several distributions. Coumou et al. (2013) used the probabilities of records by Franke et al. (2010) assuming a Gaussian distribution to make descriptive comparisons with the observed records in monthly temperatures. Wergen and Krug (2010) and Wergen et al. (2014) also used those probabilities: they found that they were useful only at a

monthly scale but found difficulties quantifying the effect of slow changes in daily temperature. Although all these approaches are useful, there is a lack of formal tests to evaluate the effect of climate change on very extreme temperatures.

The aim of this work is to propose a new approach based on the occurrence of records to detect non-stationary behaviours in the extremes of temperature series, as a tool to assess the existence of global warming. To that end, statistical tests to detect those non-stationary behaviours are required, and they have to consider the specific characteristics of climate series, such as serial correlation and seasonal behaviour. The underlying idea is to use the distribution of the occurrence of records in an i.i.d. series (X_t) to study whether the observed records are compatible with that behaviour. First, we consider the type of tests by Foster and Stuart (1954) based on the number of records, but we also propose some statistics based on the likelihood and the score function of the record indicator binary variables. In particular, we obtain the expression of a score-sum statistic based on those variables, and we prove that it is a particular case of the general family of weighted statistics based on the number of records proposed by Diersen and Trenkler (1996). The advantage of this score-sum statistic is that the weights do not have to be empirically selected since they are analytically obtained.

To improve the power of the statistics based on the upper records, Foster and Stuart (1954) considered the four types of records that can be obtained from a series: upper and lower records from the forward and the backward series. To join all the information in one test, they defined statistics based on the number of each type of record. In addition to this type of statistic, we suggest another approach to join the information, to combine the p-values of the tests for each type of record. To that end, and given the dependence between the four types of records, we calculate the covariance between the four statistics, and we apply the Brown method. Graphical tools based on the previous statistics that allow us to detect where the deviation of the null hypothesis appears are also proposed. Finally, an analysis of the size and the power of the tests under different situations is carried out, including common distributions used in the analysis of climate extremes, such as Pareto and Extreme value distributions. All the tools described in this work are implemented in the R-package *RecordTest* (Castillo-Mateo 2021), publicly available from CRAN.

The outline of the paper is as follows. Section 2 describes the motivating problem and the data. Section 3 presents two families of tests: the first uses the upper records only, and the second joins the information of four types of records. A simulation study to compare the size and power of the proposed tests is shown in Sect. 4.

Section 5 describes some graphical tools, and Sect. 6 analyses the effect of global warming on the extremes of daily temperature in Madrid (Spain). Finally, Sect. 7 summarises the main conclusions.

2 Description of the problem

The motivating problem of this work is the analysis of the effect of global warming on the extremes of a daily maximum temperature series, and the series in Madrid (Spain) is used as an example. Our aim is not only to objectively establish the existence of non-stationary behaviour at extreme temperatures but also to identify the time, the periods of the year, and the features where it occurs.

2.1 Data

The series of daily maximum temperatures in Madrid, T_x, recorded in °C from 01/01/1940 to 31/12/2019 is obtained from the European Climate Assessment & Dataset (Klein Tank et al. 2002); the observations of the 29th of February have been removed. Madrid is located in the centre of the Iberian Peninsula (40.4° N 3.7° W) at 667 m a.s.l. and has an inland Mediterranean climate (Köppen Csa). Winters are cold and summers are hot, with average temperatures in January and July of 10 and 31 °C, respectively.

The temperature series shows seasonal behaviour, and a strong serial correlation is clearly significant. Figure 1 shows the mean evolution of T_x, which has a slight trend, much lower than the variability of the series. Thus, it is not clear if this trend affects the extremes, particularly the occurrence of records. In addition, the trend in the mean temperature differs across the days of the year. This can be observed in Fig. 2 (left), which shows $\hat{\theta}_i$, the slope of a linear trend estimated by least-squares in the subseries of each day of the year, standardised in mean and standard deviation. The mean of the slopes is 0.0075, but they move from -0.0025 to 0.025.

Figure 2 (right) represents the evolution of the annual mean of T_x and its records. This plot shows that the increase in temperature and the occurrence of records at an annual scale are clearer than those at daily temperatures. This suggests that global warming manifests itself not only by global record-breaking temperatures but also by a higher number of days with extreme temperatures.

In summary, the temperature series present the following characteristics: strong seasonal behaviour, serial correlation, and different trend evolution within the year. It is noteworthy that the strong seasonal behaviour yields not only different distributions of the variables but also a high variability of the entire series. Moreover, to study the effect of global warming on a daily scale, which is essential in climate applications, the increase in the number of warmer days must be considered. In the next section, we suggest an approach to arrange the data that deals with all these problems.

2.2 Data preparation

A common approach to remove the seasonal behaviour of a daily series with annual seasonality $(X_{1,1}, X_{1,2}, \ldots, X_{1,365}, X_{2,1}, X_{2,2}, \ldots, X_{T,365})$ is to split it into 365 subseries, one for each day of the year (Hirsch et al. 1982),

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,365} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,365} \\ \vdots & \vdots & & \vdots \\ X_{T,1} & X_{T,2} & \cdots & X_{T,365} \end{pmatrix}_{T \times 365}.$$

In this way, each column in the matrix is a series formed by serially uncorrelated observations with no seasonal behaviour. Serial uncorrelatedness is checked by applying Pearson correlation tests to study whether the correlation between the series and the lagged series is null. The
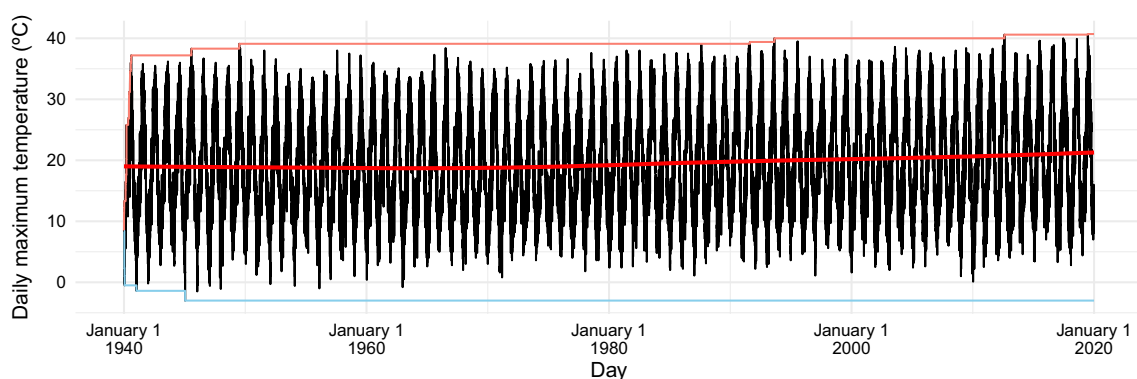

Fig. 1 Daily maximum temperature and lower (blue) and upper (red) records, Madrid
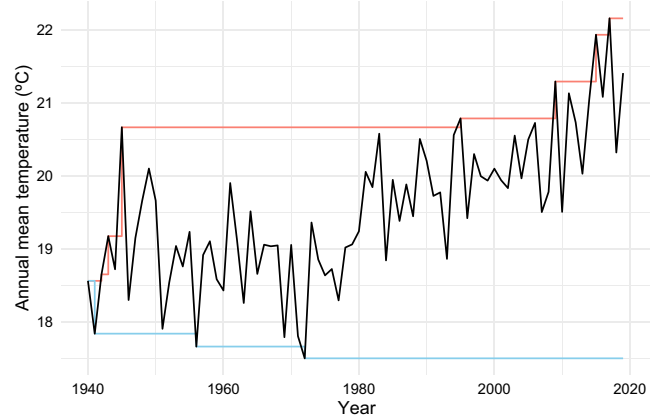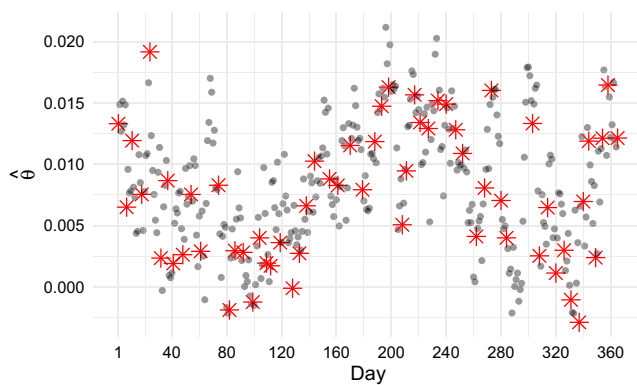
Fig. 2 (Left) Slopes of the linear trend in T_x on each day of the year; red stars mark the days whose series can be considered independent. (Right) Annual mean of T_x and upper (red) and lower (blue) records, Madrid

resulting 365 series do not have the same distribution, but given that the distribution of the occurrence of records does not depend on the distribution of i.i.d. series, the results can be aggregated.

The series of consecutive days are clearly correlated between them. To obtain a set of uncorrelated series that facilitates the development of inference tools, only a subset of the 365 series available is considered. To that end, the following approach is applied: given that the series at day k is in the subset, the correlation between series k and $k + 1$ is tested; if the correlation is not significant, series $k + 1$ is included in the subset; otherwise, the correlation between series k and $k + 2$ is tested. This step is repeated until a series $k + i$, which is not significantly correlated with series k, is found. Applying this approach to the temperature series in Madrid, and starting at 1st January, we obtain $M = 58$ series of length $T = 80$. The selected days are marked with red stars in Fig. 2 (left). One flaw of this method is that we are not using all the information in the 365 series available but, an interpretation is that it provides the effective sample size of the 365 correlated series. We also note that the set of selected subseries may depend on the method applied to assess the correlation.

In summary, the data preparation approach suggested here provides a set of M uncorrelated series with no seasonal behaviour. The transformation of one series into M subseries will be useful to obtain more evidence from the available data, as we will see in Sect. 3.

2.3 Challenges to analyse the tails of temperature series

To develop tests that are useful to detect non-stationarity in the temperature extremes, we have to consider all the characteristics of daily temperature and how data have to be arranged for analysis. Thus, we need tests with a high power to detect weak deviations of the null hypothesis, such as linear or other types of trends that may be small compared to the variability of the entire series. Second, non-parametric tests with few assumptions would be preferable so that they can be applied in a wide range of situations. Finally, due to the arrangement of the data resulting from splitting the series, the tests must be able to join information from several series, possibly with different distributions.

3 Statistical tests to study i.i.d. series

In this section, we review and propose a set of tests to study non-stationary behaviours in the occurrence of records that satisfy all the requirements described in the previous section. First, we recall some properties of the occurrence of records and the Monte Carlo method used in the development of the tests, and we state the null hypothesis to be analysed. Then, two families of tests are proposed, one based on the upper records and the other that joins the information of four types of records.

3.1 Background

3.1.1 Basic definitions and the classical record model

Given a series of variables (X_t), an observation of X_i is called an upper record (or a record) if its value exceeds all the previous observations, that is, $X_i > \max_{t<i}(X_t)$. Analogously, X_i is a lower record if $X_i < \min_{t<i}(X_t)$. Given that $\min_{t<i}(X_t) = -\max_{t<i}(-X_t)$, lower records can be defined in terms of the upper records of the negative series, and all the properties for the upper records can also be applied to the lower records.

The basic variables that characterise the occurrence of records in a series are the binary variables I_t, with $t \geq 1$, defined as

$$I_t = \begin{cases} 1 & \text{if } X_t \text{ is a record} \\ 0 & \text{if } X_t \text{ is not a record.} \end{cases}$$

Variable I_1 is always 1. N_t is defined as the number of records up to time t, $N_t = \sum_{i=1}^{t} I_i$.

The classical record model corresponds to the situation where we have the records of a series of i.i.d. continuous variables $(X_t)_{t \geq 1}$, with F being the common cumulative distribution function. An important result for the classical record model states that the distribution of record times does not depend on F (Arnold et al. 1998). The following properties, which are a consequence of the previous result, characterise in more detail the behaviour of variables I_t and N_t. They are useful for developing non-parametric tests with asymptotic distributions and easing the implementation of Monte Carlo approaches.

Property 1 Given a sequence of i.i.d. continuous variables (X_t), the sequence of variables (I_t) are mutually independent and each I_t is *Bernoulli*(p_t) distributed with,

$$p_t = P(I_t = 1) = 1/t, \quad t = 1, 2, \ldots.$$

Property 2 Given a series of i.i.d. continuous variables (X_t), the series of the corresponding variables (N_t) converges in distribution to a normal distribution,

$$\frac{N_t - \mu_t}{\sigma_t} \xrightarrow{D} N(0, 1)$$

where $\mu_t = E(N_t) = \sum_{i=1}^{t} \frac{1}{i}$ and $\sigma_t^2 = V(N_t) = \sum_{i=2}^{t} \frac{1}{i}\left(1 - \frac{1}{i}\right)$. Consequently, if t is large enough, N_t can be approximated by a $N(\mu_t, \sigma_t)$ distribution.

3.1.2 Monte Carlo method under the classical record model

Property 1 states that, under the hypothesis of i.i.d. series, the sequence of variables (I_t) are mutually independent and I_t is *Bernoulli*$(1/t)$ distributed, regardless of the continuous distribution F of (X_t). This makes easy the calculation of pivotal statistics based on (I_t). In addition, the Monte Carlo method is also easy to apply: for any F, a *Bernoulli*$(1/t)$ distribution can be used to generate samples of the variables I_t in i.i.d. series. Then, the implementation of the Monte Carlo method is standard: a series of T independent *Bernoulli*$(1/t)$ variables is generated, and the pivotal statistic $\mathcal{R}$ is obtained. Repeating this step B times, a sample of observations of $\mathcal{R}$ under the null hypothesis, $\mathcal{R}_1, \ldots, \mathcal{R}_B$, is obtained, and the p-value is estimated as

$\widehat{pv} = \sum_{b=1}^{B} I_b^{\mathcal{R}} / B$ where $I_b^{\mathcal{R}}$ is the binary variable linked to $\mathcal{R}_b \geq \mathcal{R}_o$, and $\mathcal{R}_o$ is the observed statistic.

The Monte Carlo method can also be used with statistics that are functions of dependent binary variables, as we will see in Sect. 3.3.2. The only difference from the previous approach is that a series of T independent $N(0, 1)$ variables (or any other continuous distribution) must be simulated first. Then, the corresponding series of the dependent record indicator variables I_t are obtained and used to calculate the pivotal statistic $\mathcal{R}$.

3.1.3 Notation and null hypothesis

In all the tests, we assume that there are M (≥ 1) mutually independent series of length T available, $(X_{t1}), (X_{t2}), \ldots, (X_{tM})$. These series can result from splitting the original series, or series measured at different spatial points for example. To keep the notation simple, it is assumed that all the series have the same length T, but this restriction is not necessary.

Given $(X_{t1}), (X_{t2}), \ldots, (X_{tM})$, we define the series of binary variables indicating the occurrence of records, $(I_{t1}), (I_{t2}), \ldots, (I_{tM})$, and the series of the number of records $(N_{t1}), (N_{t2}), \ldots, (N_{tM})$. We also define the number of records at time t in the M series, $S_t = I_{t1} + I_{t2} + \ldots + I_{tM}$.

The null hypothesis of all the proposed tests is

$$H_0 : p_{tm} = 1/t, \quad t = 1, \ldots, T, \quad m = 1, \ldots, M \quad (1)$$

with $p_{tm} = P(I_{tm} = 1)$. In the context of global warming, the most general alternative hypothesis is

$$H_1 : p_{tm} > 1/t, \quad \text{for at least one } t = 1, \ldots, T, \ m = 1, \ldots, M.$$
$$(2)$$

Hypothesis (2) includes the existence of a monotonous positive trend in the mean, a usual assumption of global warming, but it is more general. It also includes other types of non-stationarity, for example, nonmonotonous trends or series with increasing variability.

3.2 Tests based on the upper records

This section presents several approaches to building tests, assuming that only the upper records are available. First, we review and propose some modifications to a family of statistics based on the number of records N_t. Second, a new family of tests based on the likelihood function of the I_t variables is proposed.

3.2.1 Tests based on N_T

Property 2 states the normal asymptotic distribution of N_T, the number of records in a series. This property was used

by Foster and Stuart (1954) to define the d-statistic based on the difference between the upper and the lower records. Diersen and Trenkler (1996) found that the efficiency of the d-statistic and other similar statistics is improved by splitting a series of lengths T into k series of lengths T/k. Taking these results into account, the number of records statistic is defined as

$$\mathcal{N} = \sum_{m=1}^{M} N_{Tm} = \sum_{t=1}^{T} \sum_{m=1}^{M} I_{tm} = \sum_{t=1}^{T} S_t. \tag{3}$$

$\mathcal{N}$ is asymptotically normal under the null hypothesis since N_{Tm} are independent variables and are asymptotically normal when $T \to \infty$ (Property 2). On the other hand, by the central limit theorem, each variable S_t is also asymptotically normal under the null hypothesis, when $M \to \infty$. This means that the normal approximation of $\mathcal{N}$ is obtained when any or both M and T increase and that even with moderate values, a good approximation can be expected. The mean and the variance under the null are

$$e_{\mathcal{N}} = E(\mathcal{N}) = M \sum_{t=1}^{T} \frac{1}{t} \text{ and } v_{\mathcal{N}} = V(\mathcal{N}) = M \sum_{t=2}^{T} \frac{1}{t} \left(1 - \frac{1}{t} \right). \tag{4}$$

Using the previous distribution under the null, a test based on $\mathcal{N}$ is built. The p-value is $P(Z > (\mathcal{N}_0 - 0.5 - e_{\mathcal{N}}) / \sqrt{v_{\mathcal{N}}})$, where $\mathcal{N}_0$ is the observed value of $\mathcal{N}$ and $Z \sim N(0,1)$; note that since $\mathcal{N}$ takes integer values, a continuity correction has been applied.

It is noteworthy that using Property 1 and the independence of the M series, $S_t \sim Binomial(M, 1/t)$ under the null hypothesis. This means that the exact distribution of $\mathcal{N}$ is Poisson-Binomial, that is, the distribution of the sum of T independent $Binomial(M, 1/t)$ variables with $t = 1, \ldots T$. It does not have an explicit expression but it can be computed with numerical methods. However, it is not worth using since we have checked that the exact and asymptotic normal distributions are equivalent even with $M = 1$ and low values of T.

Weighted statistic. Diersen and Trenkler (1996) considered that more powerful statistics could be obtained by weighting the records according to their position in the series. The motivation is that the probability that an observation exceeds the actual record is inversely proportional to the number of previous observations. After an empirical study with different weights, the authors recommended the use of linear weights $w_t = t - 1$. They found that the asymptotic relative efficiency of the weighted tests improved with respect to the unweighted.

In our case, the weighted number of records statistic is

$$\mathcal{N}^w = \sum_{t=1}^{T} \sum_{m=1}^{M} w_t I_{tm} = \sum_{t=1}^{T} w_t S_t. \tag{5}$$

Under the null hypothesis, $\mathcal{N}^w$ is asymptotically normal when $M \to \infty$ since under those conditions S_t are asymptotically normal. However, we do not have normality when $T \to \infty$, since $\sum_{t=1}^{T} w_t I_{tm}$ is no longer asymptotically normal (Diersen and Trenkler 2001). The mean and the variance are $E(\mathcal{N}^w) = M \sum_{t=1}^{T} w_t \frac{1}{t}$ and $V(\mathcal{N}^w) = M \sum_{t=2}^{T} w_t^2 \frac{1}{t} \left(1 - \frac{1}{t} \right)$, and a test analogous to that based on $\mathcal{N}$ can be built.

Statistics with estimated variance The previous tests are based on the asymptotic normal distribution of the statistics with the expectation and variance obtained under the null. A disadvantage of these tests is that the expectation increases when the null hypothesis is not true but the variance also changes. This issue could diminish the power of the tests, since the statistic tends to take higher values but possibly with higher variability. We propose an alternative by standardising the statistic

$$\tilde{\mathcal{N}}_S^w = \frac{\mathcal{N}^w - E(\mathcal{N}^w)}{\sqrt{\widehat{V}(\mathcal{N}^w)}}.$$

Assuming that $V(I_{tm})$ is the same in the M series and denoting $N_T^w = \sum_{t=1}^{T} w_t I_t$, we have $V(\mathcal{N}^w) = M \, V(N_T^w)$. $V(N_T^w)$ can be estimated by the sample variance of the M values N_T^w obtained from (X_{tm}). Under the null hypothesis, $\tilde{\mathcal{N}}_S^w$ has a Student's t distribution with $M - 1$ degrees of freedom, and with $p_t > 1/t$, the statistic will tend to be far from 0. This statistic is also more robust to the existence of serial correlation. It is noteworthy that in correlated i.d. series, the probability of record is not $1/t$, but the deviations are negligible even with medium correlations.

3.2.2 Tests based on the likelihood function of I_t

Given M independent series of independent variables (I_{tm}), the likelihood is

$$\mathcal{L}(\mathbf{p_{TM}} \mid \mathbf{I_{TM}}) = \prod_{m=1}^{M} \prod_{t=2}^{T} p_{tm}^{I_{tm}} (1 - p_{tm})^{1 - I_{tm}} \tag{6}$$

where $\mathbf{p_{TM}} = (\mathbf{p_{T1}}, \ldots, \mathbf{p_{TM}})$ with $\mathbf{p_{Tm}} = (p_{1m}, \ldots, p_{Tm})$, and analogously $\mathbf{I_{TM}} = (\mathbf{I_{T1}}, \ldots, \mathbf{I_{TM}})$ with $\mathbf{I_{Tm}} = (I_{2m}, \ldots, I_{Tm})$. Note that for $t = 1$, the corresponding factor is $1^1 (1 - 1)^{1-1} = 1$.

Our aim is to propose tests based on the likelihood for studying the null hypothesis in (1). Standard likelihood tests assume that parameter values in the null hypothesis are interior points of the maintained hypothesis. However, in the one-sided alternative in (2), the parameters lie on the boundary of the parameter space so that standard regularity conditions fail to hold, and usual asymptotic distributions are no longer valid Gourieroux et al. (1982). We propose

tests under two different assumptions: first, the probabilities of record at time t of the M series are allowed to be different, and second, it is assumed that they are equal (but possibly different from $1/t$). The most common example of the second situation is when the M series have the same distribution.

General test for M independent series

If the M series are not assumed to have the same probabilities of record at time t, there are $M(T-1)$ different probabilities p_{tm}, without any restriction between them. Then, the number of unknown parameters p_{tm} is equal to the number of observations, and they cannot be estimated. An approach based on the score function, which only requires the estimation of the parameters under the null, is suggested.

Score-sum statistics. King and Wu (1997) proposed a general method of constructing a locally most mean powerful unbiased score test for one-sided alternatives. It has a small-sample optimal power property when no nuisance parameters are present, as in this case.

The statistic is based on the sum of the score vector, evaluated under H_0. In this case, using the likelihood in (6) and the null hypothesis in (1), the score vector is the $M(T-1)$ vector, $\mathbf{q}_0 = (q'_{1,0}, \ldots, q'_{M,0})$ where $q_{m,0}$ is a subvector of length $T-1$

$$q_{m,0} = \left[\frac{t}{t-1}(tI_{tm}-1)\right]_{t=2,\ldots,T}.$$

The information matrix under the null, $\mathcal{I}_0$, is a diagonal $M(T-1) \times M(T-1)$ matrix with diagonal vector $\mathcal{I}'_d = (\mathcal{I}'_{d1}, \ldots, \mathcal{I}'_{dM})$, where each subvector is

$$\mathcal{I}_{dm} = \left[\frac{t^2}{t-1}\right]_{t=2,\ldots,T}.$$

Then, we consider the statistic

$$\mathcal{S} = \frac{\sum_{m=1}^{M} q_{m,0}\mathbf{1}_{T-1}}{\sqrt{\mathbf{1}'_{M(T-1)}\mathcal{I}_0\mathbf{1}_{M(T-1)}}} = \frac{\sum_{t=2}^{T} t(tS_t - M)/(t-1)}{\sqrt{M\sum_{t=2}^{T} t^2/(t-1)}},$$

(7)

where $\mathbf{1}_K$ is the unity vector of length K. Assuming standard regularity conditions but without a requirement that the true parameter is an interior point of the parameter space and using the asymptotic normal distribution of the score vector, the distribution of $\mathcal{S}$ under the null is asymptotically $N(0, 1)$ when $M \to \infty$. Using this distribution, we can build a test, and the p-value for the alternative in (2) is $P(Z > \mathcal{S}_0)$, where $\mathcal{S}_0$ is the observed value of $\mathcal{S}$.

It is noteworthy that $\mathcal{S}$ is a linear function of variables S_t and $t = 2, \ldots, T$, with weights proportional to $t^2/(t-1)$.

This means that it is a particular case of the weighted statistic $\mathcal{N}^w$, with the advantage that the considered weights are analytically justified.

Tests for M independent series with the same distribution

The tests proposed in this section assume that the probabilities of records in the M series are equal, but they can be different from $1/t$. One advantage of this assumption is that the number of unknown parameters, $T-1$, is lower, and they can be estimated. In addition, it could be expected that in cases where the assumption is true, these tests would be more powerful than $\mathcal{S}$. However, the power study in Sect. 4 shows that this is not true. Since the use of these tests is not recommended, they are only briefly described here.

Shapiro (1988) showed that given a vector $\mathbf{y} \sim N(0, \mathbf{V})$ of dimension n and a convex cone C, the statistic

$$\bar{\chi}^2 = \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - \min_{\mathbf{b}\in C}(\mathbf{y} - \mathbf{b})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{b})$$

is distributed as $\bar{\chi}^2(V)$, a mixture of χ^2 distributions. If $\mathbf{V}$ is the identity matrix and $C = \{\mathbf{b} : \mathbf{b} \geq 0\}$, the weights of the mixture are $w_i = \binom{n}{i}2^{-n}$ for $i = 1, \ldots, n$. We apply this approach to $\mathbf{y} = \mathcal{I}_0^{-1/2}\mathbf{q}_0$, which under the null has an asymptotic distribution $N(0, 1)$. The resulting statistic is

$$\mathcal{T} = \sum_{t=2}^{T} \frac{(tS_t - M)^2}{M(t-1)}I_{S_t > M/t} \tag{8}$$

and using the asymptotic distribution, a test can be built as previously described.

We also derive a statistic based on the likelihood ratio function using the approach by Gourieroux et al. (1982) to deal with one-sided alternatives

$$\mathcal{R} = -2\sum_{t=2}^{T}\left(-S_t\log\left(\frac{tS_t}{M}\right) + (M - S_t)\left[\log\left(\frac{t-1}{t}\right)\right.\right.$$
$$\left.\left. - I_{S_t < M}\log\left(\frac{M - S_t}{M}\right)\right]\right)I_{S_t > M/t}.$$

Under the null hypothesis, it has a $\bar{\chi}^2(\mathcal{I}_0^{-1})$ asymptotic distribution. Since the numerical calculation of this distribution is computationally expensive, a Monte Carlo method is used, but the power performance is worse than $\mathcal{T}$.

3.3 Tests joining information from different types of records

When the entire series is available, the power of a test based on records can be improved by joining the information from the binary variables of the occurrence of lower records, denoted (I_t^L), and the occurrence of records in the backward series. This idea was suggested by Foster and

Stuart ([1954](#)) and Diersen and Trenkler ([1996](#)). The backward series are the series obtained when the order of the terms in the series are reversed so that we start by observing the last term,

$$I_t^B = \begin{cases} 1 & \text{if } X_{T-t+1} > \max_{i > T-t+1}(X_i) \text{ (record)} \\ 0 & \text{otherwise .} \end{cases}$$

Note that t is the index of the position in the backward series, so that I_t and I_t^B do not correspond to the same year. As variables I_t, under the null, $I_t^B \sim Bernoulli(1/t)$. Analogously, we define the binary variables for the occurrence of its lower records I_t^{BL}.

Two approaches are suggested to join the information from the four types of records. The first step is to obtain the type of statistic described in the previous section for each of the four types of records. Then, we can build a joint statistic, or we can join the resulting p-values using Fisher's type method. In both cases, it is necessary to characterise the dependence between the individual statistics and the expression of the correlation between them has to be obtained. The approaches presented here can be applied to any of the statistics in Sect. 3.2, but the results are presented for $\mathcal{S}$ since, as will be seen in Sect. 4, it is the most powerful statistic.

3.3.1 Correlation between statistics under the null

The statistic $\mathcal{S}$ in (7) can be expressed as $\mathcal{S} = \sum_{t=2}^{T} w_t S_t + K$, where K is a constant and

$$w_t = \frac{t^2}{(t-1)\sqrt{M \sum_{t=2}^{T} t^2/(t-1)}}.$$

All the calculations are expressed in terms of the weights w_t; in this way, the notation is simpler and the results are easily generalized to other $\mathcal{N}^w$ statistics. $\mathcal{S}^L$, $\mathcal{S}^B$ and $\mathcal{S}^{BL}$ denote the statistics based on the corresponding lower records or backward series. Note that using these weights, $V(\mathcal{S}) = V(\mathcal{S}^L) = V(\mathcal{S}^B) = V(\mathcal{S}^{BL}) = 1$.

Correlation between $\mathcal{S}$ and $\mathcal{S}^L$ ($\mathcal{S}^B$ and $\mathcal{S}^{BL}$). It can be proved that for $t, t' > 1$,

$$Cov(I_{tm}, I_{t'm}^L) = \begin{cases} -\dfrac{1}{t^2} & \text{if } t = t' \\ 0 & \text{if } t \neq t' \end{cases}$$

and $Cov(S_t, S_{t'}^L) = MCov(I_{tm}, I_{t'm}^L)$. Using this and $V(\mathcal{S}) = V(\mathcal{S}^L) = 1$,

$$Cor(\mathcal{S}, \mathcal{S}^L) = Cov(\mathcal{S}, \mathcal{S}^L) = -M \sum_{t=2}^{T} \frac{w_t^2}{t^2}.$$

Note that the resulting correlation does not depend on M since w_t^2 is multiplied by a factor $1/M$. These statistics are

asymptotically independent, and even for quite low T, the correlation is negligible, for example, -0.044 for $T = 50$.

Correlation between $\mathcal{S}$ and $\mathcal{S}^B$ ($\mathcal{S}^L$ and $\mathcal{S}^{BL}$). Using that $E(I_t I_{t'}^B) = P(I_t = 1, I_{t'}^B = 1)$, it is proved that for $t, t' > 1$,

$$Cov(I_{tm}, I_{t'm}^B) = \begin{cases} 0 & \text{if } t' < T - t + 1 \\ \dfrac{1}{T} - \dfrac{1}{t}\dfrac{1}{T-t+1} & \text{if } t' = T - t + 1 \\ -\dfrac{1}{tt'} & \text{if } t' > T - t + 1 \end{cases}$$

and the correlation is

$$Cor(\mathcal{S}, \mathcal{S}^{BL}) = M \sum_{t=2}^{T-1} w_t w_{T-t+1} \left(\frac{1}{T} - \frac{1}{t}\frac{1}{T-t+1} \right) - M \sum_{t=2}^{T} \sum_{t'=T-t+2}^{T} \frac{w_t w_{t'}}{tt'}.$$

These statistics show an increasing negative dependence, non-negligible; it is -0.667 for $T = 50$, and it approaches $-2/3$ with increasing T.

Correlation between $\mathcal{S}$ and $\mathcal{S}^{BL}$ ($\mathcal{S}^L$ and $\mathcal{S}^B$). It is proved that for $t, t' > 1$,

$$Cov(I_{tm}, I_{t'm}^B)$$
$$= \begin{cases} 0 & \text{if } t' < T - t + 1 \\ \dfrac{1}{t}\left(\binom{T}{t}^{-1} - \dfrac{1}{T-t+1} \right) & \text{if } t' = T - t + 1 \\ \binom{T}{t}^{-1}\left(\sum_{q=t}^{T} \binom{q}{t} \dfrac{1}{q(q-T+t'-1)} \right) - \dfrac{1}{tt'} & \text{if } t' > T - t + 1. \end{cases}$$

Then,

$$Cor(\mathcal{S}, \mathcal{S}^{BL}) = M \sum_{t=2}^{T-1} w_t w_{T-t+1} \frac{1}{t}\left(\binom{T}{t}^{-1} - \frac{1}{T-t+1} \right) +$$
$$+ M \sum_{t=2}^{T} \sum_{t'=T-t+2}^{T} w_t w_{t'} \left[\binom{T}{t}^{-1} \sum_{q=t}^{T} \binom{q}{t} \frac{1}{q(q-T+t'-1)} - \frac{1}{tt'} \right].$$
(9)

A simulation study shows that these statistics are asymptotically independent, and even for low T, the correlation is negligible, smaller than 0.03 for $T = 50$.

3.3.2 Generating a joint statistic

The idea of this approach is to join the information of $\mathcal{S}$, $\mathcal{S}^L$, $\mathcal{S}^B$ and $\mathcal{S}^{BL}$ into one statistic (Foster and Stuart [1954](#)). To that end, it must be taken into consideration that, under the alternative of an increasing trend, $\mathcal{S}$ and $\mathcal{S}^{BL}$ tend to have high positive values (since $p_t > 1/t$), while $\mathcal{S}^L$ and $\mathcal{S}^B$ have negative values (since in the corresponding series, $p_t < 1/t$). To unify the behaviour of the four statistics, we will consider linear combinations of $\mathcal{S}$; see (7), $-\mathcal{S}^L$, $-\mathcal{S}^B$

and $\mathcal{S}^{BL}$. The simplest option is to join the statistics that are asymptotically independent. For example,

$$\mathcal{S}2 = \mathcal{S} + \mathcal{S}^{BL}, \tag{10}$$

whose asymptotic distribution under the null is $N(0, \sqrt{2})$. We also consider

$$\mathcal{S}4 = \mathcal{S} - \mathcal{S}^L - \mathcal{S}^B + \mathcal{S}^{BL}, \tag{11}$$

that joins the four statistics available. Its expectation under the null is 0, and its variance is calculated using the covariances in Sect. 3.3.1. $\mathcal{S}4$ has an asymptotic normal distribution when $M \to \infty$ since it is the sum of M independent variables with the same distribution under the null. The p-value is also calculated using the Monte Carlo method in Sect. 3.1.2. The aim of this double calculation is to check the validity of the asymptotic normal distribution and state the values of M and T where the approximation can be used. The Monte Carlo method can be applied since $\mathcal{S}4$ is a pivotal statistic.

One advantage of this approach is its flexibility to define statistics, especially when the Monte Carlo method is used. Depending on the aim of the test, other statistics can be of interest, for example $\mathcal{S} - \mathcal{S}^B$ and $\mathcal{S}^{BL} - \mathcal{S}^L$ are a better option to study only the behaviour of the upper and the lower tail of the distribution, respectively. Other definitions have been tried, such as $\mathcal{S} - \mathcal{S}^L$, but the simulation analysis showed that $\mathcal{S}4$ is more powerful.

3.3.3 Generating a joint p-value

The idea of this approach is to join the information of P p-values, P_i, from the tests based on $\mathcal{S}$, $-\mathcal{S}^L$, $-\mathcal{S}^B$ and $\mathcal{S}^{BL}$ using Fisher's type method. Standard Fisher's approach states that the distribution of $\mathcal{X} = -2\sum_{i=1}^{P} \log P_i$ under the null is χ_{2P}^2, but it requires independent statistics. Using this approach, we propose the test $\mathcal{F}2$ that joins the p-values from $\mathcal{S}$ and $\mathcal{S}^{BL}$,

$$\mathcal{F}2 = -2(\log P_{\mathcal{S}} + \log P_{\mathcal{S}^{BL}}). \tag{12}$$

We also propose the test $\mathcal{B}4$ that joins the p-values of the four statistics using the modification by Kost and McDermott (2002) of the Brown approach to join dependent p-values. If the statistics are normally distributed, as the score statistics, the distribution of $\mathcal{X}$ under the null can be approximated by $c\chi_f^2$, where $c = V(\mathcal{X})/(2E(\mathcal{X}))$ and $f = 2E(\mathcal{X})^2/V(\mathcal{X})$. The expected value is $E(\mathcal{X}) = 2P$ and, using an approximation for the covariance $Cov(-2\log P_i, -2\log P_j)$,

$$V(\mathcal{X}) \approx 4P + 2\sum_{i=1}^{P}\sum_{i<j}^{P}(3.263\rho_{ij} + 0.710\rho_{ij}^2 + 0.027\rho_{ij}^3) \tag{13}$$

where ρ_{ij} is the correlation between the statistics i and j. Another restriction to join the p-values is that the statistics must have the same behaviour (increase or decrease) under the alternative. To achieve this, the p-values of the statistics $\mathcal{S}$, $-\mathcal{S}^L$, $-\mathcal{S}^B$ and $\mathcal{S}^{BL}$ must be joined,

$$\mathcal{B}4 = -2(\log P_{\mathcal{S}} + \log P_{-\mathcal{S}^L} + \log P_{-\mathcal{S}^B} + \log P_{\mathcal{S}^{BL}}). \tag{14}$$

Then, c and f are

$$c \approx 1 + \frac{1}{4}\sum_{l=1}^{3}(3.263\rho_l + 0.710\rho_l^2 + 0.027\rho_l^3)$$

$$f = 8/c$$

where $\rho_1 = \rho_{S,-S^L}$, $\rho_2 = \rho_{S,-S^B}$ and $\rho_3 = \rho_{S,S^{BL}}$, defined in Sect. 3.3.1.

4 Size and power analysis of the tests

This section summarises the main results from a simulation analysis of the size and power of the tests previously described. First, the tests based on the upper records, and second, the tests that require the entire series are analysed. The estimations are based on 10,000 replications and a significance level $\alpha = 0.05$.

In the size analysis, i.i.d. series with a $N(0, 1)$ distribution are generated without loss of generality since, under the null, the distribution of the statistics does not depend on the distribution of (X_t). The size of the tests is estimated for $M = 1, 4$ and 12 and $T = 25, 50$ and 100. The values of M correspond to common situations: non-split data, quarterly and monthly split data.

The power study focuses on comparing the tests under 'difficult conditions' since when M and T are high, or the series has a strong trend, all the statistics have a similar high power that approaches 1. Thus, only small sample sizes, $M = 1, 4, 12$ and $T = 25$ and 50, and small trends are shown in the study. The power is estimated using series with a monotonous positive trend θ since they satisfy the alternative hypothesis $H_1 : p_{tm} > 1/t$. Although other situations can lead to this alternative, a positive monotonous trend is the most common way of modelling global warming. Then, the series is generated as a trend plus a noise term

$$X_{tm} = \theta_t + Y_{tm}. \tag{15}$$

Noise terms Y_{tm} with different distributions are used to generate the series. We consider $N(0, 1)$ and other distributions commonly used in climate and environmental sciences. Series in these fields often require asymmetric and semibounded distributions (defined in $(0, \infty)$), such as the exponential or gamma distribution for rainfall, generalised Pareto (GP) for peaks over threshold applied to temperature or hydrological series, or generalised extreme value (GEV) for annual maximum temperatures or other maxima. Three kinds of monotonous trends were initially considered: linear, concave and convex trends, $\theta_t = \theta t$, $\theta_t = \theta \sqrt{tT}$, and $\theta_t = \theta t^2/T$, so that all of them increase to θT when $t \to T$. Given that very similar results are obtained in the three cases, only the results of the linear trend are shown. Concerning the magnitude, values $\theta = 0.005, 0.01, 0.025$ and 0.05 are considered. Note that in the case of $N(0, 1)$ noise, the trend θ corresponds to the $100 \times \theta$ percentage of the standard deviation.

In climate analysis, it is likely that the series used to implement the test do not exhibit the same trend, for example, in different seasons. Consequently, an alternative H_{1,θ_m} corresponding to this situation is included in the analysis: it uses M series generated with different trends $\theta_{tm} = \theta_m t$ in (15); the trend values θ_m are randomly generated from a $N(0.0075, 0.005)$ distribution. Values 0.0075 and 0.005 are the mean and standard deviation estimated from the sample of trends obtained in the temperature subseries in Madrid.

4.1 Comparison of the tests using only upper records

A thorough study of all the statistics described in Sect. 3.2 was performed. Only the results for $\mathcal{N}$, $\mathcal{S}$ and $\mathcal{T}$, see (3), (7) and (8) respectively, are summarised here since the statistics that are linear weighted functions of S_t ($\mathcal{N}^w$ and $\mathcal{S}$) yield similar results; the standardised version $\tilde{\mathcal{N}}_S^w$ neither improves the power with uncorrelated series. The performance of $\mathcal{T}$ and $\mathcal{R}$, the statistics for M series with the same probabilities of record, are also very similar between them.

Size analysis Table 1 summarises the size of the three statistics. It is adequate in all cases, even for $\mathcal{N}$ and $\mathcal{S}$,

which have an asymptotic distribution. The size of $\mathcal{S}$, which is asymptotic when $M \to \infty$, for $M = 1$ is slightly higher than the nominal value.

Power analysis Figure 3 summarises the power of the tests with series with a $N(0, 1)$ noise term and a trend. The following conclusions are obtained,

- The comparison of $\mathcal{N}$ and $\mathcal{S}$ shows that there is a clear increase in the power when weights are used. However, the value of the weights is not so relevant, and we found that the linear weights, obtained empirically, are equivalent, in terms of the power, to the theoretically derived weights from $\mathcal{S}$.
- Unexpectedly, the tests $\mathcal{T}$ and $\mathcal{R}$, which assume that the M series have the same probabilities of record, have a power lower than $\mathcal{S}$ even when that assumption is true.
- The power of the three tests increases with both T and M with a similar pattern.
- All the tests have a similar and high power with medium sample size and not very weak trends. However, $\mathcal{S}$ provides the best power under difficult conditions: it is over 0.9 for $\theta = 0.05$ with $M = 4$ and $T = 50$ or with $M = 12$ and $T = 25$ and for $\theta = 0.025$, with $M = 12$ and $T = 50$. For $\theta \leq 0.01$ and small sample sizes ($T \leq 50$ and $M \leq 12$), the tests are not useful since the power is low.

Alternative with series with different trends. The pattern of the power with series under the alternative H_{1,θ_m} is very similar to that obtained with series with a constant trend, and $\mathcal{S}$ provides the best results in all the settings (see Figure S.1 in Supplementary Material). The values are also equivalent; for example, the power of $\mathcal{S}$ under H_{1,θ_m} with random trends with mean 0.0075 and $T = 50$ and $M = 12$ is 0.3, while its counterpart with series with $\theta = 0.0075$ is 0.29.

Alternative with negative trends. The analysis of negative trends is not of great interest in climate. However, given that the occurrence of upper records in a series with a negative trend is symmetric to the occurrence of lower records with a positive trend, a brief analysis is carried out. The results are shown in Figure S.2 in the Supplementary Material, and the conclusions are analogous, although the power is slightly lower.

Table 1 Estimated size with $\alpha = 0.05$ of the tests based on the upper records

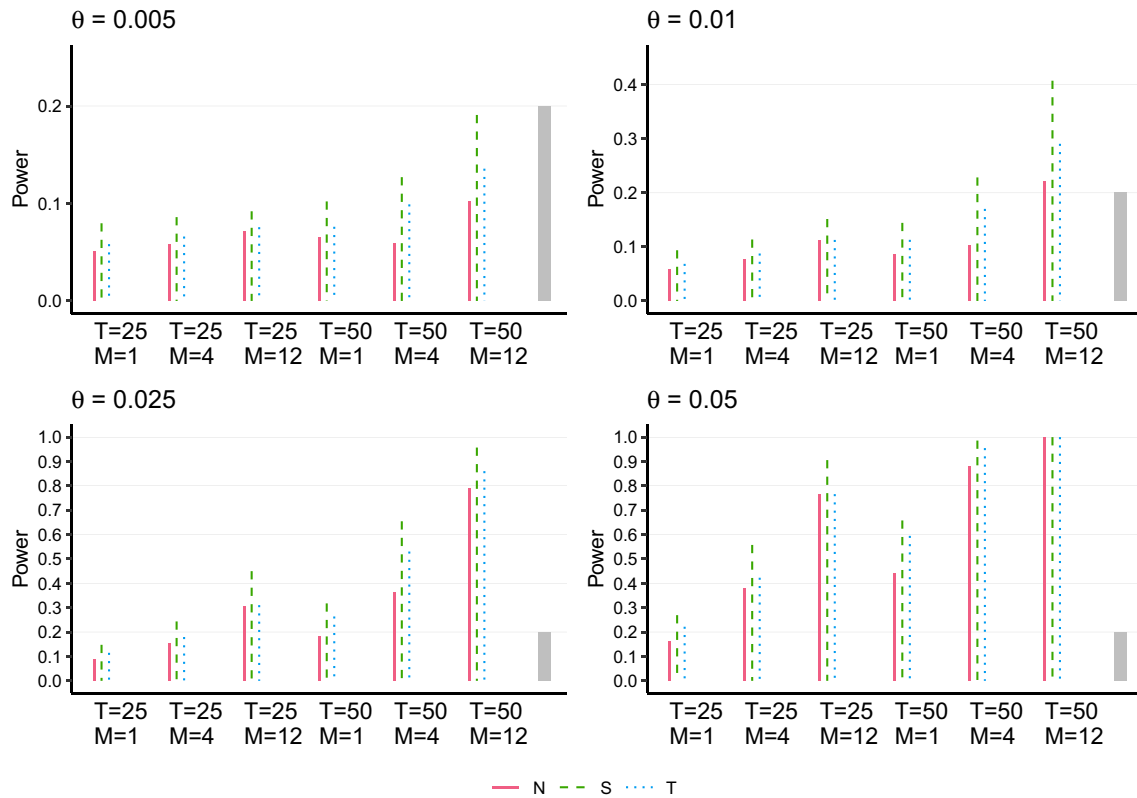| | T | 25 | 50 | 100 | 25 | 50 | 100 | 25 | 50 | 100 |
	M	1	1	1	4	4	4	12	12	12
$\mathcal{N}$		0.043	0.049	0.050	0.044	0.031	0.043	0.043	0.039	0.043
S		0.070	0.070	0.068	0.061	0.059	0.063	0.051	0.055	0.058
$\mathcal{T}$		0.049	0.049	0.050	0.050	0.047	0.052	0.047	0.051	0.052

Fig. 3 Power analysis of tests based on upper records using series with $N(0, 1)$ noise terms and the same trend

Some analyses using distributions other than the normal distribution are carried out, and as in the previous cases, $\mathcal{S}$ is the most powerful test. More results with other distributions are presented in the next section. In summary, $\mathcal{S}$ is the most powerful test in all the considered situations. Consequently, it is used to build the tests that join the information of different types of records.

4.2 Comparison of the approaches to join different types of records

This section summarises the performance of the approaches proposed to join the information from different types of records. We have a twofold objective: to analyse the improvement of the power when more than one type of record is used and to study whether it is more effective to build a joint statistic or a joint-p-value. Thus, the tests based on $\mathcal{S}2$, $\mathcal{S}4$, $\mathcal{F}2$ and $\mathcal{B}4$, see (10), (11), (12), and (14), respectively, are compared in the study.

Size analysis Table 2 summarises the size estimated for the four tests. It is adequate, although the tests based on only two statistics, especially $\mathcal{F}2$, tend to yield sizes slightly higher than the nominal value when M is low.

Power analysis Figure 4 summarises the power of the tests with series with $N(0, 1)$ noise terms and a trend; the statistic $\mathcal{S}$, based only on the upper records, is also included

for comparison purposes. The following conclusions are obtained.

- The pattern of the power is quite similar in all settings: the joint tests are clearly more powerful than $\mathcal{S}$, but the differences between them are small.
- Although the improvement resulting from joining information from different types of records is clear, the increase in the power with four types of records over two types is much lower. The power of $\mathcal{S}2$ and $\mathcal{F}2$ is slightly higher than $\mathcal{S}4$ and $\mathcal{B}4$ when $M = 1$; however, the power of $\mathcal{S}4$ increases faster with M. In any case, the differences are negligible, and given that the sizes of $\mathcal{S}4$ and $\mathcal{B}4$ are better with low sample sizes and that the computation time is similar, tests joining four types of records should be preferable.
- Concerning the approach to join the information, both joint statistics and joint p-values have a similar power in these settings. The power of $\mathcal{B}4$ is slightly lower with $M = 1$, but the differences are negligible.
- As expected, the power of the joint tests increases with both M and T, and it is over 0.8 for quite low sample sizes: for $\theta = 0.05$ with $M = 4$ and $T = 25$ or with $M = 1$ and $T = 50$ and for $\theta = 0.025$ with $M = 4$ and $T = 50$. To obtain a power higher than 0.8 for $\theta = 0.01$, a value $M = 20$ is required with $T = 50$, and with

Table 2 Estimated size with $\alpha = 0.05$ of the tests joining different types of records

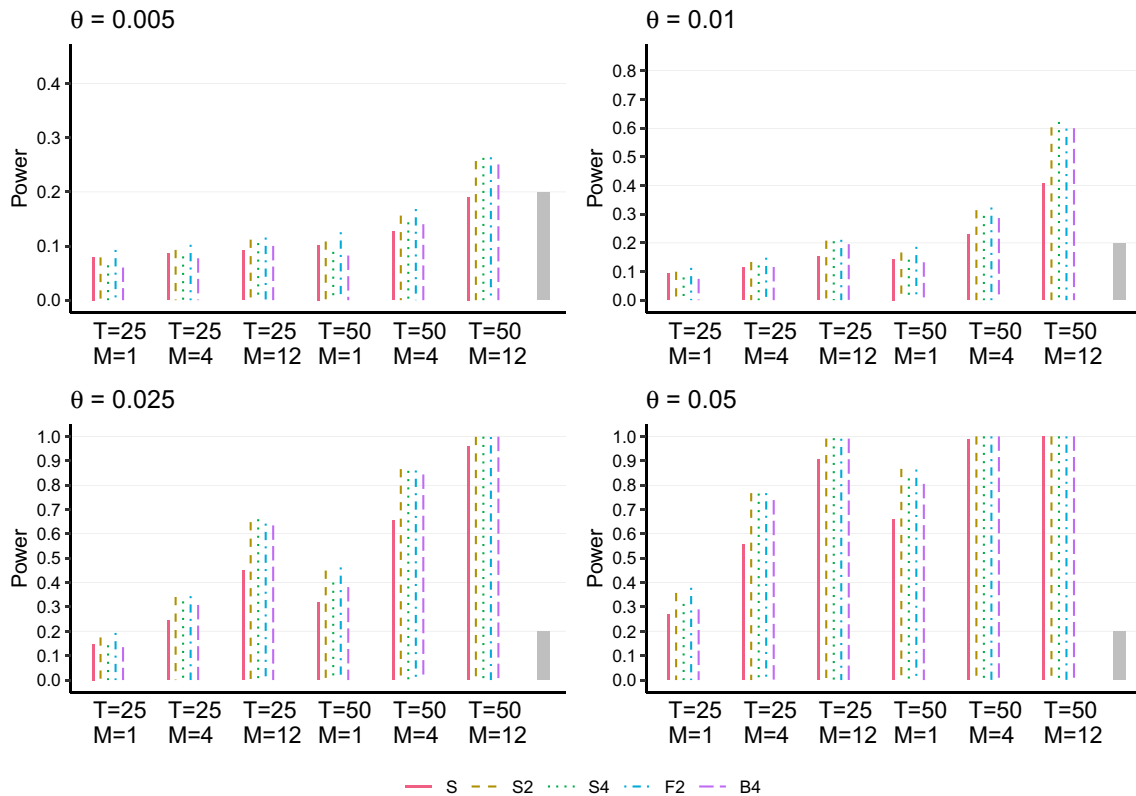| T	25	50	100	25	50	100	25	50	100
M	1	1	1	4	4	4	12	12	12
$\mathcal{S}2$	0.063	0.065	0.066	0.059	0.064	0.060	0.052	0.062	0.054
$\mathcal{S}4$	0.050	0.050	0.050	0.049	0.054	0.052	0.049	0.057	0.047
$\mathcal{F}2$	0.076	0.078	0.080	0.068	0.070	0.073	0.057	0.067	0.059
$\mathcal{B}4$	0.047	0.049	0.049	0.048	0.054	0.053	0.045	0.055	0.046

Fig. 4 Power analysis of tests based on four types of records using series with $N(0, 1)$ noise terms and the same trend

$M = 12$, a value $T = 62$. Analogously, for $\theta = 0.005$, a value $M = 80$ is required with $T = 50$, and with $M = 12$, a value $T = 115$.

Alternative with series with different trends. The conclusions about the power with series under the alternative H_{1,θ_m} are analogous. Figure 5 shows that the power of the four joint tests is higher than $\mathcal{S}$ and similar among them. The value of the power is very similar to that obtained with the M series with the same trend: it is 0.45 with $T = 50$ and $M = 12$, while the counterpart with the M series with trend $\theta = 0.0075$ is 0.43. A power higher than 0.8 is obtained with $T = 50$ and $M = 36$.

Alternative with series with non-normal distributions. The power of the tests is now estimated using distributions commonly used in climate and with different
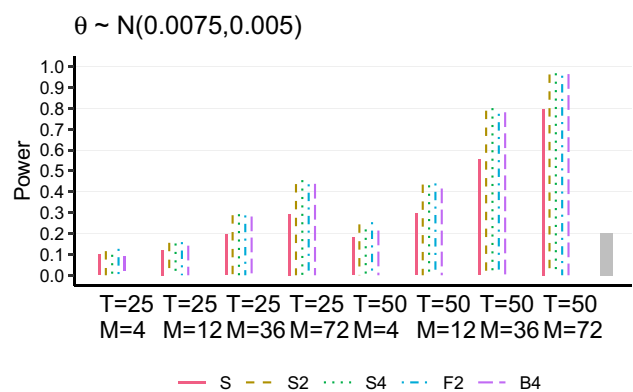

Fig. 5 Power analysis of tests based on four types of records using series with $N(0, 1)$ noise terms and different trends

types of tails (one-side bounded, heavier and lighter tails than the normal, etc.). Figure 6 summarises the power of tests $\mathcal{S}4$ and $\mathcal{B}4$ when they are applied to a series with a trend plus the one-side bounded distribution GP with shape parameter $\xi = -0.1$ (values $-0.5 < \xi < 0.5$ are common in climate) and $\sigma = 1$. In this case, the power is higher than that with normal noise terms in all settings. Even for $\theta = 0.005$ the power of $\mathcal{B}4$ is higher than 0.9 with $T = 50$ and $M = 4$, and for $\theta = 0.01$ it is also higher with $T = 25$ and $M = 12$. The power with noise terms $GP(\sigma = 1, \xi = 0)$, that is exponential distribution, $GP(\sigma = 1, \xi = -0.5)$, and $GP(\sigma = 1, \xi = -1)$, that is uniform distribution (two-side bounded), also have a higher power than the normal, see Figs. S.3, S.4 and S.5 in the Supplementary Material. The effect of the type of tails on the power is confirmed using the GEV distribution; Figs. S.6, S.7 and S.8 summarise the power for shape parameters $\xi = -0.05, 0, 0.5$ and $\sigma = 1$. The results lead to the conclusion that in one or two-side bounded distributions or in distributions with one or two tails lighter than the normal distribution, the power of the record tests is higher. With the GP distribution, the record tests are even more powerful than the MK test to detect trends in the mean. In effect, Fig. 6 shows that the power of MK is lower, specially with weak trends. This means that, in this type of distribution, the detection of trends is most

powerful if we focus on the behaviour of the bounded or light tail instead of focusing on the mean evolution. In that case, the power of $\mathcal{B}4$ is higher than $\mathcal{S}4$.

In summary, we conclude that the union of different types of records clearly improves the power of the test. We propose the use of tests based on $\mathcal{B}4$ or $\mathcal{S}4$, although, with bounded or light tail distributions, $\mathcal{B}4$ is slightly more powerful.

5 Graphical tools to detect non-stationary behaviour in records

The use of statistical tests is essential to obtain objective conclusions about the existence of non-stationary behaviours in the extremes. However, the use of graphical tools is also important to explore and characterise the existence of non-stationarity. Basic exploratory plots based on the times of occurrence of the records were used in Sect. 2, but here, we suggest some more elaborated plots together with confidence intervals (CIs). Two types of plots, based on variables N_t and estimated probabilities p_t, are proposed.
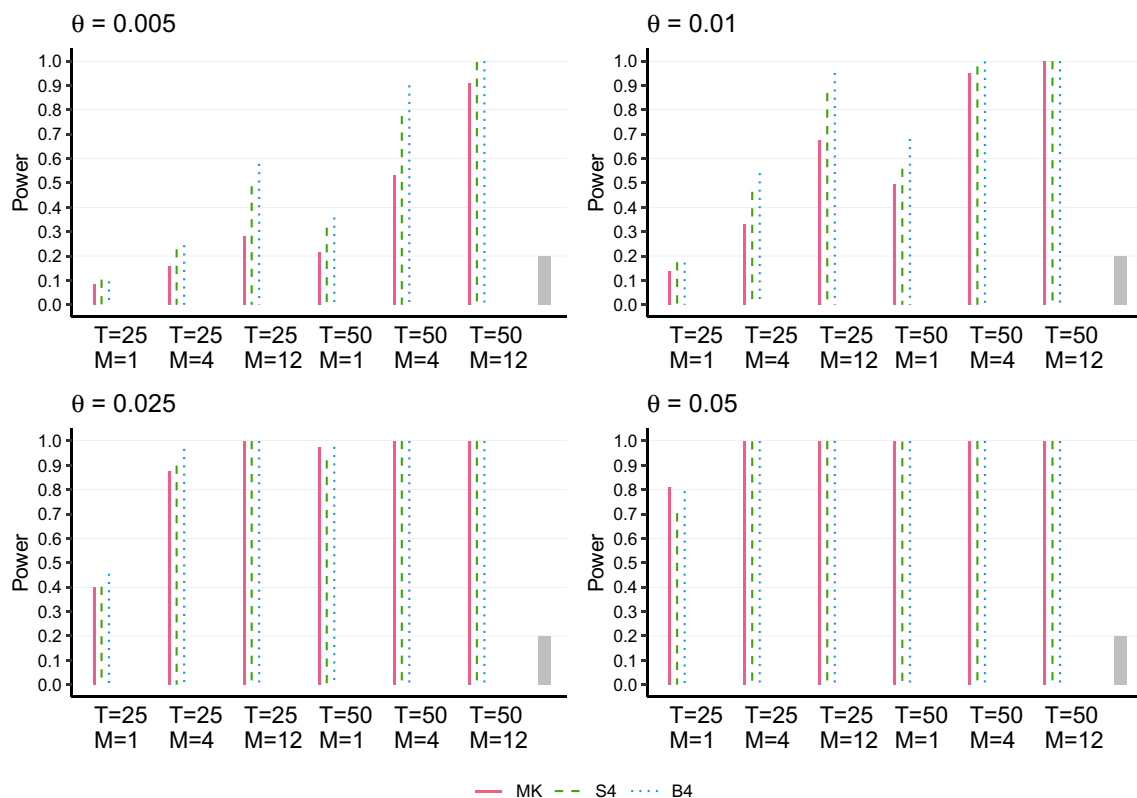

Fig. 6 Power analysis of tests based on four types of records using series with $GP(\sigma = 1, \xi = -0.1)$ noise terms

5.1 Plots based on $\bar{N}_t$

Using Property 2 and the approach in Sect. 3.2, it is obtained that the mean number of records up to time t in the M series, $\bar{N}_t = \sum_{m=1}^{M} N_{tm}/M$, is asymptotically normal (both in M and T) under the null. Using this result, approximate CI for $\mu_t = E(N_t)$, that is $\sum_{i=1}^{t} \frac{1}{i}$ under the null, are

$$\bar{N}_t \pm z_{1-\alpha/2} \sigma_t / \sqrt{M},$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of a $N(0, 1)$ distribution. These intervals together with the point estimator $\bar{N}_t$, and its expected value under the null can be plotted versus time. The resulting band is not a real confidence band of the values μ_t due to the dependence between the different $\bar{N}_t$. However, it is useful to observe deviations from stationarity in the evolution of the number of records and to identify the time point from which this deviation is significant. Another advantage is that the four types of records can be displayed in the same plot since their expected behaviour under the null hypothesis is the same.

The same approach can be used to make plots joining the number of lower and upper records in forward and backward series, which are also asymptotically normal when $M \to \infty$. It is noteworthy that at each time t, it is necessary to calculate the forward and backward records in the series observed only up to time t, not the number of backward records up to t for the series observed up to time T. If we define $\bar{D}_t = \sum_{m=1}^{M} D_{tm}/M$ with $D_{tm} = N_{tm} - N_{tm}^L - N_{tm}^{B,t} + N_{tm}^{BL,t}$, where $N_{tm}^{B,t}$ denotes the number of records in the backward series (X_{tm}) of the first t observations, the expected value of $\bar{D}_t$ under the null is 0.

5.2 Plots based on $\hat{p}_t$

The maximum likelihood estimators $\hat{p}_t = S_t/M$ satisfy $E(\hat{p}_t) = 1/t$ under the null, or equivalently, $E(t\hat{p}_t) = 1$. When (X_t) is not an i.i.d. sequence, there does not exist a general expression for p_t and $E(t\hat{p}_t)$. Assuming a series with a linear trend θ, Ballerini and Resnick (1985) proved that p_t has an asymptotically constant limit $\lim_{t\to\infty} p_t = \beta_0$ if the distribution has a finite first moment. The assumption of $p_t = \beta_0 + \beta_1/t$ is compatible with the previous result. Then, it is reasonable to consider as a general alternative, the regression model $E(t\hat{p}_t) = \beta_1 + \beta_0 t$, for $t > 1$, whose expected behaviour under the null is $\beta_0 = 0, \beta_1 = 1$. Consequently, the plot of $t\hat{p}_t$ versus time under the null hypothesis should be a random cloud of points centred around 1, and the fitted regression line should be $Y = 1$.

This model is heteroscedastic under the null, since $V(t\hat{p}_t) = (t-1)/M$. This implies that weighted least square estimators $\hat{\beta}_0^W$ and $\hat{\beta}_1^W$ with weights equal to the reciprocal of the variance must be used. CI for $E(t\hat{p}_t)$ can be obtained using that $t\hat{p}_t = tS_t/M$ and $S_t \sim Binomial(M, 1/t)$ under the null hypothesis.

6 Analysing the effect of global warming in records of daily temperature

The tests and graphical tools described in the previous sections are used to analyse the effect of global warming in the records of the series presented in Sect. 2, the daily temperature series in Madrid.

Joint analysis of the tails. Given that the complete series is available, and that our first aim is to study the existence of non-stationary behaviour in the tails of daily temperature, we start by applying tests based on different types of records. Taking into account the power study, the tests $\mathcal{S}4$ and $\mathcal{B}4$ are the most appropriate to assess $H_0 : p_{tm} = 1/t$. The resulting p-values, 2.4e–07 and 7.0e–08, lead to conclude at any usual significance level that the probability of record is higher than expected in an i.i.d. series. This means that there is evidence of non-stationarity in the occurrence of records due to an increasing trend.

Figure 7 shows the mean number of records $\bar{D}_t$ versus time, and it allows us to identify when the previous non-stationary behaviour appears and to characterise it over time. It is significant from 1980, and the value of the statistic shows an increasing trend from 2000.

Separate analysis of the upper and lower tails. We are also interested in studying whether this non-stationary behaviour appears in both tails of the distribution or only in one of them, and whether it is equally strong in both cases. To separately analyse the behaviour of the upper and lower records, we use the statistics $\mathcal{S} - \mathcal{S}^B$ and $\mathcal{S}^{BL} - \mathcal{S}^L$ respectively. The resulting p-values are 1.5e–06 and 0.006
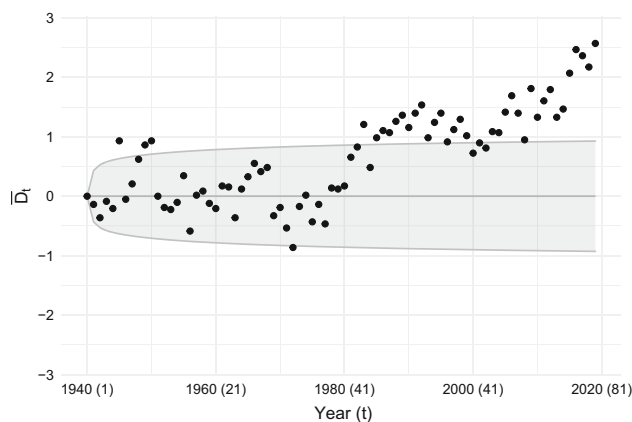

Fig. 7 Mean number of records $\bar{D}_t$ versus time and 90% CI under i.i.d. series, Madrid

so that although there is evidence of non-stationary behaviour in both tails, it is clearer in the upper records.

To study in more detail where non-stationary behaviour appears, Fig. 8 (left) shows the time plot of the number of upper and lower records in the forward and backward series, $\bar{N}_t$, $\bar{N}_t^L$, $\bar{N}_t^B$ and $\bar{N}_t^{BL}$, respectively. The forward series show weaker deviations of the i.i.d. hypothesis, and non-stationarity only appears in the upper records from 1990, becoming slightly significant from 2015. On the other hand, the backward series shows clearer deviations. This behaviour reveals that the effects of global warming are stronger in the upper tail and in the last years of the observed period. After five years of observations, the cumulative number of upper records in the backward series is significantly lower than expected in an i.i.d. series, and the consequences affect the rest of the period. However, from 1940 to 1980, the evolution of the number of records is quite parallel to the behaviour expected in a stationary series. The number of lower records is significant mainly due to the observed values between 1970 and 80, higher than expected in an i.i.d. series.

Figure 8 (right) shows the estimated probabilities of upper record $t\hat{p}_t$ for each year t together with the regression line and the confidence band. In an i.i.d. series, the slope of the regression line should be zero, while a positive slope is observed. In addition, many of the estimated probabilities are outside the CI from approximately 1980. This plot allows us to identify the specific years where the probability of record is much higher than expected. Similar plots can be made for the other types of records, and they confirm the previous conclusions.

Analysis of the tails by season. To study whether non-stationary behaviour differs across the seasons of the year and to identify the periods where it is stronger, the previous tests are applied separately to the four seasons of the year:

winter (DJF), spring (MAM), summer (JJA) and autumn (SON).

The resulting *p*-values are summarised in Table 3. If both tails are analysed jointly, using $\mathcal{S}4$ and $\mathcal{B}4$, the non-stationary behaviour is significant in all seasons except spring. However, if we study only the upper records, $\mathcal{S} - \mathcal{S}^B$, the evidence of trend is strong in summer and autumn but weak in winter. Concerning the lower records, $\mathcal{S}^{BL} - \mathcal{S}^L$, there is evidence of a decrease only in winter. No changes in lower records may be caused by an increase in variability. Figure 9 shows the cumulative number of each type of record by season. The plot in spring suggests that although the joint tests are non-significant, the number of upper backward records is significantly lower than expected, but it is compensated by the higher than expected lower forward number of records. Analysing two separate periods, it is concluded that it is due to a decreasing trend before 1970 and an increasing trend afterwards.

7 Conclusions

In the context of global warming, it is clear the interest of analysing the existence of non-stationary behaviour in the tails of a series, particularly in its records. This work reviews and proposes several statistical tests and complementary tools to detect this type of behaviour in climate series, using the properties of the occurrence of records in series i.i.d. More precisely, the tests assess the null hypothesis $H_0 : p_{tm} = 1/t$ versus the alternative $H_1 : p_{tm} > 1/t$. From a methodological point of view, the following conclusions are obtained.

- The approach proposed to arrange the data, based on splitting the series, solves two usual problems of
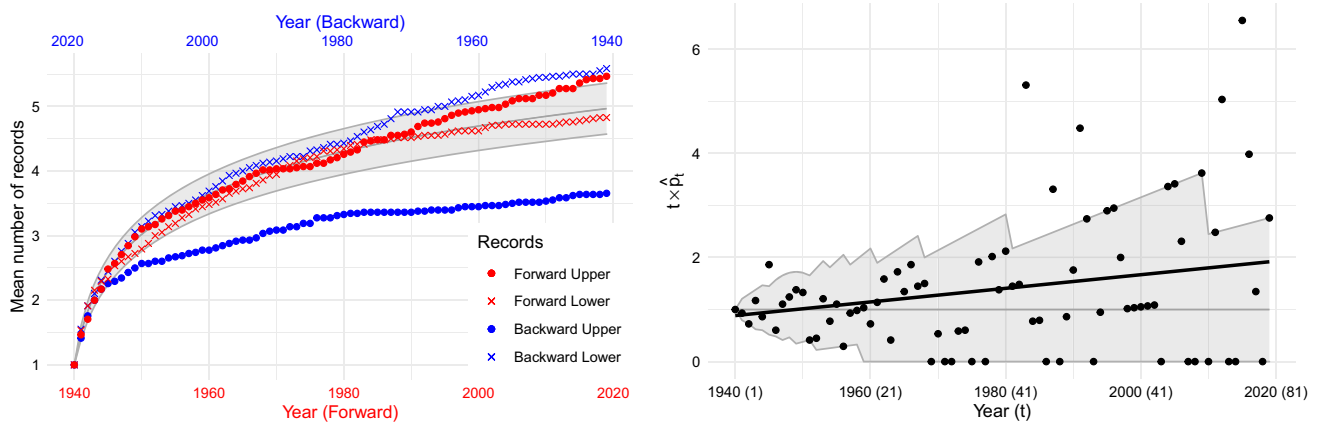

Fig. 8 Cumulative number of upper and lower records in the forward and backward series, $\bar{N}_t$, $\bar{N}_t^L$, $\bar{N}_t^B$ and $\bar{N}_t^{BL}$, and 90% CI under i.i.d. series (left). Regression line of $t\hat{p}_t$ versus time and 90% CI under i.i.d. series (right), Madrid

climate series: seasonal behaviour and serial correlation. There is another advantage of having M subseries of records available. Joining their information into one statistic, we are taking into account the increase of both, the magnitude of the highest temperatures and the number of warm days, maintaining a daily scale.

- A family of six tests based on the upper records is introduced. $\mathcal{N}$ is based on the number of records, and $\mathcal{N}^w$ and $\tilde{\mathcal{N}}_S^w$ are weighted versions of the previous one, the latter using an estimation of the variance. $\mathcal{S}$ is based on the likelihood function. Assuming that the M series

Table 3 p-values of the record tests by season

Period	Winter	Spring	Summer	Autumn
M	18	14	15	13
$\mathcal{S}4$	9.7e–05	0.52	4.1e–06	0.0002
$\mathcal{B}4$	7.1e–05	0.42	2.9e–07	4.5e–05
$\mathcal{S} - \mathcal{S}^B$	0.05	0.16	1.2e–07	5.5e–05
$\mathcal{S}^{BL} - \mathcal{S}^L$	0.0001	0.86	0.11	0.10

have the same distribution, two statistics based on the score and the likelihood ratio, $\mathcal{T}$ and $\mathcal{R}$, are considered. Asymptotic distributions are obtained for most of the previous statistics. In addition, the Monte Carlo method can be used to estimate the p-value in all cases, since they are pivotal statistics. This method is used to check the validity of the asymptotic distributions, revealing that they are valid even for $M = 1$ and low values of T. We conclude that statistics that are linear weighted combinations of variables $\mathcal{S}_t$, $\mathcal{N}^w$ and $\mathcal{S}$ are the most powerful. Statistic $\mathcal{S}$, whose weights are analytically justified, is proposed as the best test based on the upper records.

- The second family of tests aims to join the information of different types of records: the upper and the lower records of the forward and the backward series. Four statistics, $\mathcal{S}2$, $\mathcal{S}4$ (based on joint statistics) and $\mathcal{F}2$ and $\mathcal{B}4$ (based on joint p-values), that include two or four types of records are considered. A power study shows that the union of two or more types of records clearly improves the power of $\mathcal{S}$ based only on the upper records. The union of the statistics or the p-values yields tests with similar power. The power of the joint
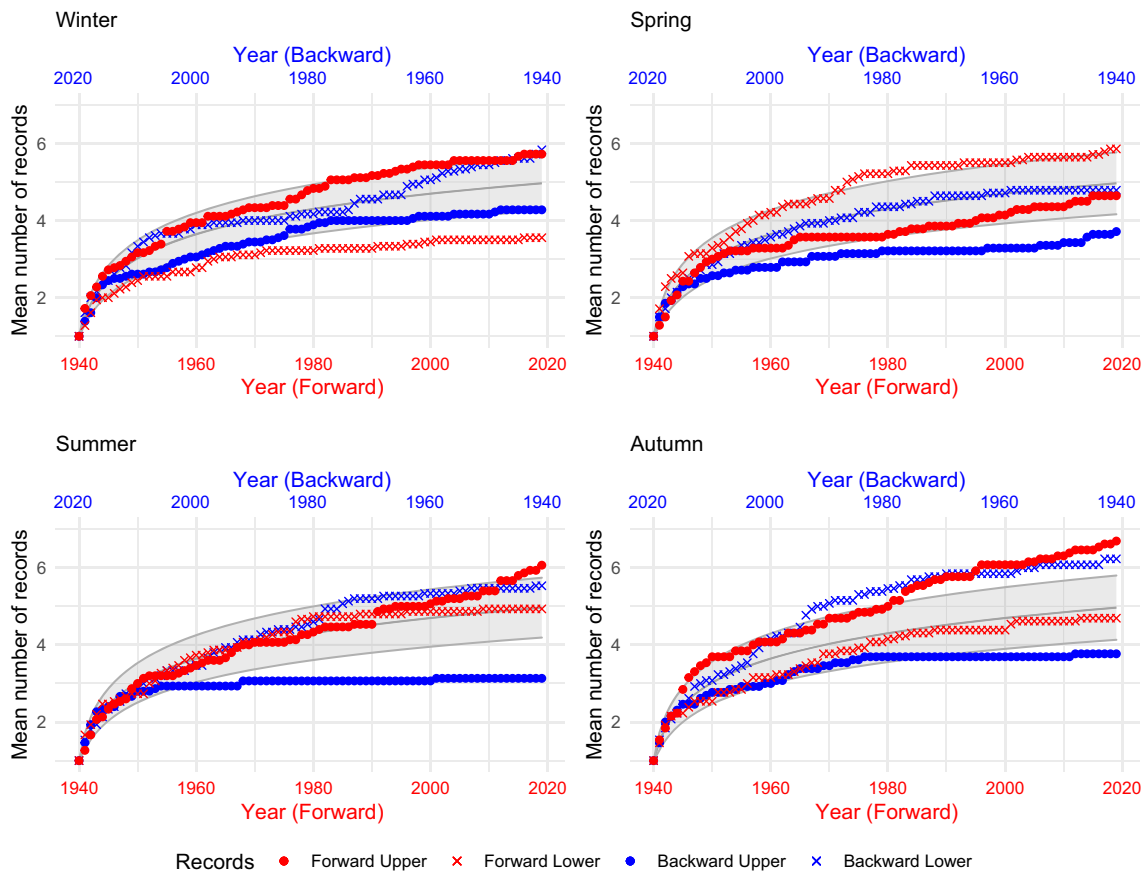

Fig. 9 Cumulative number of upper and lower records in the forward and backward series, $\bar{N}_t$, $\bar{N}_t^L$, $\bar{N}_t^B$ and $\bar{N}_t^{BL}$, and 90% CI under i.i.d series, by season, Madrid

tests with series with $N(0, 1)$ noise terms is high even with small sample sizes and weak trends. The power is higher for series with noise terms with one or two-side bounded distributions or distributions with one or two light tails, such as *GP* and *GEV*, which are often used in climate analysis. For *GP* distributions, even for a weak trend $\theta = 0.005$, the power of $\mathcal{B}4$ is higher than 0.9 with $T = 50$ and $M = 4$. The approaches suggested to join the information of different types of records are very flexible. They allow us to define other statistics to study specific features, such as non-stationary behaviour only in the upper or lower tail, or give more weight to a particular type of record.

- From the considered record tests, $\mathcal{B}4$ and $\mathcal{S}4$ are the most powerful. These tests have important advantages that make them specifically useful in the analysis of global warming. First, they have a high power to detect weak trends. They are non-parametric and require few assumptions, so that they can be applied in a wide range of situations. Moreover, they are able to join the information of M independent series, a property that is useful to deal with seasonal behaviour. This property can also be useful in the spatial analysis since it allows us to join series from different locations and to obtain global conclusions over the area of interest. The tests are complemented with graphical tools that aim to characterise where and when non-stationary behaviour occurs. Finally, all the tests and tools are easy to apply and are implemented in the R-package *RecordTest*.

The proposed inference tools are used to analyse the effect of global warming on the extremes of the daily temperature in Madrid. It is concluded that there is strong evidence of non-stationary behaviour in the tails of the distribution that affects the occurrence of records. This non-stationary behaviour is statistically significant from approximately 1980, and it increases from 2000. The behaviour is stronger in the upper tail, especially in the last years of the observed period. Moreover, the behaviour among seasons is not homogenous: it is significant in all seasons except spring. If we focus on the behaviour in the lower tail, it is only significant in winter.

The tests and graphical tools in this work are useful to analyse the extremes of observed series. In addition, they can also be used as tools for validating the capability to reproduce the most extreme values of climate models representing the entire distribution of a variable. This feature is important since a misrepresentation of the tails can yield important biases in their conclusions. The only condition to apply the tools for validating the tails of a climate model is that the model can generate trajectories of the variable under study. Examples include Earth system models (Wehner et al. 2020) and statistical models fitted by

Bayesian or other parametric methods. The general idea is to apply the tools to the trajectories generated by the climate model and to the observed series and to compare the results. The tests can also be used in other fields where the study of records is important including hydrology, finance, etc.

Author Contributions Conceptualization: AC; Data curation: JC-M, JA; Analysis: JC-M; Methodology: AC, JC-M; Software: JC-M; Supervision: AC; Validation: JC-M, AC, JA; Writing—original draft: AC; Writing—review and editing: JC-M, JA. All authors read and approved the final manuscript.

Availability of data and code. Data are available at the website of the European Climate Assessment & Dataset project (https://www.eca-d.eu). The code of the functions used in data analysis is available in the R-package *RecordTest* (Castillo-Mateo 2021). The code for the analysis of size and power is available by request.

Declarations

Conflicts of interest The authors have no financial or personal interest that can inappropriately influence this work.

References

Arnold B, Balakrishnan N, Nagaraja HN (1998) Records. Wiley

Ballerini R, Resnick SI (1985) Records from improving populations. J Appl Prob 22:487–502

Benestad RE (2004) Record-values, nonstationarity tests and extreme value distributions. Glob Planet Change 44(1–4):11–26

Castillo-Mateo J (2021) RecordTest: inference tools in time series based on record statistics. R package version 2.0.0. https://CRAN.R-project.org/package=RecordTest

Coumou D, Rahmstorf S (2012) A decade of weather extremes. Nature Clim Change 2:491–96

Coumou D, Robinson A, Rahmstorf S (2013) Global increase in record-breaking monthly-mean temperatures. Clim Change 118:771–82

Diersen J, Trenkler G (1996) Records tests for trend in location. Statistics 28(1):1–12

Diersen J, Trenkler G (2001) Weighted records tests for splitted series of observations. In: Kunert J, Trenkler G (eds) Mathematical statistics with applications in biometry: festschrift in honour of Prof. Dr. Siegfried Schach, Lohmar: Josef Eul Verlag, pp 163–178

Durran DR (2020) Can the issuance of hazardous-weather warnings inform the attribution of extreme events to climate change? Bull Am Meteor 101(8):1452–63

Foster FG, Stuart A (1954) Distribution-free tests in time-series based on the breaking of records. J R Stat Soc Ser B Stat Methodol 16(1):1–22

Franke J, Wergen G, Krug J (2010) Records and sequences of records from random variables with a linear trend. J Stat Mech Theory Exp P10013:1–21

Gourieroux C, Holly A, Monfort A (1982) Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. Econometrica 50(1):63–80

Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data. Water Resour Res 18(1):107–121

Kendall M, Gibbons JD (1990) Rank correlation methods. Oxford University Press

King M, Wu P (1997) Locally optimal one-sided tests for multiparameter hypotheses. Econom Rev 16(2):131–56

Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Mileta M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Forland E, Mietus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnar T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander LV, Petrovic P (2002) Daily surface air temperature and precipitation dataset 1901–1999 for european climate assessment (ECA). Int J Climatol 22:1441–53

Kost JT, McDermott MP (2002) Combining dependent p-values. Stat Probabil Lett 60(2):183–190

Newman W, Malamud B, Turcotte D (2010) Statistical properties of record-breaking temperatures. Phys Rev E 82:066111

Prosdocimi I, Kjeldsen T (2021) Parametrisation of change-permitting extreme value models and its impact on the description of change. Stoch Env Res Risk A 35:307–324

Redner S, Petersen M (2007) Role of global warming on the statistics of record-breaking temperatures. Phys Rev E 74:061114

Saddique N, Khaliq A, Bernhofer C (2020) Trends in temperature and precipitation extremes in historical (1961–1990) and projected. Stoch Env Res Risk A 34:1441–1455

Shapiro A (1988) Towards a unified theory of inequality constrained testing in multivariate analysis. Int Stat Rev 56:49–62

Sánchez-Lugo A, Berrisford P, Morice C, Nicolas JP (2019) Global surface temperature [in State of the climate in 2018]. Bull Am Meteor Soc 100(9):11–14

Wehner M, Gleckler P, Lee J (2020) Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: part 1, model evaluation. Weather Clim Extr 30:100283

Wergen G, Krug J (2010) Record-breaking temperatures reveal a warming climate. EPL 92:300–08

Wergen G, Hense A, Krug J (2014) Record occurrence and record values in daily and monthly temperatures. Clim Dyn 42:1275–89

Xu K, Wu C (2019) Projected changes of temperature extremes over nine major basins in China based on the cmip5 multimodel ensembles. Stoch Env Res Risk A 33:321–39

Supplementary Material: Record tests to detect non-stationarity in the tails with an application to climate change

Cebrián, A.   Castillo-Mateo, J.   Asín, J.

1   Power of tests based on the upper records

1.1   Power for the alternative hypothesis H_{1,θ_m}

Figure S.1 summarizes the power of the statistics $\mathcal{N}$, $\mathcal{S}$ and $\mathcal{T}$ based on the upper records under the alternative H_{1,θ_m} using series with Normal noise term. We use M series with different θ_m values generated from a $N(0.0075, 0.005)$ distribution. Values 0.0075 and 0.005 are the mean and standard deviation estimated from the sample of trends obtained in the temperature series in Madrid.
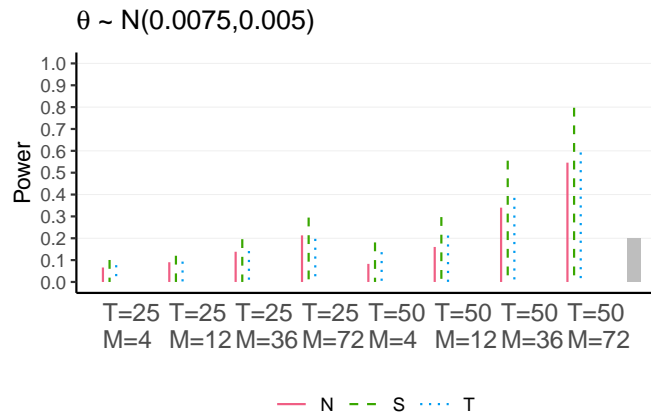
Figure S.1: Power analysis of tests based on upper records under the alternative H_{1,θ_m} using series with $N(0,1)$ noise term.

1.2   Power for negative trends

Figure S.2 summarizes the power of the statistics $\mathcal{N}$, $\mathcal{S}$ and $\mathcal{T}$ based on the upper records, with series with $N(0, 1)$ noise term and with negative trends. The occurrence of upper records in a series with a negative trend is symmetric to the occurrence of lower records with a positive trend. It is noteworthy that with this alternative, $\mathcal{S}$ cannot be applied with $M = 1$, since its power is always 0. In effect, the strongest evidence of a negative trend that can be observed is $S_t = 0$, for $t \geq 2$; in that case, the value of $\mathcal{S}$ depends on T but its minimum value is -1.48 (for $T = 21$). That means that the p-value $P(\mathcal{S} < \mathcal{S}_0)$ obtained using the Normal approximation will be always higher than 0.069, and the null hypothesis never will be rejected at a signification level $\alpha = 0.05$.
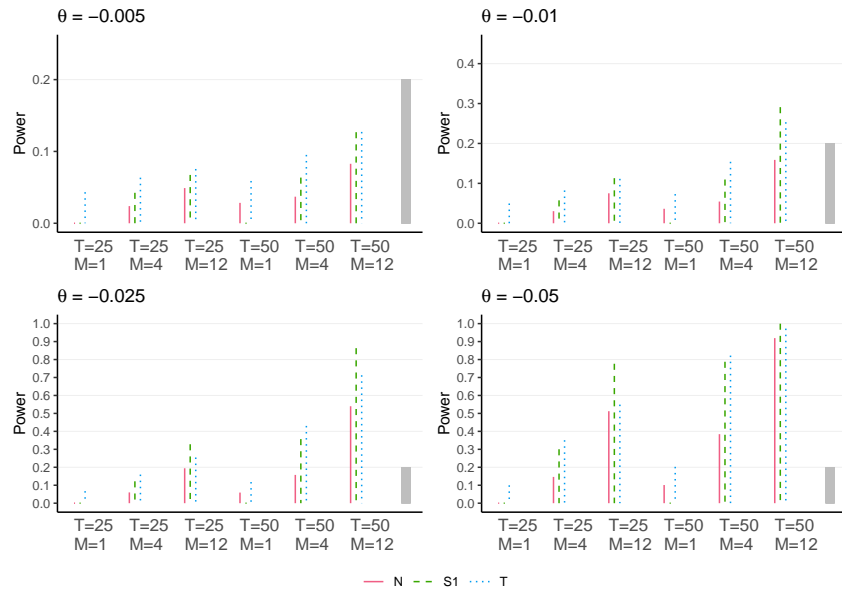
Figure S.2: Power analysis of tests based on upper records under the alternative of series with negative trends and $N(0, 1)$ noise term.

2 Power analysis of tests joining information from different types of records

2.1 Power for GP distributions

Figures S.3, S.4 and S.5 summarize the power of record tests $\mathcal{S}4$ and $\mathcal{B}4$ and rank-based MK test, when they are applied to a series with noise term following a GP distribution with different shape parameter ξ. Figure S.3 shows the case $\xi = 0$, that is the Exponential distribution, Figure S.4 shows $\xi = -0.5$, and Figure S.5, $\xi = -1$, that is the Uniform distribution (two-side bounded).
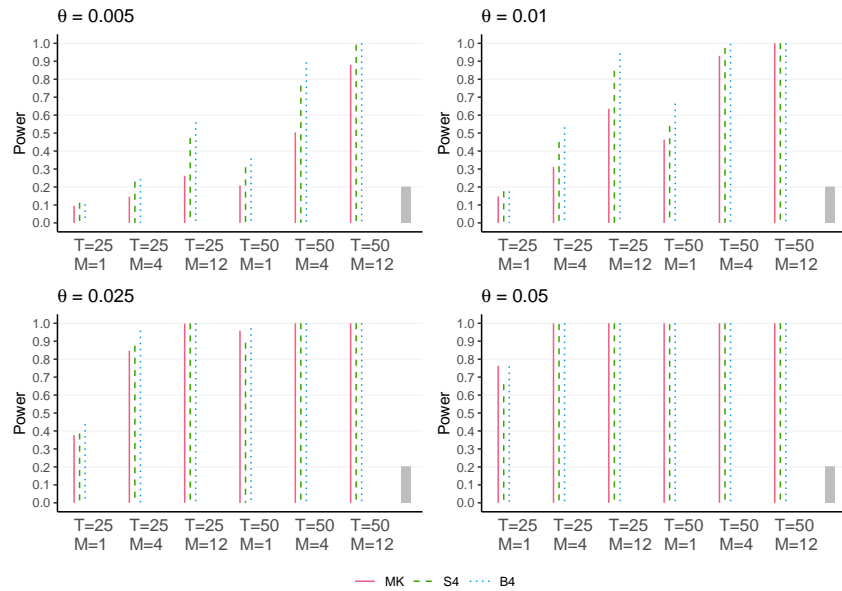
Figure S.3: Power analysis of tests under the alternative of series with $Exp(1)$ noise term.
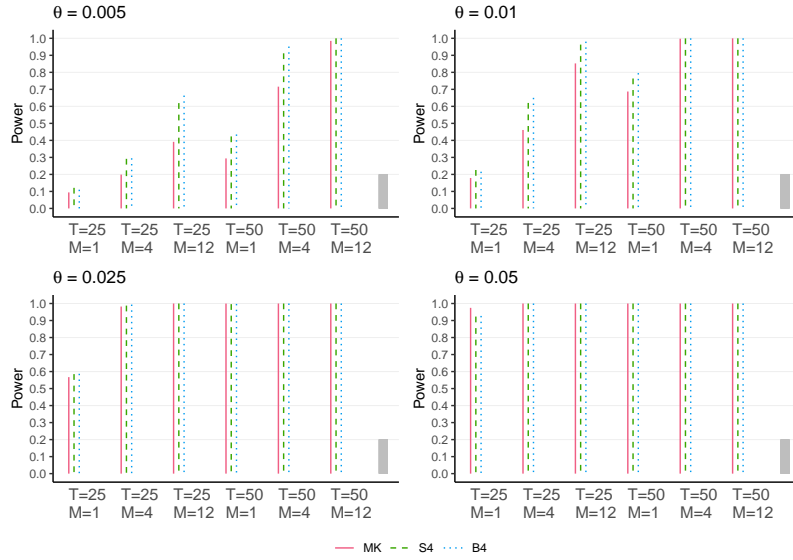
Figure S.4: Power analysis of tests under the alternative of series with $GP(-0.5)$ noise term.
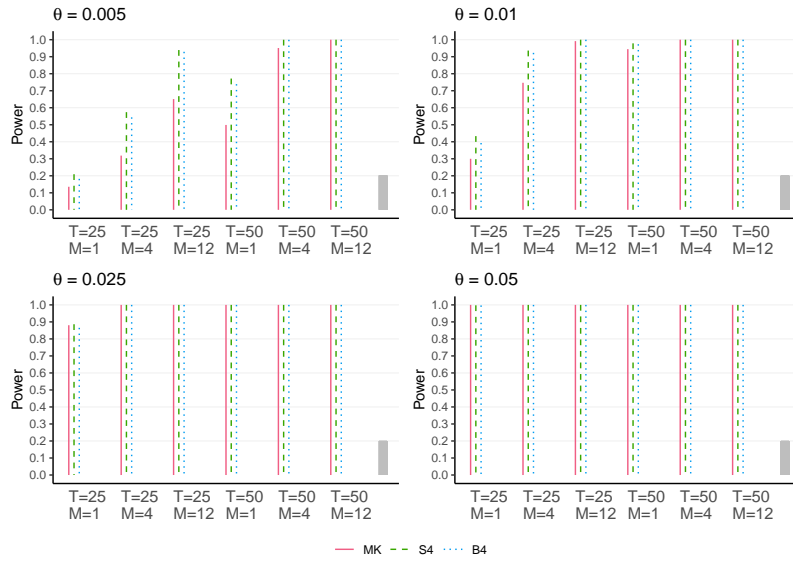
Figure S.5: Power analysis of tests under the alternative of series with $Uniform(0,1)$ noise term.

4

2.2 Power for GEV distributions

Figures S.6, S.7 and S.8 summarize the power of record tests $\mathcal{S}4$ and $\mathcal{B}4$ and rank-based MK test, when they are applied to series with noise term following a GEV distribution with different shape parameter ξ. Figure S.6 shows the case $\xi = 0$, that is the Gumbel distribution, Figure S.7 shows $\xi = 0.5$ and Figure S.8, with $\xi = 1$.
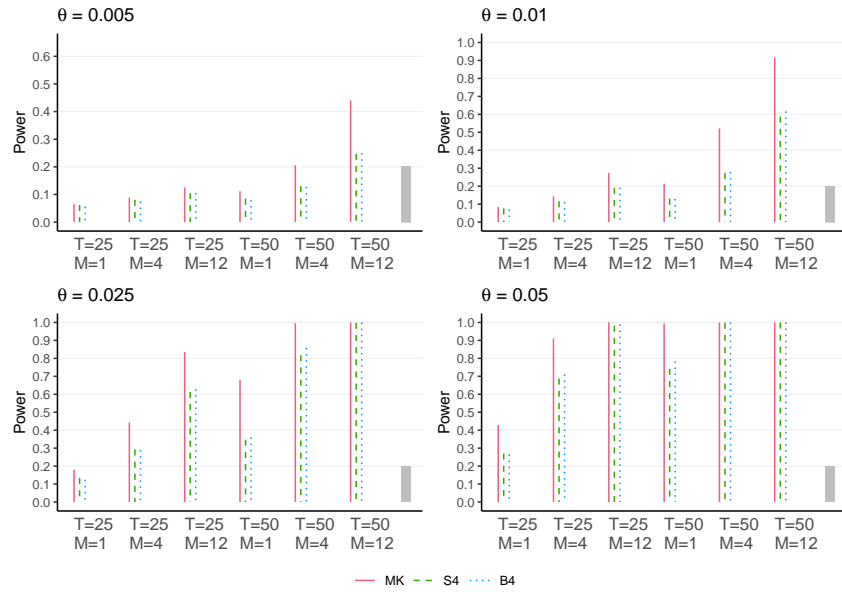
Figure S.6: Power analysis of tests under the alternative of series with Gumbel noise term.
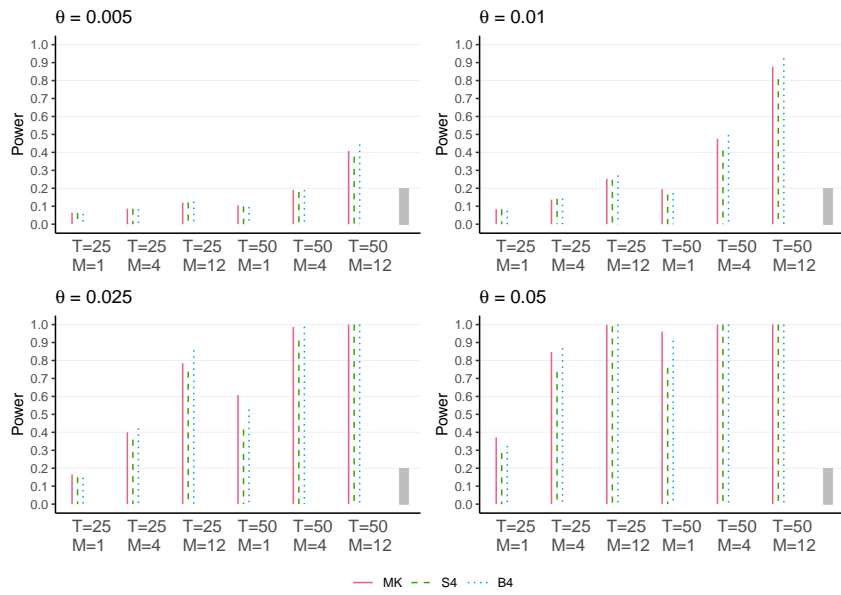
Figure S.7: Power analysis of tests under the alternative of series with GEV $\xi = 0.5$ noise term.
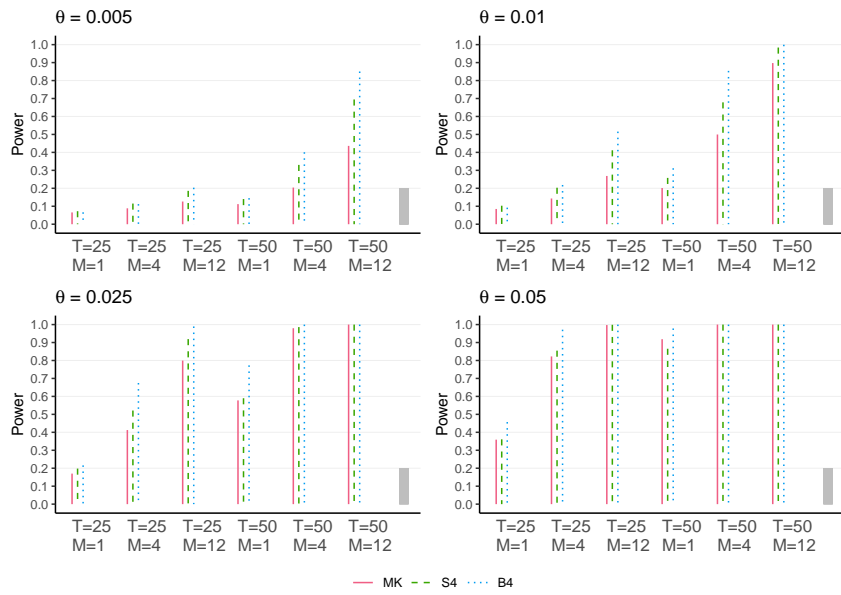
Figure S.8: Power analysis of tests under the alternative of series with GEV $\xi = 1$ noise term.

3.6 Distribution-free changepoint detection tests based on the breaking of records

This manuscript was published in:

Castillo-Mateo, J. (2022). Distribution-free changepoint detection tests based on the breaking of records. *Environmental and Ecological Statistics, 29*(3), 655–676. https://doi.org/10.1007/s10651-022-00539-2 [arXiv:2105.08186]

And it was disseminated (speaker emphasized) in:

- ***Castillo-Mateo, J.***, Cebrián, A. C., & Asín, J. (2022, September 14–16). *Records tests and applications to climate change* [Contributed talk]. VI Jornadas Científicas de Estudiantes de la SEB, Valencia, Spain. (See also in Section 3.5.)

- ***Castillo-Mateo, J.*** (2021, June 28–July 2). *Nonparametric changepoint detection tests based on the breaking of records* [Contributed talk]. Extreme Value Analysis 2021, Online.

"I read your recent paper in EES with great interest."

Lajos Horváth, in a personal communication

ORIGINAL PAPER

Distribution-free changepoint detection tests based on the breaking of records

Jorge Castillo-Mateo[1] ⬤

Abstract

The analysis of record-breaking events is of interest in fields such as climatology, hydrology or anthropology. In connection with the record occurrence, we propose three distribution-free statistics for the changepoint detection problem. They are CUSUM-type statistics based on the upper and/or lower record indicators observed in a series. Using a version of the functional central limit theorem, we show that the CUSUM-type statistics are asymptotically Kolmogorov distributed. The main results under the null hypothesis are based on series of independent and identically distributed random variables, but a statistic to deal with series with seasonal component and serial correlation is also proposed. A Monte Carlo study of size, power and changepoint estimate has been performed. Finally, the methods are illustrated by analyzing the time series of temperatures at Madrid, Spain. The R package `RecordTest` publicly available on CRAN implements the proposed methods.

Keywords Brownian bridge · Climate change · CUSUM · Nonparametric · Record-breaking · Wiener process

1 Introduction

An observation in a time series is called an upper (lower) record if it is greater (smaller) than all previous observations in the series. Therefore a new record is a remarkable event that attracts great attention in numerous applications, whether in environmental fields, economy, sports, physics or biology (see, e.g., Wergen 2013, and references therein). Particularly interesting is the study of record events in environmental sciences and their connection with climate change. For example, Benestad (2004) compared the observed and expected number of records under stationarity by means of a χ^2-test

---

Jorge Castillo-Mateo
jorgecm@unizar.es

[1] Department of Statistical Methods, University of Zaragoza, Zaragoza, Spain

and graphical tools. Respectively, Coumou et al. (2013) and Lehmann et al. (2015) found an increase in temperature and precipitation record-breaking events with respect to a stationary climate on a global scale. In addition to its many applications, the main foundations in the framework of theory of records can be found in the monographs Arnold et al. (1998) and Nevzorov (2001).

An aspect of interest is the study of the evolution of the number of records over time, in particular the identification of changes in their behavior. To analyze this type of change, changepoint detection methods that make use of the record occurrence should be considered.

The changepoint problem tries to identify times when the probability distribution function of a time series changes. In general the problem concerns both detecting whether or not a change has occurred and identifying its time of occurrence. Although several changes might be considered, our work resides in the at most one changepoint (AMOC) domain. The first results on changepoint detection start with Page (1954, 1955) who introduced a cumulative sum (CUSUM) statistic to locate a shift in the mean of independent and identically distributed (IID) normal random variables (RVs). Since then, several methods have been proposed, many of which can be found in the monographs Brodsky and Darkhovsky (1993) and Csörgő and Horváth (1997). Note-worthy is the importance of changepoint detection techniques in climatology (Reeves et al. 2007), but also in very different fields such as economy, speech processing, etc.

Traditional changepoint detection methods attempt to find changes in location or scale, more recently, changepoint detection in the extreme values has also been an active area of research. For example, Dierckx and Teugels (2010) introduced tests to detect changes in the parameters of the generalized Pareto distribution based on its likelihood for models of excesses over threshold, Kojadinovic and Naveau (2017) studied several tests for independent samples of block maxima, and e Silva et al. (2020) proposed a changepoint model for the r-largest order statistics. Ratnasingam and Ning (2021) proposed procedures based on the modified information criterion and the confidence distribution for detecting changepoints in the three-parameter Weibull distribution. Non-homogeneous Poisson processes have also been considered to study changepoints in the occurrence of peaks over threshold (Achcar et al. 2010, 2016; Rodrigues et al. 2019). To the best of our knowledge, there is no changepoint detection method based on the breaking of records. There are, however, tests for trend detection based on the breaking of records. Foster and Stuart (1954) proposed two simple statistics based on the number of records to test the hypothesis that T observations have been independently drawn from the same continuous distribution. These tests were later improved by Diersen and Trenkler (1996) and more recently new tests and graphical tools were introduced by Cebrián et al. (2022).

The aim of this paper is to develop changepoint detection tests based on the record occurrence to detect changes in the tails of the distribution. The first use of the tests introduced in this paper is to detect changes in the record occurrence and therefore in the extreme values, however, they are also useful against other types of change such as a change in location or scale. When there is a gradual change in location or scale, it will generally take time to be significantly reflected in a change in the behavior of the number of records, so the second use of the proposed methodology lies in analyzing how long it takes a series from when a changepoint is detected using another method

(see, e.g., Pettitt 1979, for a change in location), until that change is reflected in the observed records. Beyond its theoretical and descriptive interest, the third use of these changepoint detection tests based on records is that they would be uniquely appropriate whenever the original data are not available while records are.

The proposed tests make use of CUSUM-type statistics based on the record indicator RVs. The functional central limit theorem for independent but nonidentically distributed RVs is used to show that the functional evolution of the number of records adequately standardized behaves asymptotically as a Wiener process and, as a consequence, the CUSUM-type statistics follow the Kolmogorov distribution. This characterization allows to obtain exact p-values for the tests. The use of weights in the statistics can improve the power of the tests under certain scenarios. However, we prove that the weighted statistics do not have the same asymptotic properties as the previous ones and the p-value must be calculated using Monte Carlo techniques. An approach to analyze series with seasonal component or serial correlation is also proposed. The statistics based on the record indicators will allow studying the extreme values of the distribution with the advantage of not needing the specification of an underlying distribution for the data, i.e., they are distribution-free. Also, the requirement on the variance of data as in other CUSUM-type statistics is avoided here.

The rest of the paper is organized as follows. Section 2 introduces our records statistics, establishes their asymptotic distribution under the null hypothesis and proposes some generalizations. Section 3 compares these tests under various scenarios by means of Monte Carlo simulations. An application to temperature data is presented in Sect. 4, and Sect. 5 concludes the paper with final comments, conclusions and future work.

Finally, note that the proposed tests for changepoint detection are available from the R (R Core Team 2021) package `RecordTest` (Castillo-Mateo 2021).

2 Tests based on theory of records

Let $X_1, \ldots, X_T$ be a sequence of IID continuous RVs. The sequences of upper and lower record indicators, (I_t) and (I_t^L), are defined by $I_1 = I_1^L = 1$ and for $t = 2, \ldots, T$, by

$$I_t = \begin{cases} 1 & \text{if } X_t > \max\{X_1, \ldots, X_{t-1}\}, \\ 0 & \text{otherwise}, \end{cases} \qquad I_t^L = \begin{cases} 1 & \text{if } X_t < \min\{X_1, \ldots, X_{t-1}\}, \\ 0 & \text{otherwise}. \end{cases}$$

The sequence of differences in the upper and lower record indicators, (d_t), is given by $d_t = I_t - I_t^L$, while the sequence of sums, (s_t), is given by $s_t = I_t + I_t^L$.

The following lemma is a well known distribution-free result within the theory of records that characterizes the distribution of the record indicators, equally valid for upper and lower records (Arnold et al. 1998; Nevzorov 2001).

Lemma 2.1 *Let $X_1, \ldots, X_T$ be a sequence of IID continuous RVs. Then, the record indicators $I_1, \ldots, I_T$ are independent and*

$$p_t = P(I_t = 1) = \frac{1}{t}, \quad t = 1, \ldots, T.$$

It is easily checked that the expectations and variances for $t = 2, \ldots, T$, are

$$E(I_t) = \frac{1}{t}, \quad Var(I_t) = \frac{1}{t}\left(1 - \frac{1}{t}\right),$$

$$E(d_t) = 0, \quad Var(d_t) = \frac{2}{t},$$

$$E(s_t) = \frac{2}{t}, \quad Var(s_t) = \frac{2}{t}\left(1 - \frac{2}{t}\right).$$

Given p_t the probability of upper or lower record at time t, our aim is to construct asymptotic tests with null hypothesis

$$\mathcal{H}_0 : p_t = 1/t, \quad 1 \leq t \leq T,$$

against the two-sided alternative hypothesis given by

$$\mathcal{H}_1 : p_t = 1/t, \quad 1 \leq t \leq t_0 \quad \text{and} \quad p_t \neq 1/t, \quad t_0 < t \leq T, \tag{2.1}$$

where t_0 denotes the time of a possible change in the probabilities of observing new records with respect to the stationary case. The alternative hypothesis supports many nonstationary scenarios, for example a shift or a drift in location, variation or in one or both tails.

2.1 Tests based on asymptotic results

To obtain a p-value from a changepoint detection test, the exact distribution of the changepoint statistic is usually impractical, so it is generally preferable to have asymptotic results. Wiener processes, Brownian bridges and other Gaussian processes arise as asymptotic distributions in many limit problems providing exact tail probabilities. Our first objective is to build from the indicators above a random function $W_T(v)$, for $v \in [0, 1]$, in such a way that $W_T(v)$ converges in distribution to a Wiener process. For this purpose, we define the standardized record indicators, $\xi_{T1}, \ldots, \xi_{TT}$ as

$$\xi_{Tt} = \frac{I_t - E(I_t)}{\sigma_T}, \tag{2.2}$$

where $\sigma_t^2 = \sum_{k=1}^{t} Var(I_k)$. We also define the standardized number of records $S_{Tt} = \sum_{k=1}^{t} \xi_{Tk}$, its variance $v_{Tt} = \sum_{k=1}^{t} Var(\xi_{Tk}) = \sigma_t^2/\sigma_T^2$, and finally the

random function

$$W_T(\nu) = S_{Tt} + \xi_{T,t+1} \frac{\nu - \nu_{Tt}}{\nu_{T,t+1} - \nu_{Tt}} \tag{2.3}$$

for $\nu \in [\nu_{Tt}, \nu_{T,t+1}]$. Note that $S_{T1} = 0$, $\nu_{T1} = 0$ and $\nu_{TT} = 1$. It is noteworthy that the function $W_T(\nu)$ is a random broken line connecting points in the plane with coordinates (ν_{Tt}, S_{Tt}) for $t = 1, \dots, T$.

One of our major results is the asymptotic characterization of the functional evolution of the standardized number of records, $W_T(\nu)$, as a Wiener process. The result is essentially a consequence of the functional central limit theorem for independent but nonidentically distributed RVs (see, e.g., Gikhman and Skorokhod 1969). To be under the conditions of the theorem, Lindeberg's condition needs to be proved for the variables ξ_{Tt} in (2.2), which follows immediately from

$$\lim_{T \to \infty} \sum_{t=1}^{T} E\left(\xi_{Tt}^2 \times \mathbf{1}_{\{|\xi_{Tt}| > \varepsilon\}}\right) \leq \lim_{T \to \infty} \mathbf{1}_{\{1/\sigma_T > \varepsilon\}} = 0$$

for all $\varepsilon > 0$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

Theorem 2.1 *Let $X_1, \dots, X_T$ be a sequence of IID continuous RVs with $W_T(\nu)$ in (2.3). Then, as $T \to \infty$,*

$$W_T(\nu) \xrightarrow{\mathcal{D}} W(\nu), \quad \nu \in [0, 1],$$

in the metric space $\mathcal{C}[0, 1]$, where $W(\nu)$ is a standard Wiener process.

Thus, the changepoint records statistic proposed is

$$K_T = \max_{1 \leq t \leq T} |B_T(\nu_{Tt})|, \tag{2.4}$$

where $B_T(\nu) = W_T(\nu) - \nu W_T(1)$, $\nu \in [0, 1]$. The time t where (2.4) takes its maximum is the changepoint estimate $\hat{t}_0$. As a consequence of Theorem 2.1, $B_T(\nu)$ is asymptotically distributed as a standard Brownian bridge process. Moreover, the distribution of the supremum of the absolute value of a Brownian bridge is known as the Kolmogorov distribution. As $\sup_{0 \leq \nu \leq 1} |f(\nu) - \nu f(1)|$ is a continuous functional for f in $\mathcal{C}[0, 1]$, the asymptotic characterization under the null hypothesis of the statistic K_T is as follows.

Theorem 2.2 *Let $X_1, \dots, X_T$ be a sequence of IID continuous RVs with K_T in (2.4). Then, as $T \to \infty$,*

$$K_T \xrightarrow{\mathcal{D}} K = \sup_{0 \leq \nu \leq 1} |B(\nu)|,$$

where $B(\nu)$ is a standard Brownian bridge process and K is a Kolmogorov distributed RV.

The null hypothesis is rejected when K_T is too large to be explained by chance variation. In particular, if the alternative hypothesis in (2.1) is true for some time t_0, then it follows that $|B_T(\nu_{Tt_0})|$ is large and can show statistical evidence that a change occurred at time t_0. Under the null hypothesis, the p-value of the two-sided test can be calculated from any of the expressions of the Kolmogorov distribution

$$P(K \geq x) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp\left\{-2(kx)^2\right\}$$

$$= 1 - \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}\exp\left\{-\left(\frac{(2k-1)\pi}{2\sqrt{2}x}\right)^2\right\}.$$

To give a clear interpretation of K_T, we define $N_t = I_1 + \cdots + I_t$ the number of records up to time t and $N_{t_1:t_2} = I_{t_1} + \cdots + I_{t_2}$ the number of records between times $t_1 \leq t_2$. Then, $B_T(\nu_{Tt})$ can be rewritten as

$$B_T(\nu_{Tt}) = \frac{1}{\sqrt{Var(N_T)}}\left((N_t - E(N_t)) - \frac{Var(N_t)}{Var(N_T)}(N_T - E(N_T))\right).$$

Weighting for differences in the *effective* sample sizes of the number of records in two segments, $\{1, \ldots, t\}$ and $\{t+1, \ldots, T\}$, $B_T(\nu_{Tt})$ can be viewed as a scaled difference between $Var(N_t)^{-1}(N_t - E(N_t))$ and $Var(N_{(t+1):T})^{-1}(N_{(t+1):T} - E(N_{(t+1):T}))$. Consequently, K_T compares the number of records in both segments for every t and assigns as estimator, $\hat{t}_0$, the point that separates the segment that deviates the most from the null hypothesis. The mean is $E(B_T(\nu_{Tt})) = 0$ and simple calculation leads to $Var(B_T(\nu_{Tt})) = \nu_{Tt}(1 - \nu_{Tt})$. The nonuniform variance, small when it is near the ends of $\{1, \ldots, T\}$, makes changepoints occurring near the beginning or the end of the series more difficult to detect (see "Appendix A" for further details). This is a common fact in CUSUM-type statistics.

The proposed statistic only uses the information from one tail of the distribution, the right tail if upper records are used or the left tail if lower records are used. To study both tails and collect more evidence with a single statistic, it is enough to consider the variables d_t and s_t. Since the d_t's and s_t's also fulfill Lindeberg's condition, all the previous results are equally valid substituting ξ_{Tt} in (2.2) by $\xi_{Tt} = d_t/\sigma_T$ with $\sigma_t^2 = \sum_{k=1}^{t} Var(d_k)$, $t = 1, \ldots, T$; or respectively, $\xi_{Tt} = (s_t - E(s_t))/\sigma_T$ with $\sigma_t^2 = \sum_{k=1}^{t} Var(s_k)$, $t = 1, \ldots, T$. The statistic (2.4) based on d_t can be used when an increase in upper records and a decrease in lower records are expected with respect to the null hypothesis, while the statistic based on s_t can be used when an increase in both types of records is expected. In particular, the statistic based on d_t can be useful against the alternative hypothesis of a trend in location, while the statistic based on s_t can be useful against a trend in variation.

2.2 Tests with weighted statistics

Under the null hypothesis, the probability of record decreases as the series evolves. To give more importance to the most recent records and thus to be able to increase the power of the tests, we propose to give increasing weights, ω_t, to the different records according to their position in the series as

$$\xi_{Tt}^{\omega} = \omega_t \frac{I_t - E(I_t)}{\sigma_T}, \tag{2.5}$$

where $\sigma_t^2 = \sum_{k=1}^{t} \omega_k^2 Var(I_k)$, $t = 1, \ldots, T$. According to Proposition 2.1 (proved in "Appendix B"), these variables do not in general have an asymptotically normal sum, so asymptotic results such as those of Theorem 2.1 are not available.

Proposition 2.1 Let $X_1, \ldots, X_T$ be a sequence of IID continuous RVs with the sequence of RVs ξ_{Tt}^{ω} in (2.5) and $\omega_t \sim t^n$ as $t \to \infty$. If $n > 0$, then the central limit theorem does not hold for the ξ_{Tt}^{ω}'s.

Likewise, a K_T-type statistic in (2.4) associated with the weighted variables can be defined, and the distribution of which can be simulated by means of Monte Carlo techniques under the null hypothesis.

In this work we consider two different weights. First, linear weights $\omega_t = t - 1$ (see Diersen and Trenkler 1996, for a detailed explanation). Second, weights that make the discrete sequence of times of the process, $\nu_{Tt} = \sigma_t^2/\sigma_T^2$, $t = 1, \ldots, T$, equally spaced, i.e., weights proportional to the inverse of the standard deviation (SD) of I_t, i.e., $\omega_1 = 0$ and $\omega_t = Var(I_t)^{-1/2} = t/\sqrt{t-1}$ for $t = 2, \ldots, T$. These weights make the variance of $B_T(\nu_{Tt})$ symmetric in $\{1, \ldots, T\}$ (see "Appendix A").

As above, the statistic has been defined in terms of the I_t's but it is equivalent for the d_t's or s_t's. The SD in these cases suggests that the weights making the observed times of the process equally spaced are proportional to $\omega_1 = 0$, $\omega_t = \sqrt{t}$ for $t = 2, \ldots, T$, for the statistic based on d_t; and $\omega_1 = \omega_2 = 0$, $\omega_t = t/\sqrt{t-2}$ for $t = 3, \ldots, T$, for the statistic based on s_t.

2.3 Tests for seasonal series

Hirsch et al. (1982) introduced a seasonal version for tests of randomness based on ranks. Following their ideas, we propose tests which are insensitive to the existence of seasonality and serial correlation. If the time series data of interest are daily (or monthly) data, then the null hypothesis of randomness where all the observations come from the same continuous distribution may be too restrictive. For example, most series of daily temperature or precipitation show very strongly the presence of seasonality and serial correlation. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ be a sequence of series where $\mathbf{X}_m = (X_{1m}, \ldots, X_{Tm})'$ is a series of RVs. That is, $\mathbf{X}$ is the entire series, made up of subseries $\mathbf{X}_1$ through $\mathbf{X}_M$ (one for each day), and each subseries $\mathbf{X}_m$ contains annual values from day m, for $m = 1, \ldots, M$. Note that for further development the M subseries must be independent, so in general a subset of these subseries will be used. That is, below a subset of independent subseries is considered, but the notation is maintained for

simplicity. Then, we define the tth upper record indicator for the mth subseries as $I_{tm} = 1$ if $X_{tm} > \max\{X_{1m}, \ldots, X_{t-1,m}\}$ and $I_{tm} = 0$ otherwise; analogously for lower records. That is, records are calculated independently for each subseries, and the null hypothesis is relaxed allowing observations of different subseries not to come from the same distribution. To define a K_T-type statistic that joins the information of all the subseries, we simply take the ξ_{Tt}'s in (2.2) as

$$\xi_{Tt}^{\omega} = \omega_t \frac{\frac{1}{M} \sum_{m=1}^{M} I_{tm} - E(I_t)}{\sigma_T},$$

where $\sigma_t^2 = \sum_{k=1}^{t} \omega_k^2 Var(I_k)/M$; or their respective versions based on d_t or s_t. Thus, the alternative hypothesis is that of (2.1) with common changepoint t_0 for all the subseries. This approach not only allows the analysis of series with seasonal component, it also joins the information from several series, so the number of records and therefore the information used by the tests is greater.

3 Monte Carlo experiments

We investigate the empirical size, power and changepoint estimate of the changepoint tests based on the records statistics introduced in Sect. 2. Nine records statistics are considered: $N \equiv K_T$ in (2.4) with ξ_{Tt} in (2.2), d and $s \equiv K_T$ in (2.4) substituting I_t in (2.2) by d_t and s_t, respectively; and the previous statistics with weights proportional to the inverse of the SD of I_t, d_t and s_t, respectively (superscript *var*); and linear weights $t-1$ (superscript *linear*). Thus, three types of records statistics are analyzed. We denote by N-type statistics to the statistic N and its weighted versions, equivalently for d and s. Recall that, under the null hypothesis, the statistics N, d and s are asymptotically Kolmogorov distributed, while weighted statistics need Monte Carlo simulations to estimate their distribution (1000 replicates are considered).

3.1 Analysis of size

We simulate 10,000 replicates of M independent series formed by T independent samples from the standard normal distribution, i.e.,

$$Y_{tm} = \epsilon_{tm} \sim N(0, 1), \quad \text{for } t = 1, \ldots, T \quad \text{and} \quad m = 1, \ldots, M.$$

The size results are generalizable to any other continuous distribution given the distribution-free property of the tests under the null hypothesis. The size of the tests is simulated for the combination of values $T = 50, 100$, $M = 1, 12, 36$ and for a large series $T = 500$, $M = 1$.

Table 1 reports the empirical size results of the changepoint tests based on the records statistics N, d and s for nominal values $\alpha = 0.01, 0.05, 0.10$; i.e., we count how often the records statistics exceed the 99, 95, 90th percentile of the Kolmogorov distribution. We do not show the rejection frequencies of the tests based on the weighted

Table 1 Test size for $\alpha = 0.01, 0.05, 0.10$ level tests

Statistic	α	$T|M$ 50|1	100|1	500|1	50|12	100|12	50|36	100|36
N	0.01	0.011	0.011	0.012	0.006	0.008	0.005	0.005
	0.05	0.040	0.043	0.048	0.029	0.033	0.030	0.034
	0.10	0.068	0.076	0.083	0.059	0.068	0.068	0.075
d	0.01	0.004	0.005	0.008	0.005	0.006	0.005	0.006
	0.05	0.023	0.027	0.033	0.027	0.033	0.029	0.034
	0.10	0.051	0.057	0.065	0.057	0.064	0.059	0.066
s	0.01	0.009	0.009	0.010	0.005	0.006	0.004	0.006
	0.05	0.036	0.037	0.042	0.032	0.032	0.030	0.032
	0.10	0.067	0.076	0.082	0.065	0.074	0.061	0.068

statistics since their size is assured by simulating their p-value under the null hypothesis. All tests show an acceptable size for the levels α considered. Most of the tests are conservative, but their size approaches the nominal values as T increases. When M is greater than 1, the size of the statistics is considerably less than the nominal value. The size of d is particularly low, implying that these tests are very conservative.

In conservative tests, Fisher and Robbins (2019) proposed a general method to obtain a size closer to the nominal value and therefore increase the power of the tests. For our proposed tests, the method simply consists of changing the K_T-type statistic by $-\sqrt{T}\log(1 - K_T/\sqrt{T})$. Although we do not apply this method in the present paper, it may be a factor to consider in applications with low evidence since the power can increase while maintaining a proper size.

3.2 Analysis of power

The power analysis consists of 10,000 simulations of M independent series with T observations following two scenarios under the alternative hypothesis.

Scenario A. Linear drift model in the mean:

$$Y_{tm} = \mu_t + \epsilon_{tm}, \quad \text{for } t = 1, \ldots, T \quad \text{and} \quad m = 1, \ldots, M,$$

where $\epsilon_{tm} \sim N(0, 1)$, and $\mu_t = 0$ if $1 \leq t \leq t_0$ and $\mu_t = \theta(t - t_0)$ if $t_0 < t \leq T$.

Scenario B. Linear drift model in the SD:

$$Y_{tm} = \sigma_t \epsilon_{tm}, \quad \text{for } t = 1, \ldots, T \quad \text{and} \quad m = 1, \ldots, M,$$

where $\epsilon_{tm} \sim N(0, 1)$, and $\sigma_t = 1$ if $1 \leq t \leq t_0$ and $\sigma_t = 1 + \theta(t - t_0)$ if $t_0 < t \leq T$.

We report results for $T = 100$, $M = 1, 12, 36$, $t_0 = 25, 50, 75$ and the drift term $\theta = -0.10, -0.09, \ldots, -0.02, -0.01, -0.005, 0.005, 0.01, 0.02, \ldots, 0.09, 0.10$

for Scenario A and $\theta = 0.005, 0.01, 0.02, \ldots, 0.09, 0.10$ for Scenario B. N-type statistics are analyzed against both scenarios, d-type statistics against Scenario A and s-type statistics against Scenario B.

Figures 1 and 2 show, for $\alpha = 0.05$, plots of the power of the tests versus the trend θ for the Scenarios A and B, respectively. We make the following observations:

(1) All tests increase their power as the magnitude of the drift θ or the number of series M increases. In Scenario A, d-type statistics have a symmetric behavior with respect to a vertical line at $\theta = 0$, but when the drift is negative, N-type statistics have a power close to the nominal value unless M is large. This phenomenon is due to the fact that the greatest effect that a negative trend can cause is that only one record is observed in each series and under the null hypothesis it is likely to find a single record in a small number of series but it is unlikely to find a single record in many series. Finally, note that the power of tests with upper records against a positive drift is equivalent to that of tests with lower records against a negative drift.

(2) The power of the statistics according to the position of the changepoint depends on the type of weight used. The tests have a lower power for a changepoint t_0 close to the end of the series, since the accumulated trend is smaller. The unweighted statistics have a higher power when the changepoint is at the beginning of the series and lose power as it approaches the middle and especially the end of the series. The statistics with weights proportional to the inverse of the SD, in Scenario A, maintain the same power when the changepoint is in the first half of the series and lose power if the changepoint is at the end; in Scenario B, they have a higher power when the changepoint is in the middle of the series. The statistics with linear weights have a higher power when the changepoint is in the middle or the end of the series than at the beginning.

(3) For positive drifts and comparing statistics with the same type of weight. In Scenario A, N-type statistics have a higher power than d-type for low M, but when M is large this difference decreases and d-type have an equal or higher power than N-type. In Scenario B, s-type statistics have a higher power than N-type.

(4) The statistics with weights proportional to the inverse of the SD turn out to have the overall best performance with the most balanced behavior. The statistics without weights are those that have a higher power when the changepoint is at the beginning of the series, the statistics with weights proportional to the inverse of the SD have a higher power when the changepoint is in the middle of the series and the statistics with linear weights have a higher power when the changepoint is at the end of the series. While the second show a power close to the best in each case, the first and third have considerably less power than the others when they are not the most powerful.

(5) Some cases in which the records tests reach a power between 0.85 and 1 for $T = 100$ are given below. Under Scenario A, when $t_0 = 25$, we would detect $\theta = 0.05$ with $M = 1$ for statistic N or $\theta = 0.02$ with $M = 12$ for all statistics except those with linear weights or $\theta = 0.01$ with $M = 36$ for d. When $t_0 = 50$, we would detect $\theta = 0.10$ with $M = 1$ for all statistics or $\theta = 0.03$ with $M = 12$. Under Scenario B, when $t_0 = 25$, we would detect $\theta = 0.04$ with $M = 1$ for the
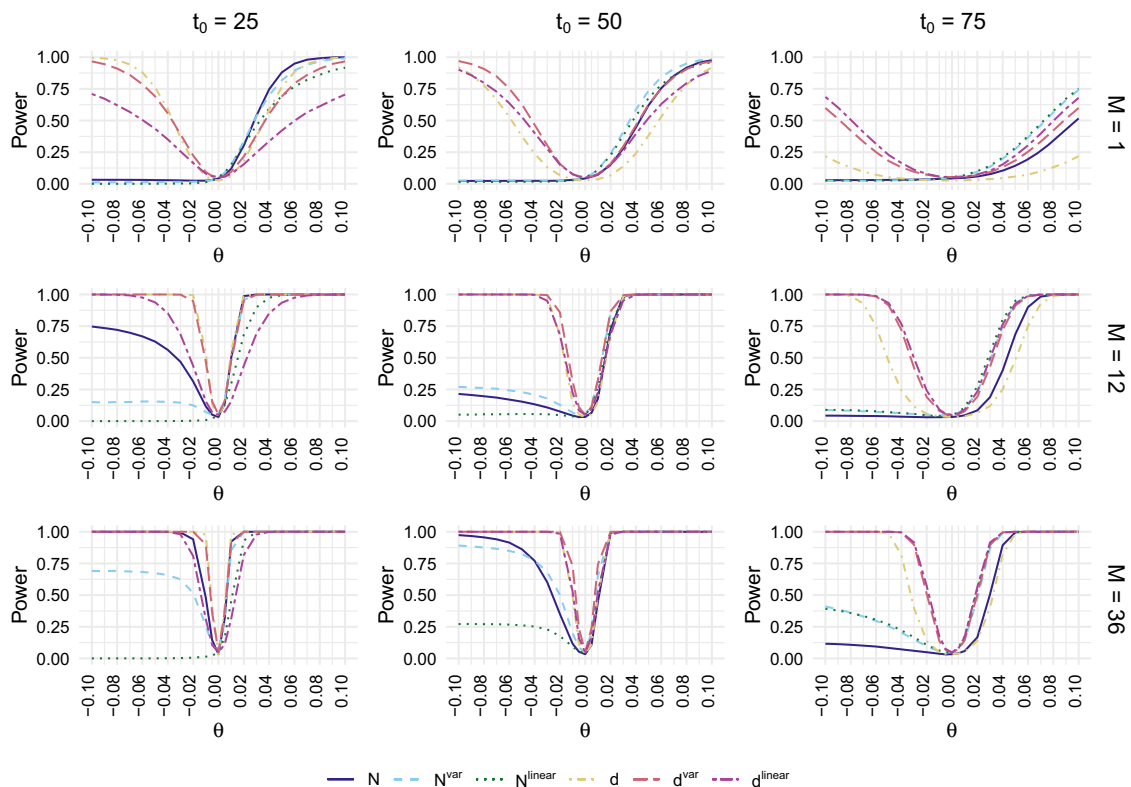
Fig. 1 Power functions of N and d-type statistics for Scenario A ($T = 100$)

statistic s or $\theta = 0.01$ with $M = 12$ or $\theta = 0.005$ with $M = 36$ for the statistics s and s^{var}. When $t_0 = 50$, we would detect $\theta = 0.05$ with $M = 1$ or $\theta = 0.01$ with $M = 12$ for the statistics s and s^{var}.

3.3 Analysis of changepoint estimation

The analysis of the changepoint estimation reports results for Scenarios A and B considered in Sect. 3.2 for $T = 100$, $M = 1$ and $\theta = 0.10$, and for $T = 100$, $M = 36$ and $\theta = 0.05$, both for a wide range of changepoints $t_0 = 10, 20, \ldots, 80, 90$.

Figures 3 and 4 show boxplots of the estimated changepoint for the Scenarios A and B, respectively. We remark the following conclusions:

(1) As it was advanced in Sect. 2.1, the nonuniform variance in CUSUM-type statistics means that changepoints occurring near the data boundaries are more difficult to detect, hence, they have trouble in detecting changes occurring away from the middle of the series. This effect is reduced as the number of series M or the magnitude of the drift θ increases.

(2) Comparing statistics with the same type of weight. In Scenario A, N-type statistics place the changepoint slightly better than d-type. In Scenario B, s-type statistics place the chagepoint better than N-type.

(3) The performance of the changepoints depends on the type of weight used. The statistics without weights properly place the changepoint when it is at the beginning or the middle of the series, but not at the end. The statistics with weights
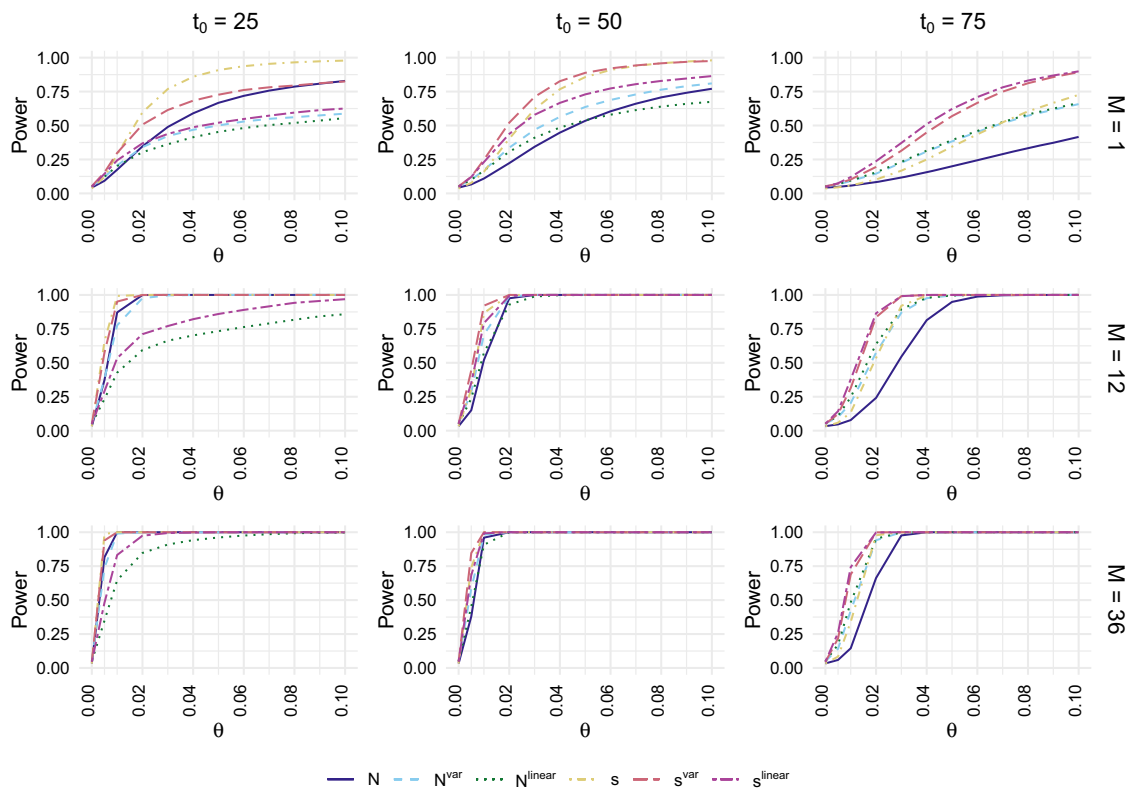
Fig. 2 Power functions of N and s-type statistics for Scenario B ($T = 100$)

proportional to the inverse of the SD properly place the changepoint when it is not found at the beginning or the end of the series. The statistics with linear weights place the majority of changepoints in the second half of the series, so its estimate is not reliable for practical use, although this effect is reduced by increasing M.

These changepoint detection tests based on the breaking of records only make use of the record occurrence to determine the changepoint estimate. For that reason, the estimated changepoint will usually be placed in the previous time of a record time, i.e., the effect of the drift is not immediately reflected in the observed record occurrence. This means that a proper estimate of the changepoint in the record occurrence will often be placed later in the series than the actual changepoint in the mean or variance. Thus, the main question here is then whether the correctly detected, but possibly displaced, changepoints are clustered near the actual value or not.

This section has been useful to illustrate the behavior of the tests against usual alternative hypotheses. Other scenarios could be considered, e.g., (C) a shift model in the mean, i.e., $\mu_t = \theta$ if $t_0 < t \leq T$ under Scenario A; or (D) a mixture model with a drift in the right tail, i.e., $Y_{tm} = \epsilon_{tm}^{(0)}$ if $u_{tm} \leq \tau$ and $Y_{tm} = \mu_t + \epsilon_{tm}^{(1)}$ if $u_{tm} > \tau$ where $u_{tm} \sim U(0, 1)$, τ a high quantile order (e.g., $\tau = 0.95$), μ_t under Scenario A, and $\epsilon_{tm}^{(0)}$ and $\epsilon_{tm}^{(1)}$ truncate $N(0, 1)$ in $(-\infty, \Phi^{-1}(\tau))$ and $(\Phi^{-1}(\tau), \infty)$, respectively. Preliminary analyzes show that the tests perform poorly against Scenario C, but have great power against Scenario D, even outperforming commonly used changepoint detection tests (e.g., the Pettitt test).
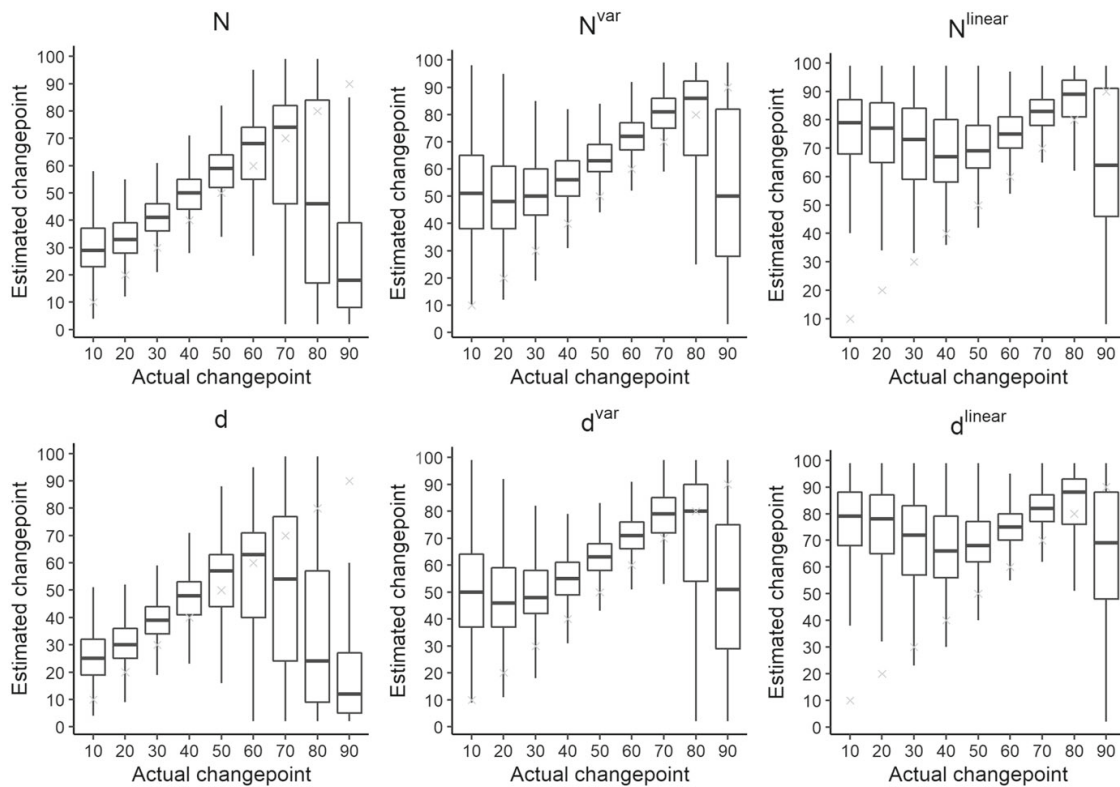
(a) $M = 1$, $\theta = 0.10$
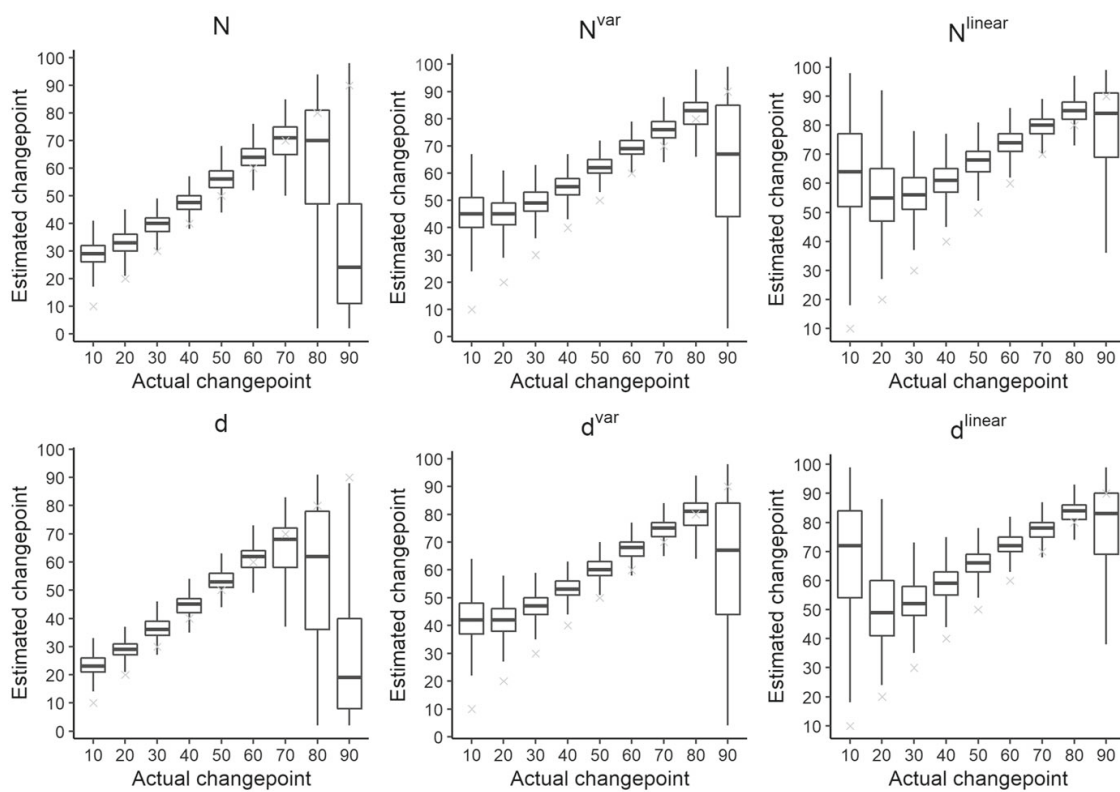
(b) $M = 36$, $\theta = 0.05$

Fig. 3 Boxplots without outliers of the estimated changepoint versus the actual changepoint of N and d-type statistics for Scenario A ($T = 100$). Crosses represent the actual changepoint

(a) $M = 1$, $\theta = 0.10$

(b) $M = 36$, $\theta = 0.05$

Fig. 4 Boxplots without outliers of the estimated changepoint versus the actual changepoint of N and s-type statistics for Scenario B ($T = 100$). Crosses represent the actual changepoint

4 Application to temperature series

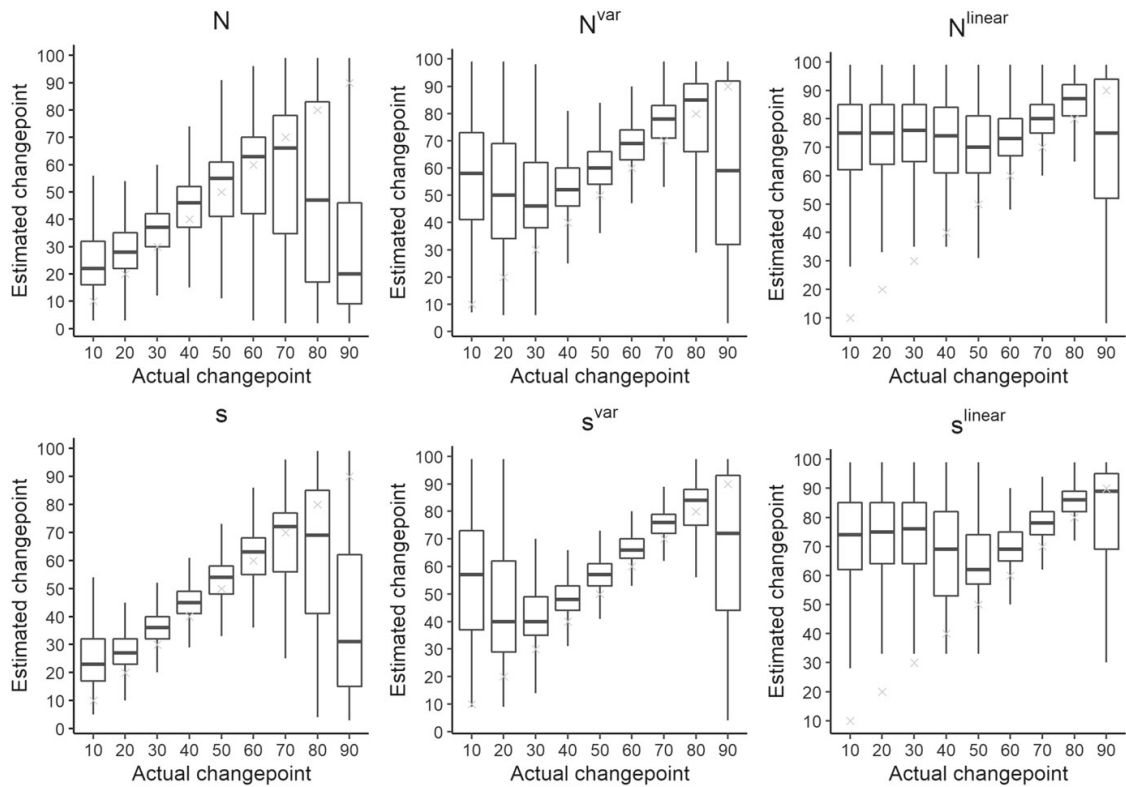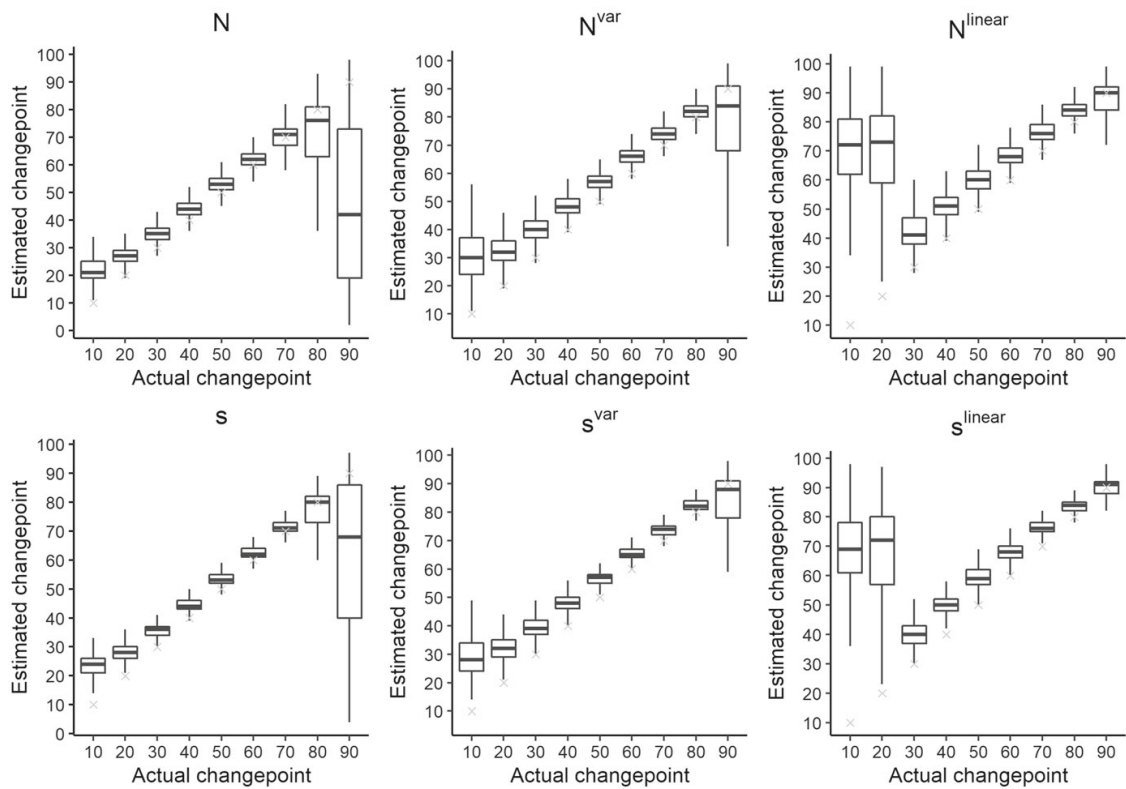To illustrate the practical use of the three types of records tests, we applied them to the daily maximum temperature series measured in degree Celsius (°C) from 1940 to 2019 at Madrid, Spain. Data are provided by European Climate Assessment & Dataset (ECA&D; Klein Tank et al. 2002) available online at https://www.ecad.eu. Madrid is located in the center of the Iberian Peninsula (40.4° N, 3.7° W) at 667 m a.s.l. and its daily temperature series has a seasonal component and a strong serial correlation. This series is analyzed using three different approaches to show the performance of the tests in different situations. The first approach considers the series of annual maximum temperatures, which corresponds to the traditional block maxima. The second approach considers the series of annual mean temperature. Finally, the series is considered on a daily scale. To do this, first we take 365 subseries each corresponding to the data of a given day across years and then we select a subset of uncorrelated subseries (Cebrián et al. 2022) on which the procedure of Sect. 2.3 is applied. The three approaches have series of length $T = 80$, the first two with $M = 1$ and the third with $M = 58$ uncorrelated series out of the 365 dependent subseries.

In the context of global warming, it is reasonable to assume an increasing trend in location that can cause an increase in the number of upper records as well as decrease the number of lower records with respect to the values expected under a stationary climate, i.e., IID series. For this reason, only results for N and d-type statistics are shown. These statistics are powerful against this scenario and obtained more evidence than s-type statistics. To compare the detection time of a changepoint in location versus a changepoint in the record occurrence, we consider the Pettitt (1979) test, which is a nonparametric rank based test widely used to detect AMOC at location.

Figure 5 shows time series plots of annual maximum **a** and annual mean **b** temperature at Madrid with their records and changepoint estimates. Table 2 shows for the two previous series and for the series in daily scale the p-values and changepoint estimates for the six records tests and the Pettitt test. Small p-values in the records tests provide evidence against the null hypothesis of stationarity, in particular, all tests are significant at a level $\alpha = 0.10$, all but one are significant at a level $\alpha = 0.05$ and fourteen out of eighteen are significant at $\alpha = 0.01$. The Pettitt test is also significant for any usual significance level in both series with $M = 1$. The estimated changepoint for the annual maximum temperature series is $\hat{t}_0 = 51$ (year 1990) for all the records statistics and $\hat{t}_0 = 38$ (1977) for the Pettitt test. The minimum p-value for the records tests is 0.0013 for the statistic N^{var}. The estimated changepoint for the annual mean temperature series is $\hat{t}_0 = 55$ (1994) with the statistics without weights and with weights proportional to the inverse of the SD, but it is $\hat{t}_0 = 69$ (2008) for the statistics with linear weights and $\hat{t}_0 = 41$ (1980) for the Pettitt test. The minimum p-value of the records tests is 0.0004 for the tests N^{var} and N^{linear}. For the daily scale series the changepoint estimate is $\hat{t}_0 = 38$ (1977) for all records statistics and here the minimum p-value is $4e$-05 for d^{var}.

The results in Table 2 agree with the results obtained in Sect. 3. When $M = 1$, N-type statistics obtain lower p-values than d-type, and the statistics with weights proportional to the inverse of the SD are those that obtained the strongest evidence. The changepoint estimate of the records tests is usually placed between 10 and 15 years
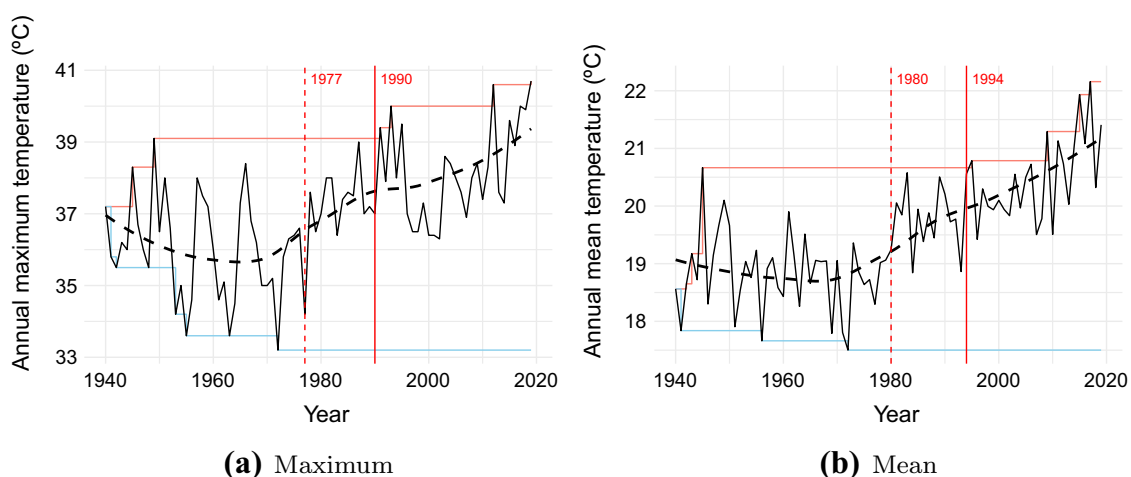
Fig. 5 Annual maximum (**a**) and mean (**b**) temperature series, and lower and upper records at Madrid, Spain. The vertical solid line is the estimated changepoint using the records tests, while the vertical dashed line is the estimated changepoint using the Pettitt test

Table 2 Estimated changepoint $\hat{t}_0$ and p-value of the records tests and the Pettit test for the annual maximum, annual mean and daily scale temperature series at Madrid, Spain

Statistic	Series					
	Annual maximum		Annual mean		Daily	
	$\hat{t}_0$	p-value	$\hat{t}_0$	p-value	$\hat{t}_0$	p-value
N	51	0.0031	55	0.0030	38	0.0003
N^{var}	51	0.0013	55	0.0004	38	0.0002
N^{linear}	51	0.0090	69	0.0004	38	0.0328
d	51	0.0443	55	0.0728	38	0.0013
d^{var}	51	0.0016	55	0.0015	38	$4e$-05
d^{linear}	51	0.0115	69	0.0018	38	0.0029
Pettitt	38	$9e$-07	41	$8e$-10	–	–

The weighted statistics used 1,000,000 replicates to estimate the p-value

after the Pettitt test estimates a changepoint in location. The changepoint estimated by the statistics with linear weights tend to locate the change very late. It is noteworthy that the changepoint is always estimated just before a record (see Fig. 5), so the changepoint estimate of a significant records test can be interpreted as the time from which there is evidence that the record occurrence is no longer stationary and the tail of the distribution begins to take on ever greater values, not previously seen. When $M > 1$ the results are more stable, the estimated changepoint appears earlier as more information is available and d-type statistics obtain smaller p-values than N-type.

Figure 6 plots the year versus the absolute value of the processes associated with the records statistics for the annual maximum **a** and mean **b** temperature series along with 95% confidence thresholds based on the Kolmogorov distribution (they are very similar even for nonKolmogorov distributed statistics). These plots allow to see the evolution of the processes and other possible points with greater record probability than under the null hypothesis. Again, the stationary null hypothesis is rejected, indicating

(a) Annual maximum

(b) Annual mean
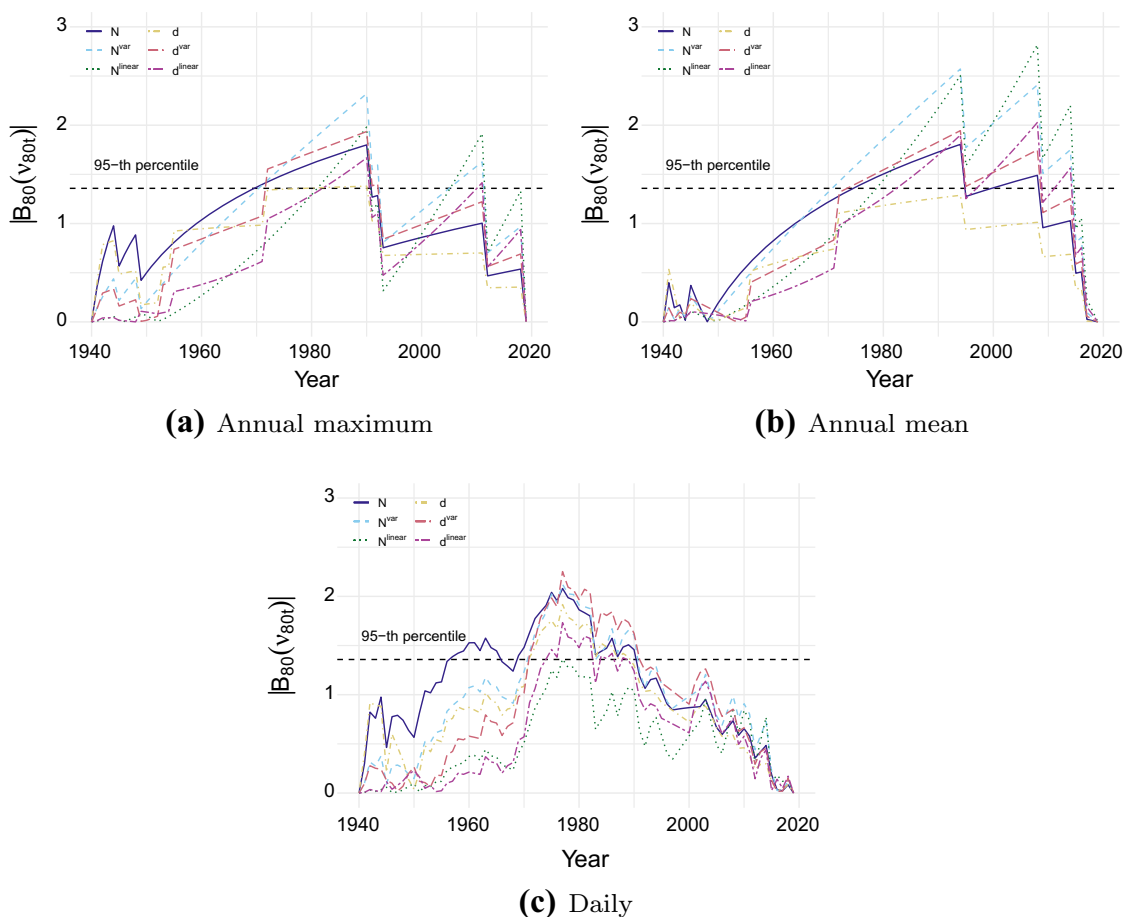
(c) Daily

Fig. 6 Absolute value of the processes associated with the records statistics for changepoints of annual maximum (**a**), annual mean (**b**) and daily (**c**) temperature series at Madrid, Spain. The horizontal dashed line represents the 95th percentile of the Kolmogorov distribution

potential changepoint 1990 and 1994, respectively. The equivalent plot for the daily scale temperature series is shown in **c**, showing a clear maximum in 1977.

5 Discussion, conclusions and future work

The interest in statistical tools to analyze nonstationary behaviors in the extreme values of the distribution is growing. While extreme value analysis has been traditionally based on block maxima and excesses over threshold, this paper proposes the use of records to study changes in the tails of the distribution. In particular, this paper proposes three novel distribution-free changepoint detection tests and some generalizations based on the breaking of records to (1) detect changes in the extreme events of the distribution, (2) learn about features of the record occurrence and (3) analyze data when only their records are available.

The proposed statistics are CUSUM-type statistics based on the record indicators. Statistics to deal with seasonal series have also been considered. Despite having a very small sample information compared to the total of the series, the Monte Carlo simulations have shown that the proposed records tests are capable of detecting deviations from the null hypothesis and a reasonable changepoint at which this deviation

becomes significant in the probabilities of record. However, care must be taken in the interpretation of the changepoint estimate, on the one hand it is usually misplaced when the actual changepoint is located in the ends of the series. On the other hand, when it is well located, it is often slightly after the actual changepoint in location or scale if it exists, i.e., the effect of the change is not immediately reflected in the observed record occurrence.

The recommendation for use according to the power and changepoint estimate accuracy of the tests is as follows. If an increase in the number of records with respect to the stationary case is expected in a single tail of the distribution, the results show that N-type statistics are usually recommended. If an increase in the number of records is expected in both tails, then s-type statistics are preferred. The statistics without weights have the advantage of having a known asymptotic distribution, while the statistics with weights proportional to the inverse of the SD have shown to have a more balanced behavior against the alternative hypothesis in the simulation results, with the disadvantage that their distribution must be calculated using Monte Carlo techniques.

The proposed tests join two important aspects in the study of climate change, changepoint detection methods and record-breaking events. This last concern has been made apparent when applying the tests on different summary series (block maxima and annual mean) and on the series on a daily scale of temperatures at Madrid, Spain; detecting significant evidence of warming since the late 1970s and early 1990s.

Future work may go in different directions. (1) Combining the information from the different statistics could be of interest to increase the power and decrease the chance of mis-detection, e.g., the harmonic mean p-value by Wilson (2019) could be used to have a single p-value of all tests. (2) The idea of splitting the series is fundamental to dealing with seasonal behavior. Here we use the method by Cebrián et al. (2022) to extract uncorrelated subseries. Another alternative to consider would be to implement permutation tests, i.e., the test statistic under the null hypothesis would be obtained by calculating all possible values of the test statistic under all possible rearrangements of the observed years, $t = 1, \ldots, T$. In this way we maintain the dependence structure between the subseries without the need to have a subset of independent subseries. (3) Our method has been developed within the AMOC domain, however its extension to the multiple changepoint domain could be of interest. The simplest procedure would be to split the series where the changepoint is detected and retest the two subseries separately. However, this can cause the number of records in the new subseries to be too small to detect new changepoints, so other alternatives should be studied.

Finally, it is noteworthy that the proposed changepoint detection records tests are not only useful for analyzing the effect of global warming on the occurrence of records, but also in other fields where records are important. Other applications of these tests are in other environmental sciences in the presence of climate change, in the study of extreme values in stock prices or in the influence that new sports equipment has on the occurrence of sports records. To facilitate its use, all the statistical tools proposed in this paper are included in the R package `RecordTest` (Castillo-Mateo 2021) available from CRAN at https://CRAN.R-project.org/package=RecordTest.

Declarations

Conflict of interest The author declares that he has no conflict of interest.

Ethical approval This work does not contain any studies with human participants and/or animals.

A Appendix: The variance of $B_T(\nu_{Tt})$

Figure 7 shows the variance of $B_T(\nu_{Tt})$ for $T = 100$ and 1000 across $t = 1, \ldots, T$. In particular, it is shown for the unweighted statistic, the statistics with weights proportional to the inverse of the SD and linear weights (see Sect. 2.2). While the second one generates a symmetric variance in $\{1, \ldots, T\}$ by construction, this does not happen with the other two. In all three cases the variance is zero for $t \in \{1, T\}$ and the maximum value is $1/4$.

The nonuniform variance makes changepoints occurring near the beginning or the end of the series (small variance times) more difficult to detect. Under the null hypothesis, the process reaches its maximum (in absolute value) with the highest probability at time t where it has the highest variance. Then, deviations from the null hypothesis at small variance times generate smaller deviations in $B_T(\nu_{Tt})$ than deviations at times of maximum variance. Thus, it is expected that the unweighted statistic will have more power when the changepoint is at the beginning of the series, the statistic with weights proportional to the inverse of the SD when the changepoint is in the middle of the series and the statistic with linear weights when the changepoint is at the end of the series. These conclusions agree with the analysis of power in Sect. 3.2.
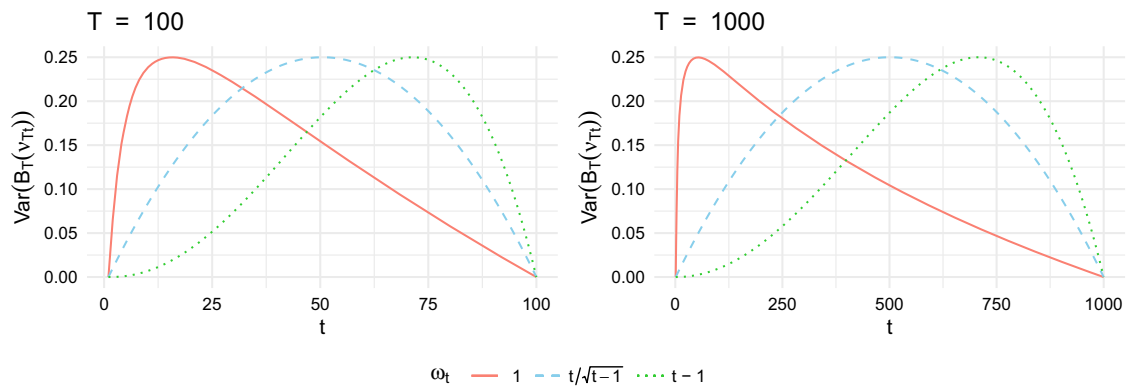
Fig. 7 The $Var(B_T(v_{Tt})) = v_{Tt}(1 - v_{Tt})$ across $t = 1, \ldots, T$ ($T = 100, 1000$) and weights ω_t given in Sect. 2.2

B Appendix: Proof of Proposition 2.1

We prove that the weighted statistics with polynomial weights $\omega_t \sim t^n$ as $t \to \infty$ for $n > 0$ do not have asymptotic Gaussian properties. In particular, the distribution of the weighted number of records do not approach the normal distribution for increasing T. This is verified by using its asymptotic skewness and showing that it is different from 0 (the skewness of any normal RV) for $n > 0$. As a consequence, the asymptotic distribution of the functional evolution of the weighted number of records does not approach that of the Wiener process.

Proof (of Proposition 2.1) To prove that the ξ_{Tt}^{ω}'s do not satisfy the central limit theorem, it is sufficient to prove that the sum $N_T^{\omega} = \sum_{t=1}^{T} \omega_t I_t$ does not have skewness 0 as $T \to \infty$. Using the basic properties of the central moments of a RV,

$$\mu \equiv \mu_1(N_T^{\omega}) = E(N_T^{\omega}) = \sum_{t=1}^{T} \omega_t \frac{1}{t},$$

$$\sigma^2 \equiv \mu_2(N_T^{\omega}) = E\left[(N_T^{\omega} - \mu)^2\right] = \sum_{t=1}^{T} \omega_t^2 \frac{t-1}{t^2},$$

$$\mu_3(N_T^{\omega}) = E\left[(N_T^{\omega} - \mu)^3\right] = \sum_{t=1}^{T} \omega_t^3 \frac{t^2 - 3t + 2}{t^3}.$$

Then, the following is a consequence of the properties of the generalized harmonic numbers, as $T \to \infty$,

$$Skew(N_T^{\omega}) = \frac{\mu_3(N_T^{\omega})}{\sigma^3} \sim \frac{\sum_{t=1}^{T} t^{3n-1}}{\left(\sum_{t=1}^{T} t^{2n-1}\right)^{3/2}} \longrightarrow \frac{2}{3}\sqrt{2n}.$$

Consequently the skewness of N_T^{ω} is asymptotically different from 0 for $n > 0$. $\square$

This proposition is easily extended to the weighted statistics based on the d_t's and s_t's. The former requires the calculation of the kurtosis since its skewness is 0 because it is a symmetric RV. We omit the details for the sake of brevity.

References

Achcar JA, Rodrigues ER, Paulino CD, Soares P (2010) Non-homogeneous Poisson models with a change-point: an application to ozone peaks in Mexico City. Environ Ecol Stat 17:303–322. https://doi.org/10.1007/s10651-009-0114-3

Achcar JA, Coelho-Barros EA, de Souza RM (2016) Use of non-homogeneous Poisson process (NHPP) in presence of change-points to analyze drought periods: a case study in Brazil. Environ Ecol Stat 23:405–419. https://doi.org/10.1007/s10651-016-0345-z

Arnold BC, Balakrishnan N, Nagaraja HN (1998) Records. Wiley series in probability and statistics. Wiley, New York. https://doi.org/10.1002/9781118150412

Benestad RE (2004) Record-values, nonstationarity tests and extreme value distributions. Glob Planet Change 44(1–4):11–26. https://doi.org/10.1016/j.gloplacha.2004.06.002

Brodsky E, Darkhovsky BS (1993) Nonparametric methods in change point problems. Springer, Dordrecht. https://doi.org/10.1007/978-94-015-8163-9

Castillo-Mateo J (2021) RecordTest: inference tools in time series based on record statistics. R package version 2.1.0. https://CRAN.R-project.org/package=RecordTest

Cebrián AC, Castillo-Mateo J, Asín J (2022) Record tests to detect non-stationarity in the tails with an application to climate change. Stoch Environ Res Risk Assess 36(2):313–330. https://doi.org/10.1007/s00477-021-02122-w

Coumou D, Robinson A, Rahmstorf S (2013) Global increase in record-breaking monthly-mean temperatures. Clim Change 118(3–4):771–782. https://doi.org/10.1007/s10584-012-0668-1

Csörgő M, Horváth L (1997) Limit theorems in change-point analysis. Wiley, Chichester

Dierckx G, Teugels JL (2010) Change point analysis of extreme values. Environmetrics 21(7–8):661–686. https://doi.org/10.1002/env.1041

Diersen J, Trenkler G (1996) Records tests for trend in location. Statistics 28(1):1–12. https://doi.org/10.1080/02331889708802543

e Silva WVM, do Nascimento FF, Bourguignon M (2020) A change-point model for the r-largest order statistics with applications to environmental and financial data. Appl Math Model 82:666–679. https://doi.org/10.1016/j.apm.2020.01.064

Fisher TJ, Robbins MW (2019) A cheap trick to improve the power of a conservative hypothesis test. Am Stat 73(3):232–242. https://doi.org/10.1080/00031305.2017.1395364

Foster FG, Stuart A (1954) Distribution-free tests in time-series based on the breaking of records. J R Stat Soc B 16(1):1–22

Gikhman II, Skorokhod AV (1969) Introduction to the theory of random processes. Saunders, Philadelphia

Hirsch RM, Slack JR, Smith RA (1982) Techniques of trend analysis for monthly water quality data. Water Resour Res 18(1):107–121. https://doi.org/10.1029/WR018i001p00107

Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Mileta M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Forland E, Mietus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnar T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander LV, Petrovic P (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int J Climatol 22(12):1441–1453. https://doi.org/10.1002/joc.773

Kojadinovic I, Naveau P (2017) Detecting distributional changes in samples of independent block maxima using probability weighted moments. Extremes 20(2):417–450. https://doi.org/10.1007/s10687-016-0273-1

Lehmann J, Coumou D, Frieler K (2015) Increased record-breaking precipitation events under global warming. Clim Change 132(4):501–515. https://doi.org/10.1007/s10584-015-1434-y

Nevzorov V (2001) Records: mathematical theory. American Mathematical Society, Rhode Island

Page ES (1954) Continuous inspection schemes. Biometrika 41(1–2):100–115. https://doi.org/10.1093/biomet/41.1-2.100

Page ES (1955) A test for a change in a parameter occurring at an unknown point. Biometrika 42(3–4):523–527. https://doi.org/10.1093/biomet/42.3-4.523

Pettitt AN (1979) A non-parametric approach to the change-point problem. J R Stat Soc C 28(2):126–135. https://doi.org/10.2307/2346729

R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Ratnasingam S, Ning W (2021) Modified information criterion for regular change point models based on confidence distribution. Environ Ecol Stat 28:303–322. https://doi.org/10.1007/s10651-021-00485-5

Reeves J, Chen J, Wang XL, Lund R, Lu QQ (2007) A review and comparison of changepoint detection techniques for climate data. J Appl Meteorol Climatol 46(6):900–915. https://doi.org/10.1175/JAM2493.1

Rodrigues ER, Nicholls G, Tarumoto MH, Tzintzun G (2019) Using a non-homogeneous Poisson model with spatial anisotropy and change-points to study air pollution data. Environ Ecol Stat 26:153–184. https://doi.org/10.1007/s10651-019-00423-6

Wergen G (2013) Records in stochastic processes—theory and applications. J Phys A 46(22):223001. https://doi.org/10.1088/1751-8113/46/22/223001

Wilson DJ (2019) The harmonic mean p-value for combining dependent tests. Proc Natl Acad Sci USA 116(4):1195–1200. https://doi.org/10.1073/pnas.1814092116

3.7 RecordTest: An R package to analyse non-stationarity in the extremes based on record-breaking events.

This manuscript was published in:

Castillo-Mateo, J., Cebrián, A. C., & Asín, J. (2023). **RecordTest**: An R package to analyse non-stationarity in the extremes based on record-breaking events. *Journal of Statistical Software, 106* (5), 1–28. https://doi.org/10.18637/jss.v106.i05

And it was disseminated (speaker emphasized) in:

- ***Castillo-Mateo, J.***, Cebrián, A. C., & Asín, J. (2022, November 23–25). ***RecordTest****: Un paquete de R para detectar comportamientos no estacionarios en la ocurrencia de eventos récord (****RecordTest****: An R package to detect non-stationary behavior in the occurrence of record events)* [Contributed talk]. I Congreso & XII Jornadas de Usuarios de R, Córdoba, Spain. *Best Oral Communication Award, Student Category.*
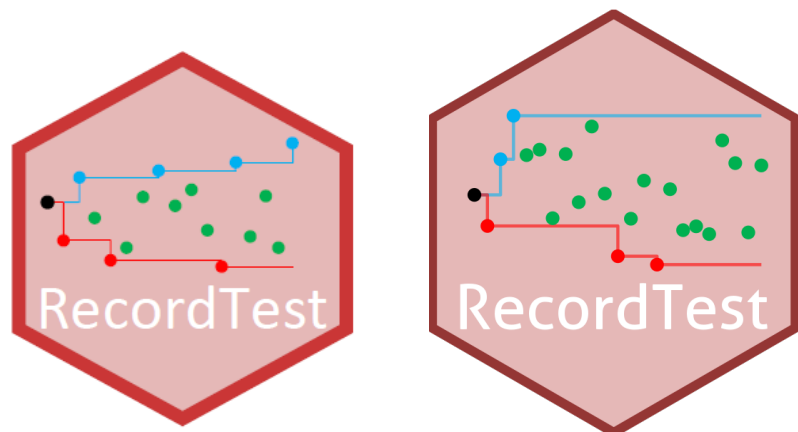
Figure 3.1: The logo of the R package **RecordTest** represents a time series with upper (blue) and lower (red) records according to `records()` in version 1.0.1. Left: first logo developed with Microsoft Word a few days after the end of the confinement due to the COVID-19 pandemic. Right: second logo developed with the R package **hexSticker** (G. Yu, 2020) for version 2.2.0

RecordTest: An **R** Package to Analyze Non-Stationarity in the Extremes Based on Record-Breaking Events

Jorge Castillo-Mateo
University of Zaragoza

Ana C. Cebrián
University of Zaragoza

Jesús Asín
University of Zaragoza

Abstract

The study of non-stationary behavior in the extremes is important to analyze data in environmental sciences, climate, finance, or sports. As an alternative to the classical extreme value theory, this analysis can be based on the study of record-breaking events. The R package **RecordTest** provides a useful framework for non-parametric analysis of non-stationary behavior in the extremes, based on the analysis of records. The underlying idea of all the non-parametric tools implemented in the package is to use the distribution of the record occurrence under series of independent and identically distributed continuous random variables, to analyze if the observed records are compatible with that behavior. Two families of tests are implemented. The first only requires the record times of the series, while the second includes more powerful tests that join the information from different types of records: upper and lower records in the forward and backward series. The package also offers functions that cover all the steps in this type of analysis such as data preparation, identification of the records, exploratory analysis, and complementary graphical tools. The applicability of the package is illustrated with the analysis of the effect of global warming on the extremes of the daily maximum temperature series in Zaragoza, Spain.

Keywords: extreme value analysis, hypothesis of stationarity, non-parametric tests, records, R.

1. Introduction

Time series data in many fields need to be examined for evidence of structural trends or shifts over time. In general, these analyses focus on the study of changes in the mean behavior, however changes in the extremes, i.e., in the tails of the distribution, are also of great interest. Extreme events typically cause large impacts because society and ecosystems are not adapted

to them, so that their study is essential. Examples of the importance of the analysis of the extremes appear in environmental sciences (large wildfires), climate (heat waves), hydrology (floods), finance (market risk), sports (limits of human capabilities), and many others.

The numerous studies about non-stationarity in the mean have been favored by the availability of easy-to-use software to compute non-parametric tests, e.g., the Mann-Kendall test (MK; Mann 1945; Kendall and Gibbons 1990). However, there does not exist similarly simple software to analyze trends, change-points, or non-stationary behavior in the extremes. Detection of this type of behaviors is complicated because extremes are rare by definition. Specific tools are required since it is difficult to link its evolution to the mean; i.e., if the magnitude of a trend in the mean is small in terms of the variability of the series, or if there are changes in the variability, the effect on the extremes might not be evident. Tools to analyze non-stationary behavior in the tails of the distribution are also required as validation tools in statistical modeling. Specific analysis of the capability to reproduce the extremes is essential in a validation analysis, since models that represent the entire distribution of a series tend to badly fit the tails and to yield important biases in extreme value statistics.

In this situation, the need for statistical tools to analyze the non-stationary behavior in the extremes and records of a series is clear. This is the aim of the R (R Core Team 2022) package **RecordTest** (Castillo-Mateo 2022b) described in this paper; the package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=RecordTest` and on GitHub at `https://github.com/JorgeCastilloMateo/RecordTest`. The package includes non-parametric tests and graphical tools based on records to detect non-stationary behavior, such as trends and change-points, in the extremes of a series. These tools can be applied to serially correlated series with seasonal behavior, if they are previously prepared applying an approach based on splitting the series. All the tools for the data preparation are also implemented. In addition, the inferential tools in the package are able to jointly analyze $M \geq 1$ series with possibly different distributions. This property is useful to analyze split series and also in spatial analysis, to study series from different locations and obtain global conclusions over the area of interest.

Classical methods for the analysis of extreme events are the block maxima and the excesses over threshold. Both of them require to fit the tails of the distribution using parametric models such as generalized extreme value (GEV) and Pareto (GP) distributions, and Poisson processes (Coles 2001). The tools in **RecordTest** are based on a different approach, the analysis of the occurrence of record events and its comparison with the behavior of records in the classical record model (CRM; Arnold, Balakrishnan, and Nagaraja 1998). The CRM describes the distribution of the records of a series $(X_1, \ldots, X_T)^\top$ of independent and identically distributed (IID) continuous random variables (RVs). An important advantage of this approach is that it yields non-parametric tools due to the probabilistic properties of records. In particular, the fact that the distribution of the record occurrence under the CRM does not depend on the underlying distribution of the X_t's allows the use of distribution-free statistics and Monte Carlo methods. **RecordTest** includes the tests of this type proposed by Foster and Stuart (1954), Diersen and Trenkler (1996), Benestad (2003, 2004), Cebrián, Castillo-Mateo, and Asín (2022), Castillo-Mateo (2022a), and some extensions thereof proposed in this paper. The package also includes useful graphical tools based on the behavior of the record occurrence under the CRM. We found in Cebrián *et al.* (2022) that the power of the records tests is high, e.g., it is between 0.80 and almost 1.00 for a sample size of $M = 12$ series of length $T = 50$ and an alternative with a linear trend in the mean which has a magnitude of about

2.5% of the standard deviation. The MK test for the mean is more powerful when the series follow a normal distribution with a linear trend in the mean. However, records tests, which focus on the tails of the distribution, tend to be more powerful than the MK test in series with one or two-side bounded distributions or distributions with one or two light tails, such as GEV and GP, often used in extreme value analysis.

Many questions of interest in the analysis of non-stationary behavior in the tails are directly related to records, and the tools in **RecordTest** are specifically useful for this type of analysis. The study of records is common in sports (Gembris, Taylor, and Suter 2002, 2007), but it is also essential in environmental sciences, for the study of floods (Vogel, Zafirakou-Koulouris, and Matalas 2001), earthquakes (Van Aalsburg, Newman, Turcotte, and Rundle 2010; Yoder, Turcotte, and Rundle 2010), or avalanches (Shcherbakov, Davidsen, and Tiampo 2013). In the context of climate change, an important question is the effect of global warming on the number of record-breaking temperatures and precipitation events (Benestad 2003, 2004; Coumou, Robinson, and Rahmstorf 2013; Wergen and Krug 2010; Lehmann, Coumou, and Frieler 2015; Lehmann, Mempel, and Coumou 2018). The study of records is also of interest in physics, in the theory of spin-glasses or high temperature superconductors, in evolutionary biology, or in finances for the study of stock prices (see Wergen 2013, and references therein).

There are many packages for analyzing the existence of trends and non-stationary behavior using non-parametric tests, but most of them focus on the analysis of the mean of the distribution. For example, in R, **Kendall** (McLeod 2022) computes the MK test, and **modifiedmk** (Patakamuri and O'Brien 2021) implements modified versions of trend tests for serially correlated data. **trend** (Pohlert 2020) includes a great variety of tests such as Cox-Stuart, (seasonal) MK and Hirsch-Slack tests for trend detection, and Lanzante, Pettitt and Buishand tests for change-point detection. **pyMannKendall** (Hussain and Mahmud 2019) is a Python (Van Rossum *et al.* 2011) implementation of non-parametric MK trend analysis, which brings together eleven types of tests. The R package **npcp** (Kojadinovic 2023) provides non-parametric CUSUM change-point detection tests sensitive to changes in the mean, the variance, the covariance, or the autocovariance in univariate or multivariate observations, as well as a test for detecting changes in the distribution of independent block maxima. Focusing on the analysis of extremes, there are quite a few R packages, such as **evir** (Pfaff and McNeil 2018), which even includes the function `evir::records()` for extracting records. Some of them include relevant tools for testing and modeling non-stationarity. For example, **extRemes** (Gilleland and Katz 2016) and **ismev** (Heffernan and Stephenson 2018) include non-stationary models for univariate extremes, **evd** (Stephenson 2002) has some functionality for non-stationary estimation but the main emphasis is on bivariate extremes, **SpatialExtremes** (Ribatet 2022) and **texmex** (Southworth, Heffernan, and Metcalfe 2020) analyze a multivariate framework, and **NHPoisson** (Cebrián, Abaurrea, and Asín 2015) fits non-homogeneous Poisson processes for peak over threshold analysis. All these packages offer a parametric approach based on the fit of GEV, GP, and Poisson processes. A similar approach is used in the Python package **pyextremes** (Bocharov 2022), which includes methods such as block maxima, peaks over threshold and fitting of GEV and GP distributions; and in the MAT-LAB (The MathWorks Inc. 2022) package **NEVA** (Cheng, AghaKouchak, Gilleland, and Katz 2014), which allows the fitting of both stationary and non-stationary GEV and GP distributions in a Bayesian framework. However, as far as we know, there does not exist any statistical software package for testing a non-stationary behavior in records. The aim of **RecordTest** is to fill this gap and provide a comprehensive toolkit to assess significant deviations from

a stationary behavior in the tails of the distribution and to characterize when it occurs and which features are affected. The study of records provides a new approach from a different point of view than the block maxima and excesses over threshold approaches. Beyond the intrinsic interest of the records, there are some differences that make this approach useful for other types of analysis. An advantage of the inference tools in **RecordTest** is that no previous modeling is needed. In addition, the study of records allows the simultaneous analysis of the lower and upper tails of the distribution by including both upper and lower records in the analysis.

The outline of the paper is as follows. Section 2 introduces some basic definitions and properties of the main variables related to the record occurrence. Section 3 describes the functions and capabilities of the package, including data preparation, exploratory analysis, statistical tests and graphical tools. Section 4 illustrates the use of **RecordTest** to analyze the non-stationary behavior in the tails of the daily maximum temperature series in Zaragoza, Spain. A summary of the paper and some future work are given in Section 5.

2. Classical record model and deviations from stationarity

The statistical tools for detecting non-stationarity in the extremes implemented in **RecordTest** are based on the properties of the record occurrence in series of IID continuous RVs, i.e., the CRM. This section reviews some basic concepts and the probabilistic results that are the basis of those tools.

2.1. Variables to characterize the record occurrence

Let $(X_1, \ldots, X_T)^\top$ be a series of RVs. An observation X_t is called an upper record (or simply a record) if its value exceeds that of all previous observations, i.e., if $X_t > \max\{X_1, \ldots, X_{t-1}\}$. By virtue of this definition, X_1 is always a trivial record. Analogously, X_t is a lower record if $X_t < \min\{X_1, \ldots, X_{t-1}\}$. Let $(I_1, \ldots, I_T)^\top$ be the series of record indicator RVs defined by

$$I_t = \begin{cases} 1 & \text{if } X_t \text{ is a record,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then, the number of records is defined by the record counting process, $(N_1, \ldots, N_T)^\top$, where

$$N_t = I_1 + I_2 + \cdots + I_t, \tag{2}$$

and subsequently, the series of record times, $(L_1, \ldots, L_{N_T})^\top$, is defined by

$$L_i = \min\{t \mid N_t = i\}. \tag{3}$$

Finally, although they are not directly related to the occurrence, the series $(R_1, \ldots, R_{N_T})^\top$ of record values is defined by $R_i = X_{L_i}$.

All the tools implemented in this package assume that there are M independent series of length T available, i.e., there is a sequence $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ of independent series where $\mathbf{X}_m = (X_{1m}, \ldots, X_{Tm})^\top$ for $m = 1, \ldots, M$. However, most of the tools can be applied even with $M = 1$. The M series can be the result of splitting the original series, or series measured at different spatial points, for example. Given $\mathbf{X}$, we define the sequences of record

indicators, $\mathbf{I} = (\mathbf{I}_1, \ldots, \mathbf{I}_M)$, and the sequences of the number of records, $\mathbf{N} = (\mathbf{N}_1, \ldots, \mathbf{N}_M)$, where $\mathbf{I}_m = (I_{1m}, \ldots, I_{Tm})^\top$ and $\mathbf{N}_m = (N_{1m}, \ldots, N_{Tm})^\top$ for $m = 1, \ldots, M$. In a similar way, obvious definitions deal with record times and record values, $\mathbf{L}$ and $\mathbf{R}$, respectively.

2.2. Probabilistic properties of the record occurrence

Arnold *et al.* (1998) present the essential topics related to the theory of records. An important result states that, in the CRM, the series $(I_1, \ldots, I_T)^\top$ consists of mutually independent RVs with *Bernoulli*(p_t) distribution where p_t, the probability of observing a new record at time t, is

$$p_t = \mathsf{P}(I_t = 1) = \frac{1}{t}, \qquad t = 1, \ldots, T. \tag{4}$$

As a consequence, the distribution of I_t, N_t, and L_i does not depend on the common continuous distribution of the X_t's. This property allows the definition of the non-parametric statistical tests and graphical tools available in **RecordTest**.

Concerning the number of records, under the CRM, N_T converges in distribution as $T \to \infty$ to a normal distribution with mean and variance

$$\mathsf{E}(N_T) = \sum_{t=1}^{T} \frac{1}{t} \qquad \text{and} \qquad \mathsf{VAR}(N_T) = \sum_{t=2}^{T} \frac{1}{t}\left(1 - \frac{1}{t}\right). \tag{5}$$

These expressions are obtained from the fact that N_T is a sum of independent Bernoulli RVs. To give some intuition about the model, note that $\mathsf{E}(N_T) \approx \log T + \gamma$ and $\mathsf{VAR}(N_T) \approx \log T + \gamma - \pi^2/6$ where γ is the Euler constant, and both expressions tend to infinity. However, under the CRM, records are not common and their occurrence becomes scarcer for larger values of T.

Turning to the notation with M independent series, if we assume that the M series have the same probabilities of record, i.e., $p_{tm} \equiv p_t$, the maximum likelihood estimates (MLEs) of these probabilities are

$$\hat{p}_t = \frac{I_{t1} + \cdots + I_{tM}}{M}, \qquad t = 1, \ldots, T, \tag{6}$$

where the sum of record indicators above follows an exact binomial distribution with M trials and probabilities of success $p_t = 1/t$. If $M = 1$, the variability associated with the estimates is large. As the number of series M increases, the estimates become more accurate and precise.

2.3. Analysis of non-stationarity in the tails of the distribution

The aim of all the inference tools in **RecordTest** is to detect a non-stationary behavior in the occurrence of records in a time series, and more generally in the upper (and lower) tail of the distribution of the series. When we refer to a non-stationary behavior we mean any deviation from the CRM in the generating system of records. The underlying idea in all the tools is to study if the occurrence of observed records is compatible with the expected behavior of the occurrence of records under the CRM, i.e., in a series of IID continuous RVs. Under the assumption that the RVs in the series are independent, any deviation from the expected behavior of records will give evidence of a change over time in the distribution, i.e., non-stationarity. In many real problems, a non-stationary behavior in a non-seasonal series is due to the existence of any type of trend.

The tests in the package are based on record probabilities. Since these probabilities are known under the CRM, the null hypothesis of all the tests is

$$H_0 : p_{tm} = 1/t, \qquad \text{for all } t = 1, \ldots, T, \text{ and } m = 1, \ldots, M, \tag{7}$$

with $p_{tm} = \mathsf{P}(I_{tm} = 1)$. Different alternative hypotheses, one-sided, two-sided or the existence of a change-point, can be of interest, and tests for each of them are proposed in **RecordTest**. The one-sided alternative claims that the probabilities of record are either greater or less than the values given by the null hypothesis. This increase or decrease may be originated by the existence of a positive or a negative trend in location, or by an increase or decrease of variability.

Two different families of tools are implemented in the package, the first only requires to know the record times of the series, while the second requires to have the entire series available. The idea of the second family, first suggested by Foster and Stuart (1954), is that more powerful tests are obtained by joining the information from different types of records instead of using only one type. More precisely, from one series $(X_1, \ldots, X_T)^\top$, four different types of records can be computed: the upper and lower records in the forward and in the backward series (or directions). The backward series $(X_T, \ldots, X_1)^\top$ is obtained by reversing the order of the terms. For example, the upper record indicators in the backward series are

$$I_t^{(BU)} = \begin{cases} 1 & \text{if } X_{T-t+1} > \max\{X_T, \ldots, X_{T-t+2}\}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

To distinguish what type of records a statistic or RV refers to, the corresponding superscripts F (forward), B (backward), U (upper), or L (lower) are added between brackets. Given the symmetry of the problem, under the CRM, the probability of record p_t is $1/t$ for the four types of records.

3. Functions and capabilities

RecordTest provides a framework for the analysis of non-stationary behavior in the extremes of a series using records. It covers all the steps in the analysis: data preparation, identification of the records, exploratory analysis, a wide range of statistical tests, and complementary graphical tools. This section describes the functions grouped according to their objective.

3.1. Data preparation and record variables

The main argument of the inference functions in the package is a vector $(X_1, \ldots, X_T)^\top$ or a matrix $\mathbf{X}$. If only the record times are available, they have to be transformed into a series $(X_1, \ldots, X_T)^\top$ with those record times. This transformation is implemented by the function `series_record()`, whose arguments are the record times, `L_upper` or `L_lower`, and optionally the record values, `R_upper` or `R_lower`. Note that inference based on this new series only makes sense for the tools that use the types of records that are introduced as an argument: upper, lower, or both.

All the functions allow missing values represented by `NA`. The way to deal with this is to replace them by `-Inf` for upper records and `Inf` for lower records, so they are records only if they appear at $t = 1$.

Split series

In many real problems, the original series has to be split into M subseries, for instance, to remove the seasonal behavior (Hirsch, Slack, and Smith 1982). As an example, if a daily series with annual seasonality, $(X_{1,1}, X_{1,2}, \ldots, X_{1,365}, X_{2,1}, X_{2,2}, \ldots, X_{T,365})^\top$ where X_{tm} is the variable on day m within year t, is split into 365 series, one for each day within year,

$$
\begin{pmatrix}
X_{1,1} & X_{1,2} & \cdots & X_{1,365} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,365} \\
\vdots & \vdots & & \vdots \\
X_{T,1} & X_{T,2} & \cdots & X_{T,365}
\end{pmatrix}_{T \times 365} ,
\tag{9}
$$

the resulting subseries (columns) do not show seasonal behavior. In addition, since the consecutive observations in each subseries are now separated by one year, the serial correlation can be assumed to be zero. The distribution of the 365 series, which correspond to series at different calendar days, will not be the same due to the seasonal effect. However, the T RVs in each series, which correspond to variables measured at the same calendar day across years, may be identically distributed. Note that the null hypothesis H_0 of the inference tools in **RecordTest** is that each of the M series available are sequences of IID continuous RVs, but no assumption about the equal distribution of the M series is required. This functionality is implemented by `series_split()` that splits the series in argument `X` into `Mcols` subseries and arrange them as the matrix in (9).

Uncorrelated series

All the statistical methods implemented in the package assume that the M series under study are independent. If we have a set of dependent series, we should extract a subset of independent series from them before applying the inference tools. The function `series_uncor()` extracts a subset of uncorrelated series from the set available. This function has the arguments `test.fun`, a function to implement the desired correlation or dependence test, and `alpha`, that establishes the significance level. The default function is the standard `stats::cor.test()` with a significance level $\alpha = 0.05$, i.e., two series are considered uncorrelated if the Pearson correlation between them is not significantly different from zero at that significance level. Although zero correlation does not imply independence, this is a usual approach in most real data problems because dependence manifests itself as some level of linear correlation and testing dependence is not possible in general. However, more sophisticated functions could be used for testing dependence in other particular situations. For example, `extRemes::taildep.test()` or `evd::evind.test()` could be considered to test dependence at the extremes of the series.

We explain the algorithm with `test.fun = stats::cor.test` as an illustration. The iterative procedure to be used for selection is specified by the argument `type`. If the series have a sequential order, for example they are measured in consecutive days, the argument `type = "adjacent"` should be used and the following approach is applied: given that the kth series is in the subset, correlation between series k and $k + 1$ is tested; if the correlation is not significant, series $k + 1$ is included in the subset, otherwise correlation between series k and $k+2$ is tested. This step is repeated until a series $k+j$ that is not significantly correlated with series k is found. This approach does not test the pairwise correlation of all the series in the final set, but it is adequate in situations where dependence is expected between consecutive series. If we want to test the pairwise correlation of all the series we use `type = "all"`.

This procedure only keeps series that are not significantly correlated with any other selected series, which gives more evidence in favor of the correlation matrix of the chosen series being diagonal.

Breaking record ties

The CRM assumes that the X_t's are continuous RVs so that the probability of ties is zero. However, ties and *weak* records (observations equal to the current record value) can occur in a series even if the variable is continuous because the measured values are rounded. The function `series_ties()` gives the percentage of weak records in a series. It is important to know this percentage since, if it is high, the number of records will be lower than expected under the CRM, even if the series is IID (Wergen, Volovik, Redner, and Krug 2012).

If the number of weak records is high, the function `series_untie()`, which applies a simple procedure to break ties, should be used. It adds to each element in the series a uniform random value in the range $(-u/2, u/2)$, where u is the precision unit of the observations, so that weak records disappear. This procedure reproduces the fact that some of the weak records would have been records if they had not been rounded.

Backward series

The function `series_rev()` gives the backward series of the argument X. X can be a vector or a matrix, and in the last case the output is the backward series of each column.

Record variables

RecordTest includes functions to compute all the record RVs introduced in Section 2.1, given a series $(X_1, \ldots, X_T)^\top$ or a matrix **X**. `I.record()` computes the record indicators **I**, using a S3 method, and `N.record()` computes the observed cumulative number of records up to time t, **N**. Additionally, the record times **L** and record values **R** are computed by `L.record()` and `R.record()`, respectively. The arguments of all these functions are: X, the vector or matrix to analyze; `record`, a character string, `"upper"` or `"lower"`, indicating the type of records to be calculated; and `weak`, a logical argument to indicate whether weak records are considered. The function `p.record()` computes the MLEs $\hat{p}_t$'s in (6).

Under the CRM, N_t follows a Poisson binomial distribution. The package includes functions, `dpoisbinom()`, `ppoisbinom()`, `qpoisbinom()`, and `rpoisbinom()`, to compute the density, distribution, and quantile function, and a random generation for the Poisson binomial distribution using the algorithm by Hong (2013).

3.2. Exploratory data analysis

Records plot. The function `records()` plots the time series $(X_1, \ldots, X_T)^\top$ and identifies the upper and lower records observed in the series; one or both directions can be specified in argument `direction = c("forward", "backward", "both")`. This plot helps to detect asymmetries between the four types of records. If we have to analyze the extreme behavior of M series, an alternative is to summarize them into a single series, calculating the mean or the maximum in each time t and apply this function. Another alternative is the following plot.

Plot of the times of record. The function `L.plot()` plots the record times in M series, (L_{im}, m), for $i = 1, \ldots, N_{Tm}$, and $m = 1, \ldots, M$. The M series are represented in the vertical axis, and the record times in each series in the horizontal axis. The display includes four panels, one for each type of records, FU, FL, BU, and BL. This plot helps to study the hypothesis of the CRM since clear differences, especially in the number of points in the second half of the plots, suggest non-stationarity in the tails.

3.3. Tests and plots for non-stationarity based on one type of records

The tools described in this section can be applied even when the only information available is the record times. Three families of tests, based on the number of records, the probabilities of record and the likelihood of the record indicators, are implemented.

Number of records

The function `N.test()` includes several tests based on statistics related to the number of records. The general expression of the statistic is

$$N^\omega = \sum_{m=1}^{M} \sum_{t=1}^{T} \omega_t I_{tm}, \tag{10}$$

where the ω_t's are weights given to the records according to their position in the series. The reason of using weights is that records at high values of t are less likely to occur so that, if they occur, they give more evidence against the null hypothesis H_0. Thus, the use of weights makes records at high t to increase more the value of the statistic. The weights are controlled by the argument `weights` that must be a function. Diersen and Trenkler (1996) recommend linear weights $\omega_t = t - 1$, i.e., `weights = function(t) t - 1`. Cebrián *et al.* (2022) propose a score-sum test that is a particular case of this statistic with weights $\omega_1 = 0$ and $\omega_t = t^2/(t-1)$ ($t = 2, \ldots, T$), i.e., `weights = function(t) ifelse(t == 1, 0, t^2 / (t - 1))`. Both types of weights are asymptotically equivalent and increase the power of the test (Cebrián *et al.* 2022).

Under the null hypothesis H_0 in (7), N^ω is asymptotically normal as $M \to \infty$ with mean and variance

$$\mathsf{E}_0(N^\omega) = M \sum_{t=1}^{T} \omega_t \frac{1}{t} \qquad \text{and} \qquad \mathsf{VAR}_0(N^\omega) = M \sum_{t=2}^{T} \omega_t^2 \frac{1}{t} \left(1 - \frac{1}{t} \right). \tag{11}$$

When $\omega_t = 1$, N^ω is the raw number of records, it follows an exact Poisson binomial distribution, and it is asymptotically normal also in T. The argument `distribution` indicates the distribution to compute the p value, `"normal"` or `"poisson-binomial"`. With w_t's which are not equal to 0 or 1, only `distribution = "normal"` can be used. Alternatively, in any situation, the p value can be estimated using Monte Carlo simulations with `simulate.p.value = TRUE`. This is not often necessary since, even when N^ω is not asymptotically normal in T, the size based on the normal distribution is reasonably satisfactory even with $M = 1$.

Another test is based on an estimation of the variance instead of the variance under the null hypothesis H_0,

$$\tilde{N}_{\tilde{S}}^\omega = \frac{N^\omega - \mathsf{E}_0(N^\omega)}{\sqrt{\widehat{\mathsf{VAR}}(N^\omega)}}, \tag{12}$$

where $\widehat{\mathsf{VAR}}(N^{\omega})$ is the unbiased sample variance. The resulting test can be applied when $M > 1$, and it is more robust against serial correlation. Under the null hypothesis H_0, $\tilde{N}_S^{\omega}$ follows an asymptotic t_{M-1} distribution, and it is implemented using `distribution = "t"`.

All the tests in `N.test()` can be applied to any of the four types of records (FU, FL, BU, or BL). The argument `record` indicates the type of records, `"upper"` or `"lower"`, to be analyzed. If backward records were desired, `series_rev(X)` has to be used as argument. Other arguments of the function are `alternative`, that indicates the alternative hypothesis, `"greater"` or `"less"`. The argument `correct` indicates whether a continuity correction should be made in the normal or t distribution approximations, which is recommended. The last arguments can be used in most of the tests in the package.

The output of `N.test()` and most tests in the package is a list of class 'htest', which contains the components `statistic`, `parameter`, `p.value`, `alternative`, `estimate`, `method`, and `data.name`.

Plot of the number of records.   The asymptotic results described above can be used to analyze graphically the null hypothesis H_0. The function `N.plot()` plots the observed values $(t, \bar{N}_t^{\omega})$, where $\bar{N}_t^{\omega} = \sum_{m=1}^{M} \sum_{j=1}^{t} \omega_j I_{jm}/M$, together with the expected values under the null hypothesis H_0, $\mathsf{E}_0(\bar{N}_t^{\omega})$. As an alternative to confidence intervals of $\mathsf{E}(\bar{N}_t^{\omega})$, reference intervals (RIs) defined by the lower and upper $\alpha/2$th percentiles of the distribution of $\bar{N}_t^{\omega}$ under the null hypothesis H_0 are plotted, i.e.,

$$\mathsf{E}_0\left(\bar{N}_t^{\omega}\right) \pm z_{\alpha/2} \sqrt{\mathsf{VAR}_0\left(\bar{N}_t^{\omega}\right)}, \tag{13}$$

with $\mathsf{E}_0(\bar{N}_t^{\omega})$ and $\mathsf{VAR}_0(\bar{N}_t^{\omega})$ in (11) taking account of the average instead of the sum, and $z_{\alpha/2}$ the upper $\alpha/2$th percentile of the standard normal distribution. If the observed data follow the null distribution, they are expected to lie inside a particular RI $100(1-\alpha)\%$ of the time. It is noteworthy that the resulting RIs are not independent and the resulting bands are not reference bands at a $1-\alpha$ confidence level. However, they are useful to observe deviations from stationarity in the evolution of the number of records, and to identify the time point from which this deviation is significant.

Different weights can be specified with the argument `weights`. Several types of records can be plotted in the same graph using the argument `record`, which is a logical vector of length four (FU, FL, BU, and BL) that specifies the records to be plotted. There are two options to calculate the backward records, `backward = "T"` indicates that the backward number of records up to time t are calculated in the series observed up to time T, $(X_T, \ldots, X_1)^{\top}$, and `backward = "t"` in the series observed up to time t, $(X_t, \ldots, X_1)^{\top}$.

Probabilities of record

F **test for linear regression.**   The function `p.regression.test()` implements a test based on the fit of a regression model to the record probabilities p_t as a function of time. Under the null hypothesis H_0, the MLEs $\hat{p}_t$'s in (6) satisfy $\mathsf{E}_0(\hat{p}_t) = 1/t$. `p.regression.test()` implements an F test to compare the null model $\mathcal{M}_0 : \mathsf{E}(t\hat{p}_t) = 1$ against a model $\mathcal{M}_1$ where the expectation is a function of t, specified by the argument `formula`. The default is $\mathcal{M}_1 : \mathsf{E}(t\hat{p}_t) = \beta_0 + \beta_1 t$, $t = 2, \ldots, T$, that is `y ~ x`. More complicated time trends can be used, e.g., a quadratic trend with `formula = y ~ poly(x, degree = 2)`.

Given that the response $\hat{p}_t$ has a non-constant variance even under the null hypothesis H_0, the regression model has to be fitted using weighted least squares with weights $1/\mathsf{VAR}_0(t\hat{p}_t)$. A simulation study showed that the size of this test is satisfactory for $M > 10$. As in previous test functions, the p value can be estimated using Monte Carlo simulations with `simulate.p.value = TRUE`.

Plot of the probabilities of record. The function `p.plot()` represents the points $(t, t\hat{p}_t)$ for $t = 1, \ldots, T$, and the fitted linear model described in `p.regression.test()`, to evaluate the goodness of fit of the model. The fitted regression line can be replaced, e.g., by a locally estimated scatterplot smoothing (LOESS), using `smooth.method = stats::loess`. RIs based on the binomial distribution of $M\hat{p}_t$ when the null hypothesis H_0 is true are added to the plot if `conf.int = TRUE`. These RIs are independent and they are helpful to detect any substantial departure from the CRM at particular times t. The plot can be displayed on different scales: using `plot = "2"`, $\hat{p}_t$ is represented against t; and using `plot = "3"`, a logarithmic scale is used in both axes.

χ^2 **goodness-of-fit test.** The function `p.chisq.test()` implements a Pearson's χ^2 test,

$$M \sum_{t=2}^{T} (\hat{p}_t - \mathsf{E}_0(\hat{p}_t))^2 \left(\frac{1}{\mathsf{E}_0(\hat{p}_t)} + \frac{1}{1 - \mathsf{E}_0(\hat{p}_t)} \right), \tag{14}$$

comparing the observed and expected probabilities of record and no-record (see Benestad 2003, 2004, for more details). Under the null hypothesis H_0, the distribution of the statistic is asymptotically χ^2_{T-1}.

The size of the test is not appropriate for small M. In those cases, the function gives a warning message and it is convenient to estimate the p value using Monte Carlo simulations with `simulate.p.value = TRUE`.

Likelihood ratio and score tests

The functions `lr.test()` and `score.test()` compute the family of tests by Cebrián *et al.* (2022) to study the null hypothesis H_0 based on the likelihood and the score function of the record indicators **I**. The main difference with the previous tests is that they can be used to test both one-sided and two-sided alternatives. Although a different statistic has to be used in each case, we only have to indicate the adequate alternative in the argument `alternative` which can take values `"two.sided"`, `"greater"`, or `"less"`.

The default alternative in the two functions is that all the $T \times M$ probabilities p_{tm} may be different, with any restriction. Using `probabilities = "equal"`, both statistics are modified to study a particular case, that the probabilities in the M series are equal, although possibly different to $1/t$. According to Cebrián *et al.* (2022), those tests are less powerful than the tests for the general alternative, even if the probabilities in the M series are equal. In general, the tests in `score.test()` are more powerful and are recommended.

3.4. Tests and plots for non-stationarity based on different types of records

The power of the tests based on one type of records is improved by joining the information from the four types of records. Two families of this type of test are implemented in **RecordTest**. In both cases, the first step is to obtain the statistic described in the previous section for each type of records, and then build a joint statistic, or combine the p values of the resulting tests.

Tests based on joint statistics

The function `foster.test()` implements the tests based on joint statistics developed by Foster and Stuart (1954), with the possibility of adding weights, as suggested by Diersen and Trenkler (1996). Seven different statistics can be selected with `statistic = c("D", "d", "S", "s", "U", "L", "W")`. All the tests, apart from `"S"` and `"s"`, study the stationarity against the alternative of a trend in the mean. The statistics `"d"`, `"D"` and `"W"` analyze non-stationary behavior in both tails using two or four types of records,

$$d^\omega \;=\; N^{\omega,(U)} - N^{\omega,(L)} = \sum_{m=1}^{M} \sum_{t=1}^{T} \omega_t \left(I_{tm}^{(U)} - I_{tm}^{(L)} \right), \tag{15}$$

$$D^\omega \;=\; d^{\omega,(F)} - d^{\omega,(B)} = \sum_{m=1}^{M} \sum_{t=1}^{T} \omega_t \left(I_{tm}^{(FU)} - I_{tm}^{(FL)} - I_{tm}^{(BU)} + I_{tm}^{(BL)} \right), \tag{16}$$

$$W^\omega \;=\; \sum_{t=1}^{T} \omega_t \left(I_{tm}^{(FU)} + I_{tm}^{(BL)} \right). \tag{17}$$

The statistics in `"U"`, $U^\omega = N^{\omega,(FU)} - N^{\omega,(BU)}$, and `"L"`, $L^\omega = N^{\omega,(BL)} - N^{\omega,(FL)}$, only use the two types of upper or lower records and they are useful to detect trends only in the right or the left tail, respectively. The statistics in `"S"` and `"s"` study the existence of a trend in variation, and they are defined as $s^\omega = N^{\omega,(U)} + N^{\omega,(L)}$ and $S^\omega = s^{\omega,(F)} - s^{\omega,(B)}$. The statistics without weights are asymptotically normal in both T and M, and the weighted statistics only in M; although their size based on the normal distribution is satisfactory even with $M = 1$.

As explained in the definition of $\tilde{N}_S^\omega$ in (12), more robust statistics against serial correlation are obtained when the above statistics are standardized in mean and sample variance. The new statistics have an asymptotic t_{M-1} distribution and they are computed with `distribution = "t"`.

Plot of the Foster-Stuart statistics. The function `foster.plot()` plots the observed values of one of the statistics defined above, selected with `statistic`, obtained with the series observed up to time t, for every $t = 1, \ldots, T$. The plot also includes the expected values and RIs based on the normal approximation of the distribution of the statistic under the null hypothesis H_0. It is useful to detect the time t for which the stationarity hypothesis fails.

Global test. The function `global.test()` also computes a joint statistic, but it combines the statistic selected in `FUN`, say $\mathcal{X}$, which must be one of the two-sided tests in `p.regression.test()`, `p.chisq.test()`, `lr.test()`, or `score.test()`. By default, the global statistic $\mathcal{X}^G = \mathcal{X}^{(FU)} + \mathcal{X}^{(FL)} + \mathcal{X}^{(BU)} + \mathcal{X}^{(BL)}$ is used, but some terms can be omitted using argument `record`. The distribution of $\mathcal{X}^G$ is unknown, but the p value is estimated by Monte Carlo simulations.

Tests based on combined p values

The functions `brown.method()` and `fisher.method()` compute tests that combine the p values resulting from applying the tests with asymptotic normal distribution to the different types of records.

The function `fisher.method()` implements the general Fisher's method to combine the p values from any set of independent tests with the same null hypothesis; the vector of p values is the only argument. It can be applied to any test but, in the context of records, it is used to join the p values of the records that are asymptotically independent, i.e., FU and FL, BU and BL, FU and BL, or FL and BU.

The function `brown.method()` implements an algorithm to combine the p values of the tests in `N.test()` to any subset of the four types of records, selected with `record`. Since the p values are dependent, the algorithm is based on the Brown's method: the combined p values,

$$-2 \left(\log(pv^{(FU)}) + \log(pv^{(FL)}) + \log(pv^{(BU)}) + \log(pv^{(BL)}) \right), \qquad (18)$$

follow a $c\chi_f^2$ distribution with scale parameter c and degrees of freedom f that depend on the covariances of the p values. In general, this test is more powerful than the previous ones and the (seasonal) MK test when the series follow a GP or some types of GEV distributions with a linear drift in location (see Cebrián *et al.* 2022, for the details).

3.5. Tests for change-point detection

The function `change.point()` implements a family of tests to study the null hypothesis H_0 against the alternative hypothesis of an unknown change-point t_0, i.e.,

$$H_1 : p_{tm} = 1/t, \quad t = 1, \ldots, t_0 \quad \text{and} \quad p_{tm} \neq 1/t, \quad t = t_0 + 1, \ldots, T, \qquad (19)$$

for $m = 1, \ldots, M$. Note that these tests aim to detect the beginning of the non-stationary behavior in the tails, not in the mean. The test statistic given by Castillo-Mateo (2022a) is

$$K^\omega = \max_{1 \leq t \leq T} \left| \frac{N_t^\omega - \mathsf{E}_0(N_t^\omega)}{\sqrt{\mathsf{VAR}_0(N_T^\omega)}} - \frac{\mathsf{VAR}_0(N_t^\omega)}{\mathsf{VAR}_0(N_T^\omega)} \frac{N_T^\omega - \mathsf{E}_0(N_T^\omega)}{\sqrt{\mathsf{VAR}_0(N_T^\omega)}} \right|, \qquad (20)$$

where $N_t^\omega = \sum_{m=1}^M \sum_{j=1}^t \omega_j I_{jm}$, and the estimated change-point $\hat{t}_0$ is the value t where K^ω attains its maximum. K^ω is asymptotically Kolmogorov distributed as $T \to \infty$ if $\omega_t = 1$; otherwise, the p value has to be estimated by Monte Carlo simulations using `simulate.p.value = TRUE`. Weights equal to 1 or proportional to the inverse of the standard deviation of I_t are recommended. The test has been defined in terms of the number of upper or lower records N_t^ω, but it can also be defined in terms of $d_t^\omega = N_t^{\omega,(U)} - N_t^{\omega,(L)}$ or $s_t^\omega = N_t^{\omega,(U)} + N_t^{\omega,(L)}$, depending on `record = c("upper", "lower", "d", "s")`.

4. An example: Daily maximum temperature in Zaragoza

This section illustrates how package **RecordTest** can be used to analyze the effect of global warming on the records and extremes of a daily maximum temperature series, the series in Zaragoza, Spain. It is shown how the functions in the package cover all the steps of the analysis: data preparation, exploratory analysis, and inference to study the non-stationary behavior of the extremes and to identify the time, the periods of the year, and the features where the non-stationary behavior appears.

Record values and record times
Data: Tx
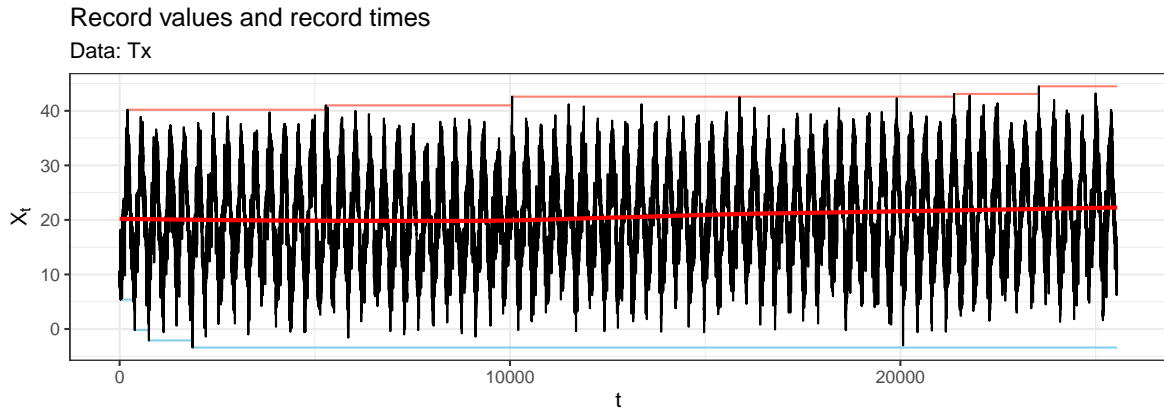
Figure 1:   Daily maximum temperature at Zaragoza, Spain (1951–2020). LOESS (solid red), and upper (red) and lower (blue) records.

4.1. Dataset

The 'data.frame' TX_Zaragoza included in **RecordTest** contains two columns: DATE, the dates in 'Date' format, spanning from "1951-01-01" to "2020-12-31"; and TX, the daily maximum temperature series at Zaragoza Airport (Spain), in tenths of degree Celsius (°C). The dataset has been downloaded from the European Climate Assessment & Dataset (ECA&D; Klein Tank *et al.* 2002) and modified by eliminating all the observations from the 29th of February. This is because when the series is split, these days would yield a four-year time series that is difficult to join to the analysis of the other yearly time series. The series with the 29th of February is also available as TX_Zaragoza29F. The series has three missing observations indicated by NA corresponding to "1951-03-31", "1965-01-04", and "1965-10-05". The dataset can be accessed after loading the package:

```
R> library("RecordTest")
R> Tx <- TX_Zaragoza$TX / 10
```

4.2. Data preparation and exploratory analysis

Most daily temperature series present a clear seasonal component and a high serial correlation. That also is the case for the Zaragoza series, which can easily be seen by plotting the series using the function records() (see Figure 1). The output of this and all the plot functions in **RecordTest** are 'ggplot' objects. Consequently, the plots can be easily improved using **ggplot2** (Wickham 2016) functions; an example of how to add a LOESS is shown in the following chunk,

```
R> records(Tx) +
+    ggplot2::geom_smooth(formula = y ~ x, method = stats::loess,
+       mapping = ggplot2::aes(y = Tx), se = FALSE, col = "red")
```

The plot reveals the seasonal behavior and a weak long-term time trend in the mean, summarized by the LOESS. The upper and lower records in the forward direction are also plotted, but their behavior is difficult to be analyzed due to the seasonality of the series.

Since the series shows annual seasonality, we use the function `series_split()` with `Mcols = 365` to apply the split procedure described in Section 3.1, and it returns 365 subseries of length $T = 70$. Then, we use `series_uncor()` to extract from them a subset of uncorrelated subseries, and it returns $M = 76$ subseries of length $T = 70$. We do not change the default option `test.fun = stats::cor.test` given that dependence in temperature data is usually linked to correlation. The default significance level is `alpha = 0.05`. Multiple testing corrections are not implemented in this example since, in order to apply the tests, it is more important to guarantee that the selected series are uncorrelated, while the problem of rejecting series which are uncorrelated is less problematic. The indexes of the subset of the selected series are shown with the argument `return.value = "indexes"`. As we can see, a separation between 2 and 14 days yields uncorrelated series.

```
R> TxZ365 <- series_split(Tx, Mcols = 365)
R> TxZ <- series_uncor(TxZ365)
R> dim(TxZ)

[1] 70 76

R> series_uncor(TxZ365, return.value = "indexes")

 [1]    1    5   13   16   20   25   29   34   40   48   54   62   76   80   83   88   95
[18]  100  104  108  111  118  122  126  133  138  143  152  156  162  169  176  179  188
[35]  194  197  199  202  205  207  209  212  215  220  223  226  229  238  241  245  251
[52]  257  261  267  270  274  277  286  292  298  306  308  312  314  317  321  325  333
[69]  337  339  342  345  350  357  361  364
```

The temperature series is rounded to the tenth of °C. Given that the rounding is small with respect to the value of the temperature, the assumption of continuous RVs is reasonable. The existence of a high percentage of ties may affect the validity of the tests, that are developed for continuous RVs, so it is advisable to apply `series_ties()` to check this percentage. The output is a 'list' with three elements: (i) `number`, a vector with the total number of records, the number of "strong" and weak records, and the expected number of records under IID series, (ii) `percentage`, the percentage of record ties out of the total number of records, and (iii) `percentage.position`, the percentage of record ties by time t, e.g., 4.76% of records were ties at $t = 5$. NaNs indicate that there were no records of any type, e.g., at $t = 19$.

```
R> lapply(series_ties(TxZ, record = "upper"), round, digits = 2)

$number
       Total           Strong            Weak        Expected under IID
       391.0            375.0            16.0                     367.3

$percentage
[1] 4.09

$percentage.position
       1        2        3        4        5        6        7        8        9       10
```
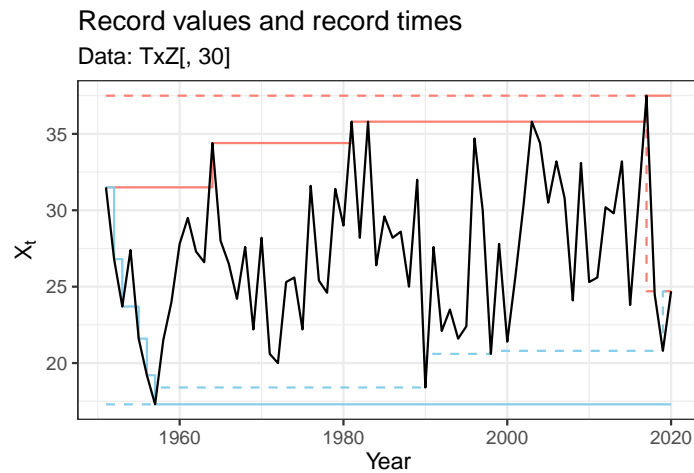
Figure 2:   Daily maximum temperature of the 11th of June at Zaragoza, Spain (1951–2020). Upper (red) and lower (blue) records in the forward (solid) and backward (dashed) directions.

```
 0.00    2.44    0.00    0.00    4.76    0.00    0.00    0.00    0.00    0.00
   11      12      13      14      15      16      17      18      19      20
11.11   14.29   20.00    0.00   16.67    0.00   50.00    0.00     NaN    0.00
   21      22      23      24      25      26      27      28      29      30
 0.00    0.00    0.00     NaN    0.00    0.00    0.00    0.00    0.00    0.00
   31      32      33      34      35      36      37      38      39      40
20.00    0.00   33.33   50.00   40.00  100.00    0.00    0.00    0.00    0.00
   41      42      43      44      45      46      47      48      49      50
 0.00    0.00    0.00    0.00   16.67    0.00     NaN    0.00    0.00     NaN
   51      52      53      54      55      56      57      58      59      60
 0.00    0.00   25.00    0.00    0.00    0.00     NaN     NaN    0.00    0.00
   61      62      63      64      65      66      67      68      69      70
 0.00    0.00    0.00   20.00    0.00    0.00    0.00    0.00    0.00    0.00
```

```
R> lapply(series_ties(TxZ, record = "lower"), round, digits = 2)
```

Since the percentage of ties is about 4.09% for upper records and 3.74% for lower records (output omitted), it does not seem to be necessary to apply the function `series_untie()` to break the ties.

In series without seasonality, the analysis of records is easier. As an example, Figure 2 shows the plot obtained with the chunk below, which includes the upper and lower records in the forward and backward directions of the temperature measured on the 11th of June (30th column in `TxZ`).

```
R> records(TxZ[, 30], direction = "both") +
+    ggplot2::scale_x_continuous(name = "Year", breaks = c(10, 30, 50, 70),
+      labels = c("1960", "1980", "2000", "2020"))
```

The plot shows evidence of an increasing trend, since no lower records occur after 7 time units in the forward series, and the last upper record occurs at time point 3 in the backward series.
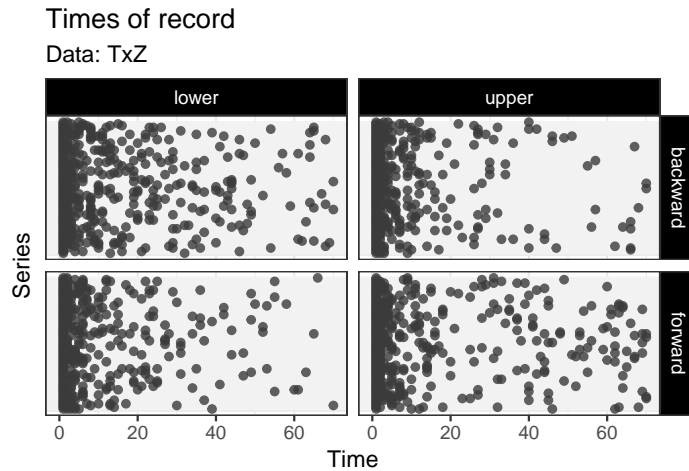
Figure 3: Plot of the times of record for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020).

The function `L.plot()` is used to summarize the times of the four types of records in each of the 76 uncorrelated subseries by means of the plot of the times of record,

```
R> L.plot(TxZ)
```

Figure 3 shows that as time evolves there are less FL and BU records than FU and BL records. This suggests non-stationary behavior since the pattern in the four plots should be similar in IID series. This effect is difficult to be observed in only one subseries; however, when the 76 subseries are plotted together the evidence is clearer.

4.3. Inference tools to study non-stationarity in the extremes

Tests to detect deviations from stationarity

The effect of global warming on the extremes of the temperature series may appear in the upper, the lower, or in both tails. We are interested in analyzing those hypotheses both jointly and individually, but in all cases by means of the null hypothesis H_0 in (7). The $M = 76$ subseries available correspond to different days within a year so they are not identically distributed. Consequently, under the alternative hypothesis the probabilities of record p_{tm} may be different in the M subseries.

Analysis of one tail. To analyze the behavior of the upper tail, we study the upper records. In the context of global warming, the alternative hypothesis of interest is

$$H_1 : p_{tm}^{(FU)} > 1/t, \qquad \text{for at least one } t = 1, \dots, T, \text{ and } m = 1, \dots, M. \tag{21}$$

This alternative hypothesis is quite general since it includes the existence of a monotonous positive trend in the mean, but also other types of non-stationarity, such as some non-monotonous trends. To test this hypothesis, we implement the weighted test in `N.test()` using the simple linear weights $\omega_t = t - 1$ and default arguments,

```
R> N.test(TxZ, weights = function(t) t - 1)
```

```
        Test on the weighted number of upper records with weights = t - 1
```

```
data:  TxZ
Z = 3.3138, p-value = 0.0004602
alternative hypothesis: true 'N' is greater than 4952.704
sample estimates:
          N            E          VAR
  6335.000    4952.704 173877.950
```

The hypothesis of stationarity is rejected at any usual significance level α. `N.test()` calculates the value of the statistic `N`, its standardized version `Z` with a continuity correction by default, and its expected value and variance under the null hypothesis H_0, `E` and `VAR`, respectively. The Monte Carlo p value obtained with the argument `simulate.p.value = TRUE` is very similar (not shown), indicating that the normal approximation is good.

To analyze the behavior of the lower tail, we also use `N.test()`. Under the alternative hypothesis of a positive trend in the mean, the probability of lower records is less than under the null hypothesis H_0, and we have to use the arguments `record = "lower"` and `alternative = "less"`. A significant behavior against stationarity is also observed in the lower tail, since the following yields a p value equal to `0.001044` (output omitted).

```
R> N.test(TxZ, weights = function(t) t - 1, record = "lower",
+     alternative = "less")
```

Given that the complete series is available, we can add more information to the study of one tail using the backward series and the more powerful tests implemented in `foster.test()`. Here, we apply the statistic U^ω defined in Section 3.4, based on the forward and backward upper records. The alternative for a positive trend in the mean must be the default `alternative = "greater"`. The p value is lower than that obtained with `N.test()` so more evidence to reject the null hypothesis H_0 is found,

```
R> foster.test(TxZ, weights = function(t) t - 1, statistic = "U")
```

```
        Forward - backward upper records test with weights = t - 1
```

```
data:  TxZ
Z = 4.0641, p-value = 2.411e-05
alternative hypothesis: true 'statistic' is greater than 0
sample estimates:
statistic        E        VAR
   3110.0      0.0  585579.8
```

Analysis of both tails.   To carry out a joint analysis of both tails, we use the D^ω statistic in (16) based on the four types of records. The alternative for a positive trend is again the default `alternative = "greater"`,

```
R> foster.test(TxZ, weights = function(t) t - 1, statistic = "D")

        Foster-Stuart D-statistic test with weights = t - 1

data:  TxZ
Z = 5.1889, p-value = 1.058e-07
alternative hypothesis: true 'statistic' is greater than 0
sample estimates:
statistic        E       VAR
     5692        0   1203318
```

The more robust version of the statistic against serial correlation as defined in (12) but for D^ω can be calculated using the argument `distribution = "t"`. The null hypothesis H_0 is also rejected at any usual significance level α,

```
R> foster.test(TxZ, weights = function(t) t - 1, statistic = "D",
+    distribution = "t")

        Foster-Stuart D-statistic test with weights = t - 1

data:  TxZ
t = 5.263, df = 75, p-value = 6.507e-07
alternative hypothesis: true 't' is greater than 0
```

Another option to carry out a joint analysis is to apply Brown's method using the default option that combines the p values of `N.test()` for the four types of records. Although it is the default option, we specify the alternative hypothesis for the four types of records with `alternative`, as an example of use,

```
R> brown.method(TxZ, weights = function(t) t - 1,
+    alternative = c("FU" = "greater", "FL" = "less", "BU" = "less",
+    "BL" = "greater"))

    Brown's method on the weighted number of records with weights = t - 1

data:  TxZ
X-squared = 38.669, df = 4.7592, c = 1.6810, p-value = 2.088e-07
```

It is noteworthy that the tests joining the information of the four types of records give the lowest p values, on the order of 10^{-7}. They lead to conclude, at any usual significance level α, that the probabilities of record are greater for FU and BL records, and less for FL and BU records, than expected under the CRM. This gives evidence of non-stationarity in the occurrence of records in the subseries and, consequently, the existence of an increasing positive trend in daily maximum temperature that affects the occurrence of extremes.

Graphical tools to detect deviations from stationarity

We have formally tested the existence of a significant non-stationary behavior both in the upper and lower tails of temperature. Our next aim is to characterize that behavior using graphical tools, to identify when it appears, which features are affected, etc.
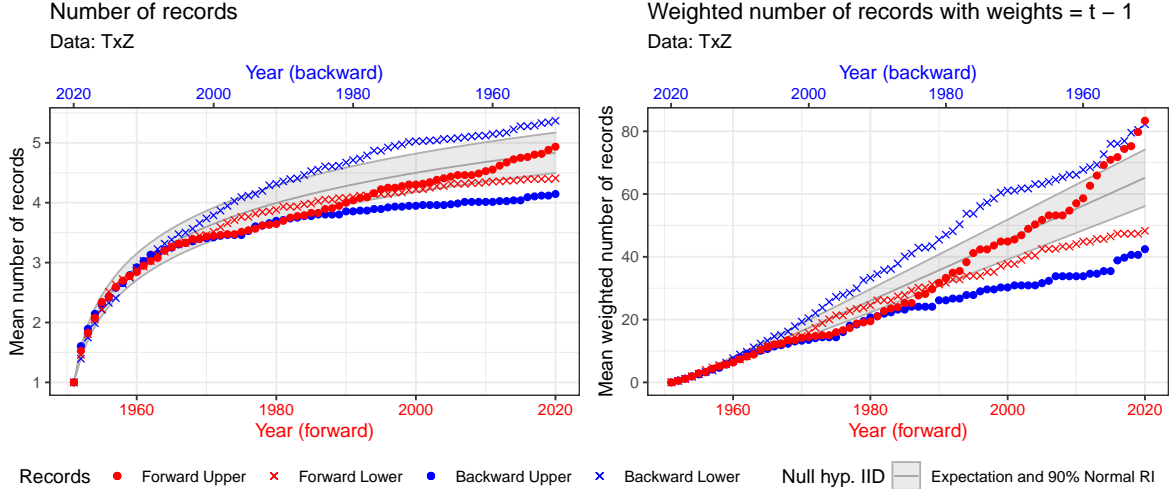
Figure 4:   Plot of the number of records for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020). Expected values and 90% RIs (mean, 5th and 95th percentiles of the distribution of $\bar{N}_t^\omega$ under the null hypothesis H_0) for the four types of records (gray shaded area). Left: Unweighted statistics. Right: Weighted statistics with linear weights.

First, we analyze the behavior of the four types of records with the plot of the number of records using `N.plot()`. To facilitate the comparison, all types of records are displayed together using the default argument `record`.

```
R> N.plot(TxZ)
R> N.plot(TxZ, weights = function(t) t - 1)
```

The chunk above gives the plots by default, but Figure 4 is obtained adding **ggplot2** functions to draw the time axis for the forward and backward series; the complete code is available in the supplementary material. The left plot shows that the number of FU records in the 80s is slightly lower than expected in a stationary series. From that point onward, the number of records increases until the end, although it does not become significantly high. FL records have a stationary behavior up to the 90s, but its number starts to be lower than expected thereafter. Backward records show more clear deviations of stationarity, and this suggests that non-stationary behavior is stronger in the last part of the observed period. Both types of backward records are outside the RIs from the first 30 observations, which correspond to the period spanning from 1991 to 2020. The non-stationary behavior observed in the four types of records is the behavior expected in a series with a positive trend in the mean. The right plot is obtained using linear weights to give more importance to the occurrence of records in high values of t, where the probability of record is lower. It shows that the use of weights leads to clearer evidence of non-stationarity: the deviation of the forward records is now significant and the deviation in the backward series is detected even earlier.

To analyze both tails jointly, we combine the information of the four types of records in one signal. We can show a plot equivalent to the previous one based on the Foster-Stuart D^ω statistic in (16) with `foster.plot()`, whose expected value under the null hypothesis H_0 is zero. We do not show the **ggplot2** functions for simplicity,

```
R> foster.plot(TxZ, weights = function(t) t - 1)
```
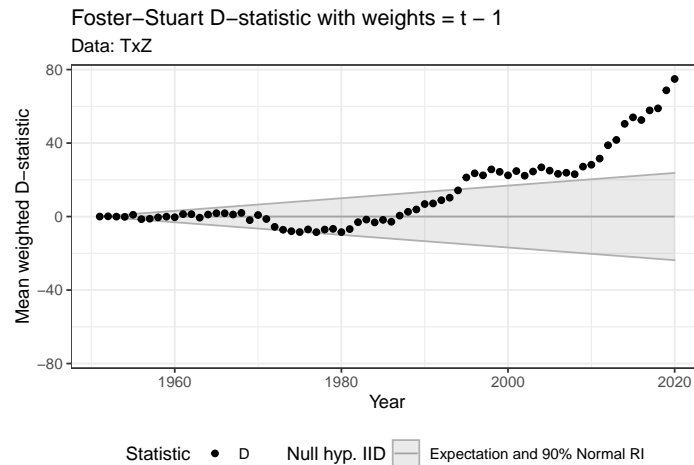
Figure 5: Plot of the mean value of the Foster-Stuart D^ω statistic with linear weights for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020), observed up to time t, $t = 1, \ldots, T$. Expected values and 90% RIs under the null hypothesis H_0 (gray shaded area).

Figure 5 shows a significant non-stationary behavior from the latest 90s onward and the statistic shows a strong increasing trend starting around 2010.

Another approach to characterize non-stationarity is the analysis of the probabilities of records. We plot $t\hat{p}_t$ against t using `p.plot()` with the default argument `plot = "1"`, for FU records. Under the null hypothesis H_0, the fitted regression to those points should be a horizontal line, but different alternatives may be fitted; here, a quadratic trend is considered,

```
R> p.plot(TxZ, record = c("FU" = 1, "FL" = 0, "BU" = 0, "BL" = 0),
+    smooth.formula = y ~ poly(x, degree = 2))
```

The top plot in Figure 6 shows that the fitted curve is clearly different from zero and many values $t\hat{p}_t$ from the late 90s onward are outside the RIs. This plot helps us to identify the years where the probability of record is much higher than expected. To characterize the lower tail, the FL and BL records are shown in the same plot but with different colors using `point.col`,

```
R> p.plot(TxZ, record = c("FU" = 0, "FL" = 1, "BU" = 0, "BL" = 1),
+    point.col = c("FU" = NA, "FL" = "blue", "BU" = NA, "BL" = "red"))
```

The bottom plot in Figure 6 shows that FL records are less informative in the case of an increasing trend. In effect, in that case, FL probabilities tend to decrease, but given that they are always bounded by zero, points lower than the low interval bound cannot appear.

To formally check if the deviation from the CRM is significant, we apply the F test in `p.regression.test()` to study $\mathsf{E}(t\hat{p}_t) = 1$. Since the previous function `p.plot()` suggests a quadratic trend as an alternative, we use

```
R> p.regression.test(TxZ, formula = y ~ poly(x, degree = 2))

        Regression test on the upper records probabilities
```
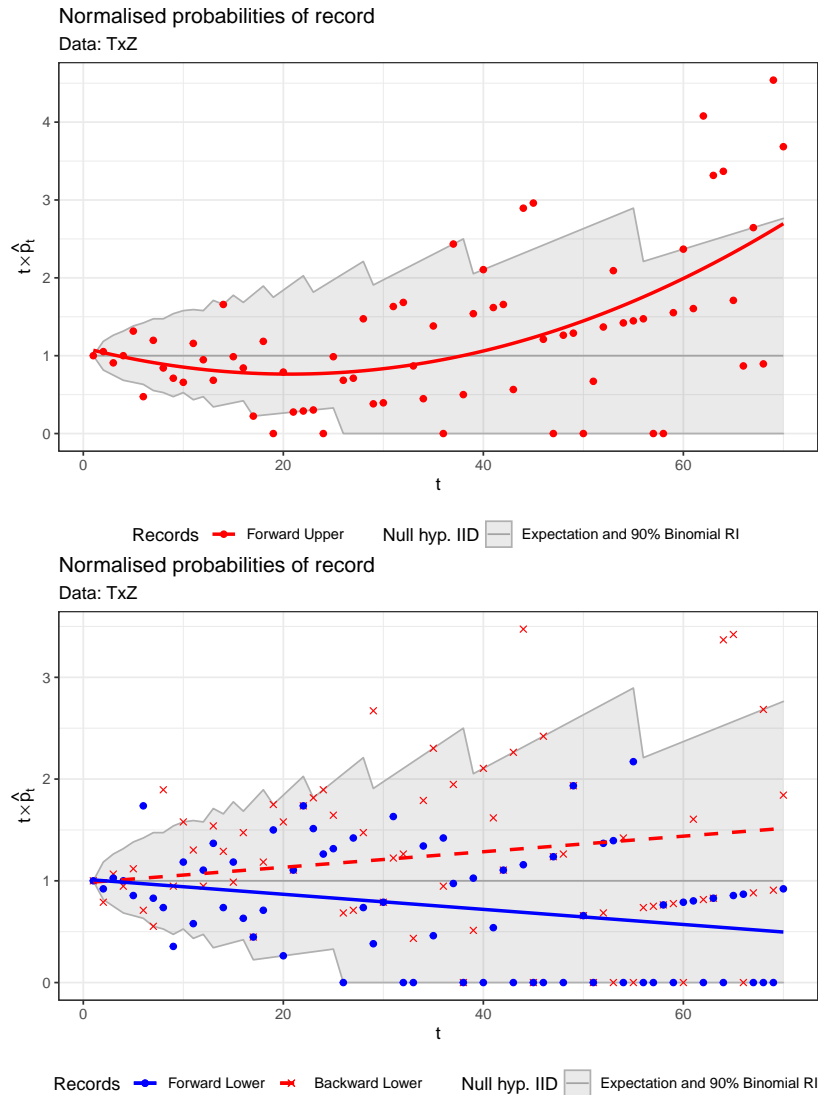
Figure 6: Plot of the normalized probabilities of record for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020). Expected values and 90% RIs for $t \times \hat{p}_t$ (gray shaded area). Top: FU records with quadratic time trend. Bottom: FL and BL records with linear time trend.

```
data:  TxZ
F = 9.0496, df1 = 3, df2 = 66, p-value = 4.225e-05
alternative hypothesis: two-sided for record probabilities
null values:
        (Intercept) poly(x, degree = 2)1 poly(x, degree = 2)2
                  1                    0                    0
sample estimates:
        (Intercept) poly(x, degree = 2)1 poly(x, degree = 2)2
           1.265065             4.028521             2.337520
```
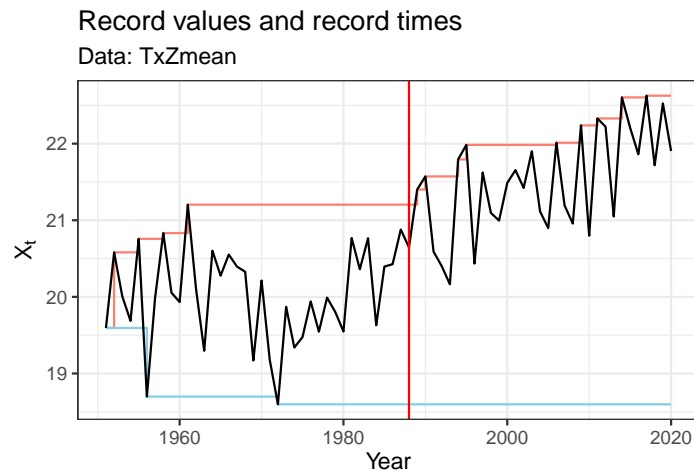
Record values and record times
Data: TxZmean

Figure 7: Annual mean temperature at Zaragoza, Spain (1951–2020). Change-point estimate (vertical solid red), and upper (red) and lower (blue) records.

We also apply the *F* test to the BL records using `series_rev()` to calculate the backward series, and the argument `records = "lower"` (output omitted),

```
R> p.regression.test(series_rev(TxZ), record = "lower")
```

The resulting p value is `0.039`. Both tests suggest that the p_{tm}'s are significantly different from $1/t$. The evidence in BL records is not so strong but still significant, as it is observed in the number of points outside the RIs in the bottom plot in Figure 6.

Tests for change-point detection

Once we have found evidence of a trend in the tails of the temperature distribution, our aim is to identify the time point where this trend starts. First, we consider a series without seasonal behavior, the annual mean temperature. Figure 7 shows the annual mean temperature at Zaragoza, together with the change-point estimate, and its upper and lower records, resulting from

```
R> TxZmean <- rowMeans(TxZ365, na.rm = TRUE)
R> records(TxZmean) +
+    ggplot2::scale_x_continuous(name = "Year", breaks = c(10, 30, 50, 70),
+      labels = c("1960", "1980", "2000", "2020")) +
+    ggplot2::geom_vline(xintercept = change.point(TxZmean)$estimate,
+      color = "red")
```

It seems reasonable to check the null hypothesis H_0 against the alternative hypothesis H_1 in (19) using a change-point test based on upper records without weights,

```
R> change.point(TxZmean)

        Records test for single changepoint detection
```

```
data:  TxZmean
Kolmogorov = 3.7425, p-value = 1.366e-12
alternative hypothesis: two.sided
sample estimates:
probable changepoint time
                        38
```

The p value of order 10^{-12} yields to reject the null hypothesis H_0 at any usual significance level α, and the estimated change-point $\hat{t}_0 = 38$ corresponds to the year 1988.

The change-point test can also be applied to the 76 uncorrelated subseries as given in the chunk below (output omitted). The estimated change-point on a daily scale is $\hat{t}_0 = 36$ (1986), and the p value `0.0003547` is significant at any usual significance level α.

R> change.point(TxZ)

5. Summary and future work

The study of non-stationary behavior in the extremes and the tails of a distribution is important in data analysis in many fields, such as environmental sciences, climate, finance, or sports. However, most of the available software packages to analyze non-stationarity focuses on the study of the mean. As far as we know, the R package **RecordTest** is the only available software package for the analysis of record-breaking events. In addition, the use of records provides a useful general framework for a fully non-parametric analysis of non-stationary behavior in the extremes. The underlying idea of all the inference tools implemented in the package is to use the distribution of the record occurrences under the classical record model, and study if the observed records are compatible with that behavior.

The package offers functions that cover all the steps in this type of analysis. This includes functions to prepare the data, obtaining a set of uncorrelated series with no seasonal behavior from the original series, identify the variables for characterizing the record occurrence, and implement graphical tools for exploratory analysis. The main functionality of the package is the implementation of all the tests to detect non-stationarity based on records currently available in the literature, and complementary graphical tools. The null hypothesis H_0 of all the tests is that the series are sequences of IID continuous RVs, expressed in terms of the probabilities of record, i.e., $p_t = 1/t$. There are two main families implemented, the first one can be applied even when the only information available are the times of record, and this includes tests based on the number of records, the probabilities of record, and the likelihood of the record indicators. The second family requires to know the entire series but it includes the most powerful tests. The underlying idea is to combine the information from four types of records, the upper and lower records in the forward and backward series, using joint statistics or joint p values. Different alternative hypothesis, one-sided, two-sided or even the existence of a change-point can be studied with the wide range of available tests.

The applicability of the package to analyze real data is illustrated with the analysis of the effect of global warming on the extremes of the daily maximum temperature in Zaragoza, Spain. The availability of the tools implemented in **RecordTest** will favor the realization of studies for analyzing records and non-stationarity in the extremes in many fields.

Future work will focus on the implementation of permutation tests, although this approach requires further development in the literature. This procedure will capture the dependence between the M series, so the tests will not require independent series. It will be especially useful to jointly analyze series with spatio-temporal dependence.

Acknowledgments

References

Arnold BC, Balakrishnan N, Nagaraja HN (1998). *Records*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York. doi:10.1002/9781118150412.

Benestad RE (2003). "How Often Can We Expect a Record Event?" *Climate Research*, **25**(1), 3–13. doi:10.3354/cr025003.

Benestad RE (2004). "Record-Values, Nonstationarity Tests and Extreme Value Distributions." *Global and Planetary Change*, **44**(1–4), 11–26. doi:10.1016/j.gloplacha.2004.06.002.

Bocharov G (2022). *pyextremes: Extreme Value Analysis (EVA) in Python*. Python package version 2.2.5, URL https://github.com/georgebv/pyextremes.

Castillo-Mateo J (2022a). "Distribution-Free Changepoint Detection Tests Based on the Breaking of Records." *Environmental and Ecological Statistics*, **29**(3), 655–676. doi:10.1007/s10651-022-00539-2.

Castillo-Mateo J (2022b). *RecordTest: Inference Tools in Time Series Based on Record Statistics*. R package version 2.1.1, URL https://CRAN.R-project.org/package=RecordTest.

Cebrián AC, Abaurrea J, Asín J (2015). "**NHPoisson**: An R Package for Fitting and Validating Nonhomogeneous Poisson Processes." *Journal of Statistical Software*, **64**(6), 1–25. doi:10.18637/jss.v064.i06.

Cebrián AC, Castillo-Mateo J, Asín J (2022). "Record Tests to Detect Non-Stationarity in the Tails with an Application to Climate Change." *Stochastic Environmental Research and Risk Assessment*, **36**(2), 313–330. doi:10.1007/s00477-021-02122-w.

Cheng L, AghaKouchak A, Gilleland E, Katz RW (2014). "Non-Stationary Extreme Value Analysis in a Changing Climate." *Climatic Change*, **127**(2), 353–369. doi:10.1007/s10584-014-1254-5.

Coles S (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag, London. doi:10.1007/978-1-4471-3675-0.

Coumou D, Robinson A, Rahmstorf S (2013). "Global Increase in Record-Breaking Monthly-Mean Temperatures." *Climatic Change*, **118**(3–4), 771–782. doi:10.1007/s10584-012-0668-1.

Diersen J, Trenkler G (1996). "Records Tests for Trend in Location." *Statistics*, **28**(1), 1–12. doi:10.1080/02331889708802543.

Foster FG, Stuart A (1954). "Distribution-Free Tests in Time-Series Based on the Breaking of Records." *Journal of the Royal Statistical Society B*, **16**(1), 1–22. doi:10.1111/j.2517-6161.1954.tb00143.x.

Gembris D, Taylor JG, Suter D (2002). "Trends and Random Fluctuations in Athletics." *Nature*, **417**(6888), 506. doi:10.1038/417506a.

Gembris D, Taylor JG, Suter D (2007). "Evolution of Athletic Records: Statistical Effects versus Real Improvements." *Journal of Applied Statistics*, **34**(5), 529–545. doi:10.1080/02664760701234850.

Gilleland E, Katz RW (2016). "**extRemes** 2.0: An Extreme Value Analysis Package in R." *Journal of Statistical Software*, **72**(8), 1–39. doi:10.18637/jss.v072.i08.

Heffernan JE, Stephenson AG (2018). ***ismev**: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.42, URL https://CRAN.R-project.org/package=ismev.

Hirsch RM, Slack JR, Smith RA (1982). "Techniques of Trend Analysis for Monthly Water Quality Data." *Water Resources Research*, **18**(1), 107–121. doi:10.1029/wr018i001p00107.

Hong Y (2013). "On Computing the Distribution Function for the Poisson Binomial Distribution." *Computational Statistics & Data Analysis*, **59**, 41–51. doi:10.1016/j.csda.2012.10.006.

Hussain M, Mahmud I (2019). "**pyMannKendall**: A Python Package for Non Parametric Mann Kendall Family of Trend Tests." *Journal of Open Source Software*, **4**(39), 1556. doi:10.21105/joss.01556.

Kendall MG, Gibbons JD (1990). *Rank Correlation Methods*. A Charles Griffin Title, 5th edition. Oxford University Press, London.

Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Mileta M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Forland E, Mietus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnar T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander LV, Petrovic P (2002). "Daily Dataset of 20th-Century Surface Air Temperature and Precipitation Series for the European Climate Assessment." *International Journal of Climatology*, **22**(12), 1441–1453. doi:10.1002/joc.773.

Kojadinovic I (2023). **npcp**: *Some Nonparametric CUSUM Tests for Change-Point Detection in Possibly Multivariate Observations.* R package version 0.2-5, URL `https://CRAN.R-project.org/package=npcp`.

Lehmann J, Coumou D, Frieler K (2015). "Increased Record-Breaking Precipitation Events under Global Warming." *Climatic Change*, **132**(4), 501–515. `doi:10.1007/s10584-015-1434-y`.

Lehmann J, Mempel F, Coumou D (2018). "Increased Occurrence of Record-Wet and Record-Dry Months Reflect Changes in Mean Rainfall." *Geophysical Research Letters*, **45**(24), 13468–13476. `doi:10.1029/2018gl079439`.

Mann HB (1945). "Nonparametric Tests against Trend." *Econometrica*, **13**(3), 245–259. `doi:10.2307/1907187`.

McLeod AI (2022). **Kendall**: *Kendall Rank Correlation and Mann-Kendall Trend Test.* R package version 2.2.1, URL `https://CRAN.R-project.org/package=Kendall`.

Patakamuri SK, O'Brien N (2021). **modifiedmk**: *Modified Versions of Mann Kendall and Spearman's Rho Trend Tests.* R package version 1.6, URL `https://CRAN.R-project.org/package=modifiedmk`.

Pfaff B, McNeil A (2018). **evir**: *Extreme Values in R.* R package version 1.7-4, URL `https://CRAN.R-project.org/package=evir`.

Pohlert T (2020). **trend**: *Non-Parametric Trend Tests and Change-Point Detection.* R package version 1.1.4, URL `https://CRAN.R-project.org/package=trend`.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ribatet M (2022). **SpatialExtremes**: *Modelling Spatial Extremes.* R package version 2.1-0, URL `https://CRAN.R-project.org/package=SpatialExtremes`.

Shcherbakov R, Davidsen J, Tiampo KF (2013). "Record-Breaking Avalanches in Driven Threshold Systems." *Physical Review E*, **87**(5), 052811. `doi:10.1103/physreve.87.052811`.

Southworth H, Heffernan JE, Metcalfe PD (2020). **texmex**: *Statistical Modelling of Extreme Values.* R package version 2.4.8, URL `https://CRAN.R-project.org/package=texmex`.

Stephenson AG (2002). "**evd**: Extreme Value Distributions." *R News*, **2**(2), 31–32. URL `https://journal.R-project.org/articles/RN-2002-015/`.

The MathWorks Inc (2022). *MATLAB – The Language of Technical Computing, Version R2022b.* Natick. URL `https://www.mathworks.com/products/matlab/`.

Van Aalsburg J, Newman WI, Turcotte DL, Rundle JB (2010). "Record-Breaking Earthquakes." *Bulletin of the Seismological Society of America*, **100**(4), 1800–1805. `doi:10.1785/0120090015`.

Van Rossum G, *et al.* (2011). *Python Programming Language.* URL `https://www.python.org/`.

Vogel RM, Zafirakou-Koulouris A, Matalas NC (2001). "Frequency of Record-Breaking Floods in the United States." *Water Resources Research*, **37**(6), 1723–1731. doi:10.1029/2001wr900019.

Wergen G (2013). "Records in Stochastic Processes – Theory and Applications." *Journal of Physics A: Mathematical and Theoretical*, **46**(22), 223001. doi:10.1088/1751-8113/46/22/223001.

Wergen G, Krug J (2010). "Record-Breaking Temperatures Reveal a Warming Climate." *EPL (Europhysics Letters)*, **92**(3), 30008. doi:10.1209/0295-5075/92/30008.

Wergen G, Volovik D, Redner S, Krug J (2012). "Rounding Effects in Record Statistics." *Physical Review Letters*, **109**(16), 164102. doi:10.1103/physrevlett.109.164102.

Wickham H (2016). **ggplot2**: *Elegant Graphics for Data Analysis*. 2nd edition. Springer-Verlag, New York. doi:10.1007/978-3-319-24277-4.

Yoder MR, Turcotte DL, Rundle JB (2010). "Record-Breaking Earthquake Intervals in a Global Catalogue and an Aftershock Sequence." *Nonlinear Processes in Geophysics*, **17**(2), 169–176. doi:10.5194/npg-17-169-2010.

Affiliation:

Jorge Castillo-Mateo, Ana C. Cebrián, Jesús Asín
Department of Statistical Methods
University of Zaragoza
Pedro Cerbuna 12
50009 Zaragoza, Spain
E-mail: jorgecm@unizar.es, acebrian@unizar.es, jasin@unizar.es

3.8 Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021

This manuscript was published in:

Castillo-Mateo, J., Cebrián, A. C., & Asín, J. (2023). Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021. *Atmospheric Research, 293*, 106934. https://doi.org/10.1016/j.atmosres.2023.106934

And it was disseminated (speaker emphasized) in:

- **Castillo-Mateo, J.**, Cebrián, A. C., & *Asín, J.* (2022, October 19–21). *Patrones estacionarios y no estacionarios en la ocurrencia de récords de la temperatura diaria en la Península Ibérica (Stationary and non-stationary patterns in the occurrence of daily temperature records in the Iberian Peninsula)* [Contributed talk]. 12 Congreso de la Asociación Española de Climatología, Santiago de Compostela, Spain.

"Faith is a torment, did you know that? It is like loving someone who is out there in the darkness but never appears, no matter how loudly you call."

E. Ingmar Bergman, in *The Seventh Seal*

ELSEVIER

Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021

Jorge Castillo-Mateo [*], Ana C. Cebrián, Jesús Asín

Department of Statistical Methods, University of Zaragoza, Pedro Cerbuna 12, 50009 Zaragoza, Spain

ABSTRACT

This work analyses the effects of global warming in the upper extremes of daily temperature series over Spain. This objective implies specific analysis, since time evolution of mean temperature is not always parallel to evolution of the extremes. We propose the use of several record tests to study the behavior of the extreme and record-breaking events in different temperature signals, at different time and spatial scales. The underlying idea of the tests is to compare the occurrence of the extreme events in the observed series and the occurrence in a stationary climate. Given that under global warming, an increasing trend, or an increasing variability, can be expected, the alternative is that the probability of the extremes is higher than in a stationary climate. Some of the tests, based on a permutation approach, can be applied to sets of correlated series and this allows the analysis of short periods of time and regional analysis, where series are measured in close days and/or locations. Using these tests, we evaluate and compare the effects of climate change in temperature extreme and record-breaking events using 36 series of daily maximum temperature from 1960 to 2021, all over peninsular Spain. We also compare the behavior in different Spanish regions, in different periods of the year, and in different signals such as the annual maximum temperature. Significant evidences of the effect of an increasing trend in the occurrence of upper extremes are found in most of Spain. The effects are heterogeneous within the year, being autumn the season where the effects are weaker and summer where they are stronger. Concerning the spatial variability, the Mediterranean and the North Atlantic region are the areas where the effects are more and less clear, respectively.

1. Introduction

In the framework of climate change, there are many works that analyze the evolution of mean temperature over time, and the existence of an increasing trend is generally accepted; see Sánchez-Lugo et al. (2019) for a review and Peña-Angulo et al. (2021) for a study on the Spanish mainland. However, changes in variability and extremes of temperature are also relevant (Schär et al., 2004). The interest of analyzing whether the occurrence of extreme and record-breaking temperatures is affected by climate change is clear (Kysely, 2010; Coumou et al., 2013; Saddique et al., 2020; Om et al., 2022). The reason is that some of the most serious consequences of global warming on human health and other fields, such as agriculture or energy consumption, are often related to the occurrence of increasingly intense extremes (Tan et al., 2007; Coumou and Rahmstorf, 2012). The Mediterranean region has been referenced as a hot spot of climate change, this is, a region whose climate is especially responsive to global warming (Diffenbaugh and Giorgi, 2012; Lionello and Scarascia, 2018; Tuel and Eltahir, 2020).

In fact, this region has suffered warming 20% faster than the rest of the globe (MedECC, 2020). This makes the study of temperature extremes in the Spanish mainland of special interest (Cos et al., 2022).

The numerous studies to analyze the evolution of mean temperature have been favored by the availability of simple distribution-free statistical tests, such as Mann–Kendall (MK) test (Mann, 1945; Kendall and Gibbons, 1990) and easy-to-use software to compute them. Extreme- and record-event type statistics exist since the 70s (Feller, 1991; Arnold et al., 1998; Bunge and Goldie, 2001) but they have not been widely applied within the climate community. The existing analysis of record-breaking temperatures usually aim to describe observed records and to identify the role of different factors in their occurrence (Xu et al., 2021; Zhang et al., 2021). Other works analyze the projected behavior of extreme records under different emission scenarios (Xu and Wu, 2019; Fischer et al., 2021; Yu et al., 2023).

Specific tools are required to analyze the tails of temperature distribution since their evolution may not be parallel to the evolution of the mean. In addition, if the magnitude of a trend in the mean is small in

terms of the variability of the series, the effect on the extremes of that trend may be difficult to detect. This is the case of global warming trends in daily temperatures. On the other hand, in aggregated data like global mean temperatures, the natural variability is smaller and the effect of the climate change is easier to detect. For this reason, the studies to quantify the effects of global warming in the extremes have focused on the analysis of annual or monthly summaries of temperatures (Zorita et al., 2008; Coumou et al., 2013; van der Wiel and Bintanja, 2021; Salameh et al., 2019), although in this context the study of the effect in daily temperatures is more important because averaged or summarized data can under-represent warm or cool periods with a persistence of only a few days (Yosef et al., 2021).

Some works to analyze the effects of climate change on record-breaking temperatures use probabilistic properties of the occurrence of records in independent and identically distributed (i.i.d.) series to quantify their evolution (Redner and Petersen, 2006; Coumou et al., 2013; Wergen et al., 2014). Gouet et al. (2020) characterize the probabilities of record in sequences of variables with a linear trend in location, but they require restrictive assumptions about the distribution of temperature. In this context, it is of great interest to use the probabilistic properties of records to develop formal statistical tools, however, only a few works try to provide hypothesis tests to objectively establish the effects of climate change in the extreme and record-breaking temperatures. In this line, Benestad (2003), Benestad (2004) proposed record-statistic tests to detect non-stationarities based on records in forward and backward series and Monte Carlo integration, and applied them to analyze spatially aggregated monthly mean temperatures. Benestad (2008) underscored the utility of these tests in terms of evaluating trends in extremes. Meehl et al. (2009) and Anderson and Kostinski (2011) studied the ratio of daily and monthly record high maximum temperatures to record low minimum temperatures averaged across the US and compared it with the expected ratio under stationary conditions. Cebrián et al. (2022) and Castillo-Mateo (2022) provide a wide family of distribution-free record tests which can be applied without any assumption about the temperature distribution. Another important advantage of these tests is their high power even when the underling trend is small compared to the variability of data, what makes them useful to the analysis of any type of temperature signals, including daily signals. They can be applied both locally and regionally. Given that the effects of climate change show important differences depending on the climate and region, these tests are useful tools to assess and evaluate those differences. The analysis using these tools is highly facilitated by the R package (R Core Team, 2022) RecordTest (Castillo-Mateo, 2023; Castillo-Mateo et al., 2023). All the tests and graphical tools used in this work, as well as other tools based on records (Foster and Stuart, 1954; Diersen and Trenkler, 1996; Benestad, 2003; Benestad, 2004), are implemented in this package.

The contribution of this work is the use of the statistical tools developed by Cebrián et al. (2022) and Castillo-Mateo (2022) to assess and analyze the effect of global warming in the extreme and record-breaking events in different signals of 36 daily maximum temperature series in peninsular Spain over 1960–2021. In addition, a modification of the tests based on a permutation approach is proposed. This new approach can be used with correlated series, which allows the application of the tests in smaller regions and shorter periods of time. Using these tests, we evaluate and compare the effects of climate change in the extremes over different Spanish regions, globally and in different periods of the year.

The outline of the paper is as follows. Section 2 presents the time series of observed temperature data in 36 Spanish stations and an exploratory data analysis. Subsequently, the section introduces the methodology to detect non-stationarity in the occurrence of extreme and record-breaking events. Section 3 shows the analysis of the evolution over time of extreme and record-breaking temperatures in different daily temperature signals using Spanish series. It also shows relationships between records in the observed series and in temperature series at

geopotential levels from ERA5 reanalysis. Section 4 concludes the paper with a discussion and conclusions.

2. Data and methods

2.1. Data

The database in this study has been extracted from the European Climate Assessment & Dataset (ECA&D; Klein Tank et al., 2002). It includes surface observations of daily maximum temperature (T_x) in °C in 36 stations located in the Spanish territory within the Iberian Peninsula, in the period 1960–2021, see Fig. 1. The Iberian Peninsula includes areas with very different climates (Chazarra-Bernabé et al., 2022). To a great extent, climate in inland Spain is temperate with a dry summer (Cs, Köppen classification), but some areas in the Central Plateau, Southeast of Andalusia and Ebro valley have a semi-arid climate (BS). North coast and Atlantic North coast show a temperate climate with no dry season (Cf). Mediterranean coast has a Cs climate in the North and BS in the Southeast. Our database includes series representing the different climate zones and also different elevation, 13 stations are over 500 m a. s.l., 5 stations over 800 m a.s.l. and one in the Central Mountains reaches 1894 m a.s.l. Climate summary measures of the locations are shown in Table S1 from Supplementary material.

All the series in the database have less than 100 missing observations in the considered period, this is, less than 0.5% of data are missing. Missing data are not removed from the series since the statistical tools applied in the analysis are weakly affected by a small percentage of missing data. Observations corresponding to February 29 are removed from the dataset for convenience. Furthermore, observed temperature data was rounded to the nearest tenth of a °C, resulting in a small probability of ties. Our analysis only considers strong records, this is, values strictly higher than the existing record. This means that the results are conservative in the sense that the number of records without the rounding effect could be around a 3% higher.

2.1.1. Exploratory data analysis

This sections shows some exploratory analysis to describe how the distribution of T_x has changed in time over the observed period 1960–2021. In particular, we compare the evolution over time of the median and the 5th and 95th percentiles of daily temperature. The aim of this analysis is to study if the effects of climate change are similar in the center and in the tails (extremes) of temperature distribution. If there exist relevant differences, the conclusions about the effects of global warming in the mean would not be valid for the tails, revealing the need of a specific analysis for the evolution of extreme and record-breaking temperatures.

To compare the evolution of the distribution of T_x, we plot a kernel density estimation of T_x for July 15, in two reference periods 1961–1990 and 1991–2020, and the corresponding 5th, 50th and 95th empirical percentiles. The densities are estimated using data in a centered window of 21 days. Fig. 2 shows the plots for Sevilla and Barcelona-Fabra, as an illustration of two different climates. Same plots for San Sebastián and Zaragoza, and plots for January 15 are shown in Figs. S1 and S2 from Supplementary material, respectively. In Sevilla, a clear positive shift of the entire distribution is observed in the last period. However, Barcelona-Fabra shows a different evolution in the center of the distribution and in the extremes, since the increase of the median is higher than in the 5th and 95th percentiles. Figs. S3 and S4 from Supplementary material show the difference between both 30-year periods in the mean value and the standard deviation across days within year for Sevilla, Barcelona-Fabra, Zaragoza, and San Sebastián.

Fig. 3 represents the difference between the period 1991–2020 minus the period 1961–1990 of the median and the 95th percentile of T_x. The percentiles are computed daily using a centered window of 31 days. The different evolution of the differences between the two periods reveals that the long-term trend in the center of the distribution is
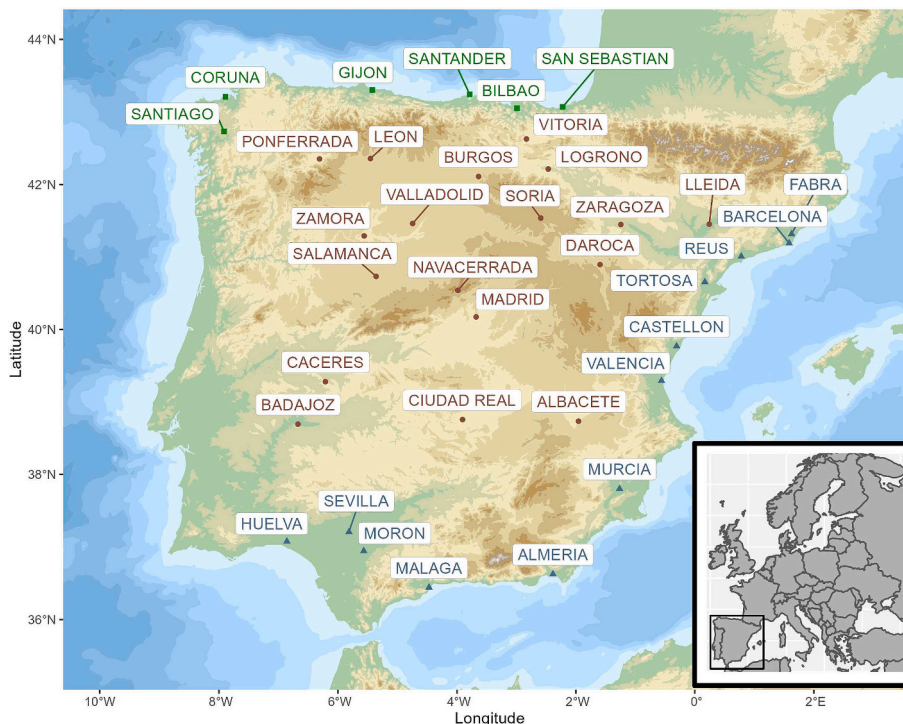
Fig. 1. Map of the 36 Spanish stations, with different climate regions: North Atlantic (green square), Continental (red dot), and Mediterranean (blue triangle).
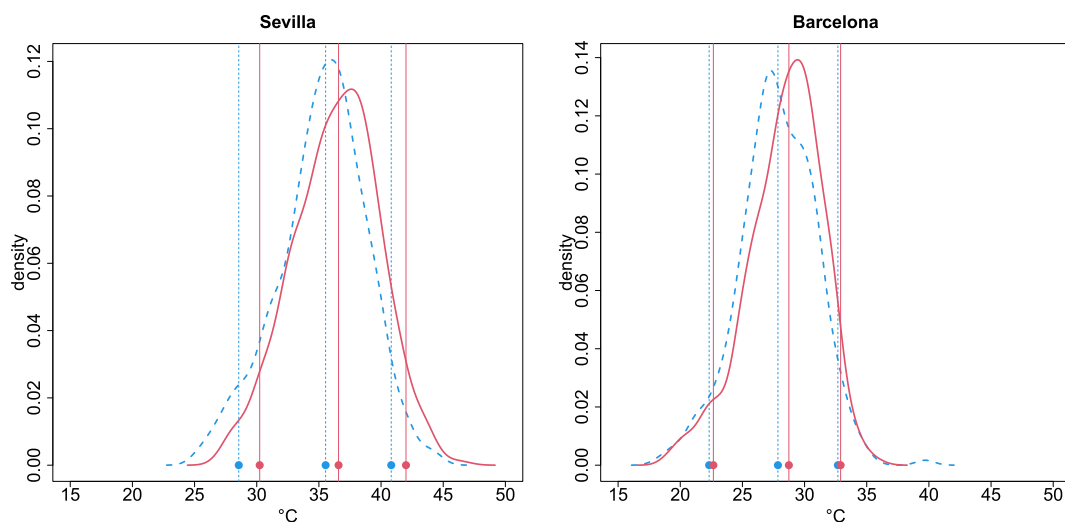

Fig. 2. Kernel density estimation of T_x in Sevilla and Barcelona-Fabra on July 15 in the periods 1961–1990 (blue dashed) and 1991–2020 (red solid). Vertical lines show the 5th, 50th and 95th empirical percentiles.

different to the trend in the extremes. In Barcelona-Fabra, the difference between the two periods is always positive in both the median and the 95th percentile. In Sevilla, the difference is also positive except for the 95th percentile in the early autumn. In Sevilla, the difference in the mean is primarily higher or equal than the difference in the 95th percentile throughout the year. This effect is observed in Barcelona-Fabra only during the summer.

To analyze spatially the different evolution of the median and the 95th percentile, we estimate a linear trend over time (°C/decade) of the median and the 95th percentile of T_x in each location and month using quantile regression. We calculate the difference between those trends and we plot, for each location, the mean of those differences in the 12 months versus elevation (m a.s.l.) at the location, see left plot in Fig. 4. Positive values indicating trends of the median higher than trends in the

95th percentile are observed in 29 locations, in particular in all locations over 500 m a.s.l. Negative values are observed only in 7 locations near the coast: San Sebastián, Bilbao and Santiago (Cantabrian coast), and Tortosa, Barcelona-Fabra, Barcelona-Airport and Almería (Mediterranean coast). This suggests that, in some areas near the coast, the increase of the extremes tends to be stronger than the increase in the mean. More details for Sevilla, Barcelona-Fabra, Zaragoza, and San Sebastián are shown in Fig. S5 from Supplementary material.

To study if the behavior of the trends is different within the year, the right plot in Fig. 4 summarizes, for each month, the differences between trends in the 36 locations using a boxplot. It shows that positive differences are observed in most of the locations (positive boxplot-median) in all the months except, January, April and November. That means in particular that, in most locations, trend in the median is higher than in
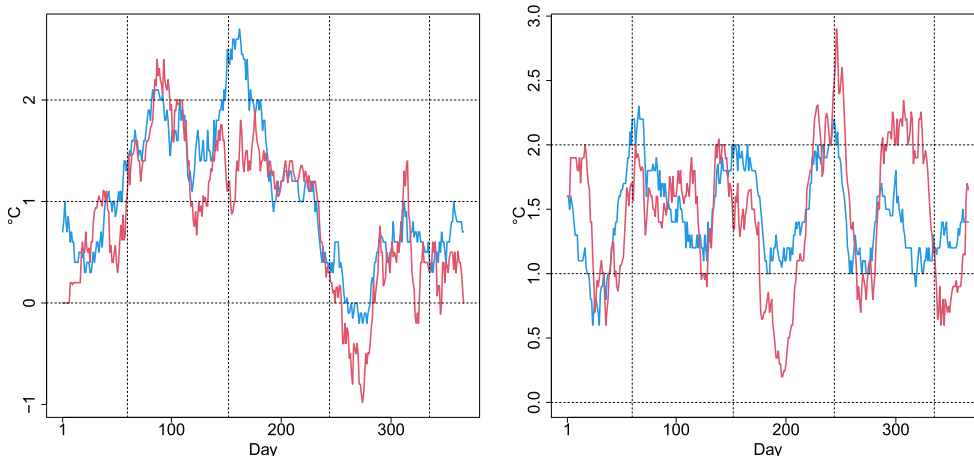
Fig. 3. Differences of the median (blue) and the 95th percentile (red) between the periods 1961–1990 and 1991–2020 of T_x in Sevilla (left) and Barcelona-Fabra (right).
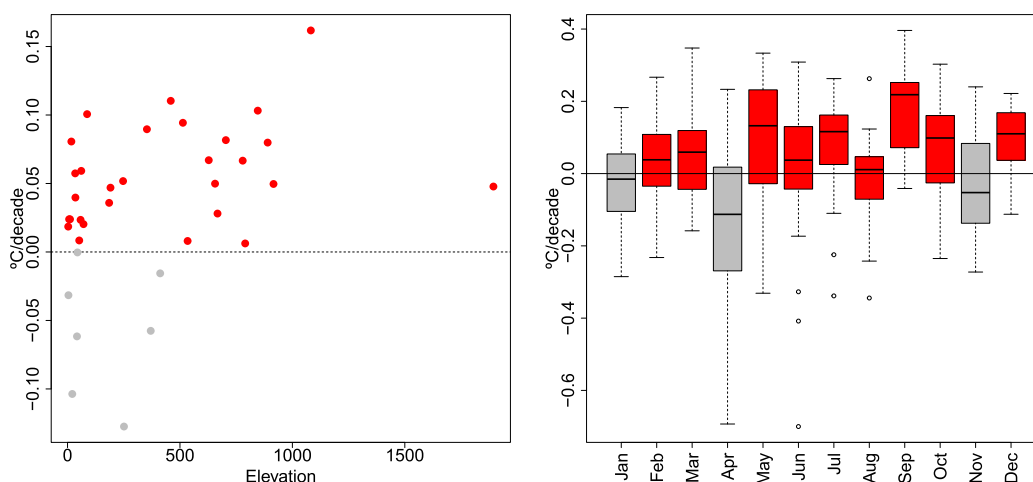

Fig. 4. Mean over the 12 months of the differences between the time trend of the median and the 95th percentile of T_x for each location versus elevation; positive values in red (left). Boxplots by month of the previous differences between trends in the 36 locations; months with positive medians of the differences in red (right).

the 95th percentile in the extended summer period from May to October.

To sum up, the exploratory analysis shows that the time evolution of the central part of the distribution of T_x is not always parallel to the evolution of the tails. Consequently, the estimated effect of global warming in the mean should not be used to describe the effects in records and extreme temperatures. Due to the characteristics of extremes, which are rare by definition, specific statistical tools are needed to analyze their evolution. Another important feature of the series to be considered in their analysis is the spatial and temporal dependence between them.

2.2. Methods

Given a series (X_t), an observation X_i is called an upper record if it has a higher value than the previous observations, this is, if $X_i > \max_{t<i}\{X_t\}$. Analogously, X_i is called a lower record if $X_i < \min_{t<i}\{X_t\}$. All the properties for upper records are valid for lower records because $\min_{t<i}\{X_t\} = -\max_{t<i}\{-X_t\}$. The sequence of record indicator binary variables (I_t), with I_t taking value 1 if a record is observed at time t and 0 otherwise, characterizes the occurrence of records in a series. From this sequence, the number of records up to time t is defined by $N_t = \sum_{i=1}^{t} I_i$.

There are some probabilistic results that characterize the occurrence of records in a series (X_t) of i.i.d. continuous random variables. The first result states that the variables (I_t) are mutually independent with I_t following a *Bernoulli*(p_t) distribution, where the probability of record at time t is

$$p_t = P(I_t = 1) = 1/t, \qquad t = 1, 2, \ldots.$$

This means that in a stationary climate, the probability of observing a record decreases over time but there is always a positive probability of occurrence. Concerning the behavior of the number of records in i.i.d. series, the variables N_t have an asymptotic normal distribution where the expected number of records is $E[N_t] = \sum_{i=1}^{t} p_i = \sum_{i=1}^{t} 1/i$ and the variance $Var[N_t] = \sum_{i=2}^{t} p_i(1-p_i)$.

2.2.1. Record tests for non-stationarity detection

The null hypothesis H_0 of the tests used in this work is that the probability of record at each time t in a series of length T is the probability of record under the stationary condition characterized by i.i.d. series, this is, $1/t$. Under climate change, it is expected that temperatures show an increasing trend and/or an increasing variability. Under those conditions, the probabilities of upper record are higher than in i.i.d. series. Consequently, the following one-sided alternative hypothesis is

considered,

$$H_0 : p_t = 1/t, \quad t = 2, \dots, T.$$
$$H_1 : p_t > 1/t, \quad \text{for at least one} \quad t = 2, \dots, T. \tag{1}$$

The details of the tests can be found in Cebrián et al. (2022). They proposed a family of distribution-free tests to detect deviations from i.i. d. series in the tails of the distribution using the probabilistic properties of the occurrence of records. The underlying idea of these tests is to compare the expected behavior of the occurrence of records in i.i.d. series with the behavior over time of the observed records. Although the tests detect any deviation from i.i.d. series, if the analyzed series is formed by independent observations with no seasonal behavior, deviations from the i.i.d. hypothesis suggest the existence of trends or changing variability, which are the usual features expected under climate change.

The tests assume the availability of $M \geqslant 1$ mutually independent series of length T. These series can be series measured at different spatial points or series obtained from splitting the original data. It is noteworthy that the tests do not require that the M series have the same distribution. From these M series, the series of binary variables $(I_{t1}), (I_{t2}), \dots, (I_{tM})$, and the series of number of records $(N_{t1}), (N_{t2}), \dots, (N_{tM})$ are obtained.

Record tests based on N_T. The most basic statistic is the total number of records in the observed period of length T in the M series:

$$N = \sum_{m=1}^{M} N_{Tm} = \sum_{m=1}^{M} \sum_{t=1}^{T} I_{tm}.$$

Under the null hypothesis H_0, N_T is asymptotically normal when T and/ or M tend to ∞. The power of N is improved by weighting the record indicators according to their position in the series, this is,

$$\mathcal{N} = \sum_{m=1}^{M} \sum_{t=1}^{T} w_t I_{tm}.$$

The idea is to use weights which are increasing functions of t since, given that records become less likely for increasing time, the occurrence of a record at a high t gives more evidence against the null hypothesis H_0. Although different weights can be used, Cebrián et al. (2022) shows that the weights $w_t = t^2/(t-1)$ ($w_1 = 0$) give the locally most powerful unbiased score test. Under the null hypothesis H_0, $\mathcal{N}$ is still asymptotically normal in M. Using this asymptotic distribution, the p-value of the test to study the previous hypothesis is $P(Z > (\mathcal{N}_0 - 0.5 - \mu)/\sigma)$ where μ and σ^2 are the mean and variance of the statistic under the null hypothesis H_0, $\mathcal{N}_0$ is the observed statistic, Z is a standard normal variable and 0.5 is a continuity correction. These tests are implemented with the function `N. test` in `Recordtest`.

Further, the asymptotic normal distribution of the statistics N and $\mathcal{N}$

can be used to compute reference intervals (RI's) of the number of records and of the weighted number of records up to time t. This is useful to implement plots to analyze the evolution of the number of records over time. These graphical tools are implemented with the function `N. plot` (Castillo-Mateo et al., 2023).

Joining information from different types of records. It is noteworthy that four different types of records can be obtained from one series. The upper and lower records in the forward series, $X_1, X_2, \dots, X_T$, and in the backward series obtained when the order of the variables is reversed, $X_T, \dots, X_2, X_1$. Fig. 5 illustrates the times of occurrence of the four types of records in T_x series for June 30 in Barcelona-Fabra. The advantage of considering four types of records is that many different hypotheses can be studied combining them adequately. Further, the power of the tests based on different types of records is higher than those using only one type (Foster and Stuart, 1954; Diersen and Trenkler, 1996; Cebrián et al., 2022). Herein, the superscripts L and B indicate lower records and records in backward series. For example, I_t^L denotes the binary variables for lower records, and I_t^{BL} for lower records in the backward series.

Under the null hypothesis of i.i.d. series, all variables $I_t, I_t^L, I_t^B, I_t^{BL}$ follow a *Bernoulli*$(1/t)$ distribution. Using this property, it is easy to combine the information provided by them in different statistics that allow the study of different alternative hypotheses. It is noteworthy that by considering different types of records in different subseries and in the forward and the backward series, we obtain information from the tails of temperature distribution, not only from the observed record events. Similarly to other approaches to study the extremes of a distribution, such as annual maxima or peak over threshold methods, this information can be used to characterize the behavior of the upper tail, or even both tails, of temperature distribution.

In particular, we can define statistics to analyze:

- The effect of an increasing trend in the upper tail, by considering the upper records in the forward and also in the backward subseries, $\mathcal{N}_{upp} = \mathcal{N} - \mathcal{N}^B$.
- The effect of an increasing trend in the lower tail, by considering now the lower records, $\mathcal{N}_{low} = \mathcal{N}^{BL} - \mathcal{N}^L$.
- The effect of an increasing trend in both tails, by considering both the lower and the upper records in both the forward and the backward subseries, $\mathcal{N}_{both} = \mathcal{N} - \mathcal{N}^L - \mathcal{N}^B + \mathcal{N}^{BL}$.

Note that the sign of the statistic for each type of record is positive or negative according to whether a higher or lower number of records is expected under the considered alternative hypothesis H_1. The effect of an increasing trend in the lower tail implies that the probability of lower record is smaller than $1/t$. In the backward series an increasing trend becomes a decreasing trend so the probabilities of upper and lower re-
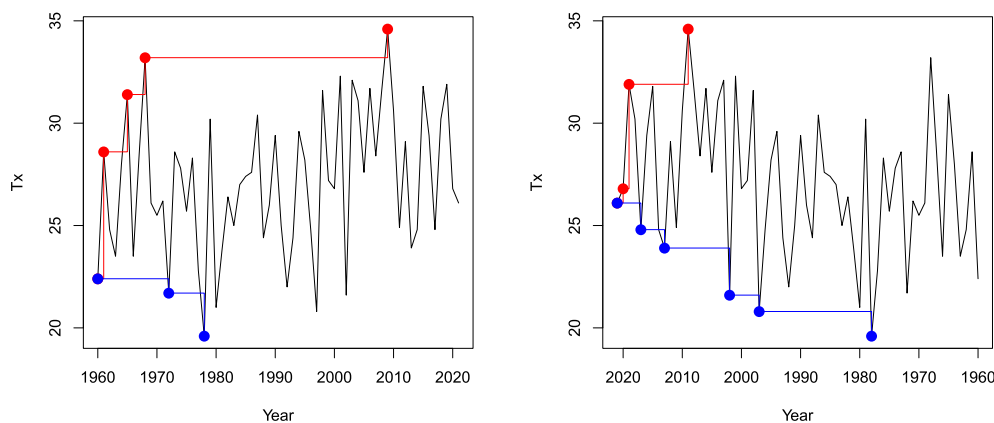

Fig. 5. T_x series for June 30 in Barcelona-Fabra and occurrence times of upper records (red) and lower records (blue) in the forward (left) and backward (right) series.

cords under H_1 are respectively lower and higher than $1/t$; see Fig. 5 as illustration.

We can also analyze the effect of an increasing variability in the tails, taking into account that the increase will lead to a higher number of lower and upper records in the forward series and lower in the backward,

$$\mathscr{N}_{var} = \mathscr{N} + \mathscr{N}^L - \mathscr{N}^B - \mathscr{N}^{BL}.$$

Under H_0, all these statistics have an asymptotic normal distribution with zero mean, and the corresponding p-values can be obtained as usual. The function `foster.test` in `RecordTest` implements these tests. The power of these tests with $M > 12$ and $T > 50$ is high even with trends in location around 1% of the standard deviation but it decreases with lower M (Cebrián et al., 2022).

2.2.2. Record tests for change-point detection

Castillo-Mateo (2022) proposed three distribution-free statistics to detect a change-point if the record occurrence stops being stationary. The statistics test the null hypothesis H_0 in (1) against the two-sided alternative hypothesis

$$H_1 : p_t = 1/t, \quad t = 1, ..., t_0, \qquad \text{and} \qquad p_t \neq 1/t, \quad t = t_0 + 1, ..., T,$$

where t_0 is the change-point. The test statistic for change-point detection in the record occurrence is $\mathscr{K} = \max_{1 \leqslant t \leqslant T} |K_T(t)|$, where

$$K_T(t) = \frac{N_t - E(N_t)}{\sqrt{Var(N_T)}} - \frac{Var(N_t)}{Var(N_T)} \frac{N_T - E(N_T)}{\sqrt{Var(N_T)}}.$$

The change-point $\widehat{t}_0$ is defined as $\widehat{t}_0 = \arg \max_{1 \leqslant t \leqslant T} |K_T(t)|$. Under the null hypothesis H_0, the distribution of $\mathscr{K}$ is Kolmogorov in the limit as $T \rightarrow \infty$. Thus in a two-sided test for a change-point in the record occurrence, the null hypothesis H_0 is rejected if $\mathscr{K} > k_{\alpha/2}$, where $k_{\alpha/2}$ is the upper $\alpha/2$th quantile of the Kolmogorov distribution and α is the significance level for the test; and a significant change-point occurs at time $\widehat{t}_0$ if the null hypothesis H_0 is rejected.

Henceforth, according to Castillo-Mateo (2022), the change-point statistic also considers M series and weights for the record indicators with $w_t = \sqrt{t^2/(t-1)}$ ($w_1 = 0$). The statistic does not follow the Kolmogorov distribution, but the p-value can be estimated using Monte Carlo simulations. Castillo-Mateo (2022) also includes a detailed Monte Carlo analysis of the power and ability of this statistic to detect the actual change-point. The estimator is right-sided biased, which can be usefully interpreted as the time when the underlying process that drives the temperature distribution truly affects the records in the observed data. In other words, the change-point is determined by the occurrence of a record; if no records are observed, there is no change-point. This bias decreases significantly when the number of series M is increased or when the effect of the change becomes greater under the alternative hypothesis H_1. This test is implemented with the function `change. point` in `RecordTest`.

2.2.3. Applying the tests to temperature series

Temperatures, as most environmental time series, measured in an intra-annual temporal scale show seasonal behavior and often serial correlation. Since the previous record tests detect any deviation from i.i. d. series, when the aim is to detect deviations provoked by climate change (increasing trend or changing variability), we need to remove first seasonality and serial correlation from the series. Another practical limitation of the previous tests is that they require M independent series, and independence is a quite restrictive assumption in climate series. This section shows how to prepare the data and a modification of the tests to solve these limitations.

Seasonal and serially correlated series. A common approach in environmental studies to deal with seasonality and serial correlation is to split the series of daily observations into 365 series, so that each series contains the observations from one calendar day across T years:

$$
\begin{array}{c}
\begin{array}{cccc}
\text{Day 1} & \text{Day 2} & \cdots & \text{Day 365}
\end{array} \\
\begin{array}{c}
\text{Year 1} \\
\text{Year 2} \\
\vdots \\
\text{Year } T
\end{array}
\left(
\begin{array}{cccc}
X_{1,1} & X_{1,2} & \cdots & X_{1,365} \\
X_{2,1} & X_{2,2} & \cdots & X_{2,365} \\
\vdots & \vdots & & \vdots \\
X_{T,1} & X_{T,2} & \cdots & X_{T,365}
\end{array}
\right).
\end{array} \tag{2}
$$

The series in each column consists of independent observations with no seasonal behavior and no serial correlation. The resulting 365 series do not necessarily have the same distribution because the distribution in December is probably different from the distribution in August. However, this is not an assumption of the tests, since the probability of record under the null hypothesis H_0 does not depend on the distribution of the series.

The transformation of one series into $M = 365$ subseries is also useful to obtain a high number M of series to apply the tests, and consequently to increase their power. The problem is that, given that the series are measured in consecutive times (days in this case), they will probably show a strong dependence between them, and the tests require independent series. If the number of dependent series is high enough, the simplest option is to extract among them a subset of M independent series. This can be done, selecting series separated by a fixed distance (for example, 10 days) or applying an approach based on Pearson correlation tests (Castillo-Mateo et al., 2023). A second option is to use tests that allow dependent series, as described below.

Tests for a set of M dependent series. The asymptotic distribution of the record statistics relies on the assumption that the M studied series are independent. To avoid this restriction, we propose an alternative approach where the p-values of the statistics defined in Section 2.2.1 are computed using permutation techniques.

Permutation tests only rely on the assumption of exchangeability under the null hypothesis (Welch, 1990). A sample is exchangeable if any permutation of it has the same joint probability distribution. In the record tests, there is a sample of $t = 1, ..., T$ observations of a vector of M variables $(X_{t1}, X_{t2}, ..., X_{tM})$. Under the null hypothesis, the T observations of the vector $(X_{t1}, X_{t2}, ..., X_{tM})$ are independent with the same multivariate distribution, so that permutations of rows, see the data structure in (2), are exchangeable. Note that the M variables in the vector may be correlated between them or may have different marginal distributions, the only required assumption is that the multivariate distribution of the vector is the same over the T years.

If observations are exchangeable under H_0, and all the possible permutations are considered to compute the p-value, the resulting test yields the exact significance level. In general, a random sample of possible permutations must be used because the total number of different permutations is too high. The p-value is computed as the proportion of samples whose statistic value is greater or equal than the observed statistic. A number of 10,000 permutations gives a good approximation to the p-value.

3. Results

3.1. Analysis of daily temperature series in one location

This section aims to show how the record tests previously defined work, and the different hypotheses that can be analyzed with them. To that end, a detailed analysis is presented using the longest series available, the daily maximum temperature series in Barcelona-Fabra from 1914 to 2021.

The observed number of upper and lower records in the period 2001–2021 in the 365 daily series is 224 and 23, respectively. The number of upper records is 10 times the number of lower records while, in a stationary climate, the expected number of records in both cases is

the same and equal to 78.5. In this case, where the observed period is $T = 108$ years long, even this simple exploratory analysis suggests the effects of an increasing trend, but in many series, the observed period is shorter and more formal inference tools are needed to evaluate the effect of global warming.

Before applying the record tests, given the seasonal behavior and serial correlation of daily temperature in Barcelona-Fabra, the processing tools described in Section 2.2 have to be used. The resulting final data is a subset of $M = 46$ uncorrelated series of length $T = 108$ with no seasonal behavior and no serial correlation.

We start by analyzing graphically the evolution of the number of records over time, to detect possible deviations from the stationary behavior and to identify when they appear. To that end, the mean, in M series, of the number of records up to time t, $\overline{N}_t = \sum_{m=1}^{M} \sum_{i=1}^{t} I_{im}/M$, for $t = 1, \ldots, T$, is plotted versus t together with RI's for the number of records in i.i.d. series. An analogous plot is obtained for the weighted number of records $\overline{\mathcal{N}}_t = \sum_{m=1}^{M} \sum_{i=1}^{t} w_i I_{im}/M$. Fig. 6 summarizes the evolution of the four types of records, in both cases. In the four types of records there is a significant evidence of a non-stationary behavior in the upper tail (forward and backward upper records), and even earlier in the lower tail (forward and backward lower records). The evidence is clearer, and consequently it is detected earlier, in the backward series and using weights.

To obtain more formal conclusions about the effect of global warming in extreme temperature, we can apply the record tests and study different hypotheses. To analyze the behavior of the upper records, we apply the test $\mathcal{N}$ with the alternative that the probabilities of the upper records are higher than in i.i.d. series. The p-value is $1.8e-05$, so that the null hypothesis is rejected at any usual significance level. To analyze the behavior of the upper extreme temperatures, not only the records, we apply the test $\mathcal{N}_{upp}$; the null hypothesis is also rejected with a p-value $8.4e-7$. To study the lower tail and both tails simultaneously, we have to use the statistics $\mathcal{N}_{low}$, and $\mathcal{N}_{both}$, respectively, they yield p-values equal to $7.5e-7$, and $8.8e-12$. We conclude that there is evidence of the effect of an increasing trend in both the upper and the lower tail and this evidence is stronger when information of both tails is joined. Since effects of an increasing trend are detected in both tails, it can be expected that the variability is not increasing; this is confirmed with the p-value of $\mathcal{N}_{var}$ test, 0.5066, that leads to not reject the null hypothesis. Finally, the change-point test detects a significant change-point at time $\hat{t}_0 = 64$ (1977) with a p-value 0.0029.

3.2. Spatial analysis of temperatures

Climate change is a spatial phenomenon, so the interest lies in the analysis of the behavior of the temperature records and extremes over the peninsular Spain. To make fair comparisons of the behavior across different locations, the same period of time should be analyzed in all of them, so that the set of 36 series from 1960 to 2021 introduced in Section 2.1 is considered.

It is noteworthy that since 36 series are analyzed, corrections for multiple comparisons should be applied. Given that the test statistics from nearby locations will be possibly dependent, and although this dependence may become negligible between the farthest locations, we opted for applying the conservative Benjamini-Yekutieli method (Benjamini and Yekutieli, 2005). This method controls the expected proportion of false rejections and can be applied to p-values from dependent statistics.

3.2.1. Analysis of daily temperature series

The first hypotheses that we study are the effect of an increasing trend in the upper records and in the upper tail using statistics $\mathcal{N}$ and $\mathcal{N}_{upp}$, respectively. Fig. 7 shows the maps summarizing the results. These maps show with points in a red color scale the locations where the p-value is lower than $\alpha = 0.1$, and in a blue color scale, higher values. For the upper tail there are 29, 25, and 22 series with significant p-values at level $\alpha = 0.10, 0.05$ and 0.01 respectively, and for the upper records, 23, the same 23, and 18. There are evidences of the effect of an increasing trend in both upper records and upper tail in most of the studied locations; as expected, there are more evidences in the upper tail, where more information is available. Only in the North area, and specially in the Cantabrian coast, there is no significant evidence of the effects of an increasing trend. The effects are weaker also in other locations in the coast, for example Huelva, in the South, and Castellón and nearby locations in the Levante coast.

Analysis within the year. It is known that effects of climate change may not be homogeneous over the year, so that it is necessary to analyze the potential effects of a trend in shorter periods of time within the year, seasons or even months. In those cases, M, the number of independent series available is usually small, so that the power of the asymptotic tests is low with weak trends. Then, it is preferable to apply the tests based on permutations described in Section 2.2.1, using all the daily series available in the considered period. Observations from subsequent days are dependent, but permutation tests allow for the analysis of dependent columns in matrices like (2). Fig. 8 shows the p-values from the tests to study the effect in the upper tail of the daily maximum temperature in each season: winter (DJF), spring (MAM), summer (JJA) and autumn (SON). It is clear that the effects are not homogeneous over the year: autumn is the season where the effects are weaker, being clearly significant only in the North part of the Mediterranean coast. On the other hand, in summer, significant evidences are found all over Spain except in Cantabrian coast and some nearby locations. Note that in this area, only in winter, there are some weak evidences of the effect of the trend in the extremes. The analysis of the upper records shows a similar pattern to the upper extremes, although the p-values tend to be slightly higher. In particular, in autumn, significant evidences at a $\alpha = 0.05$ significance
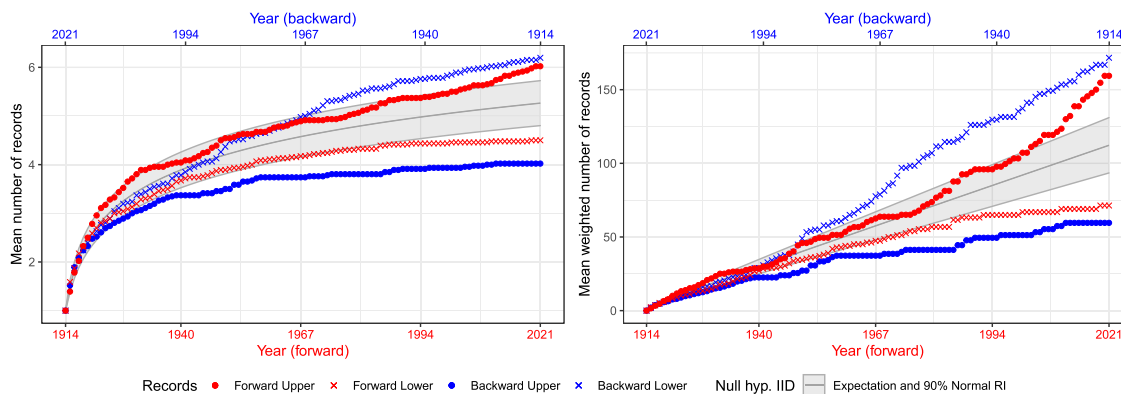

Fig. 6. Mean (left) and mean weighted (right) number of records up to time t versus t, in daily maximum temperature in Barcelona-Fabra (1914–2021).
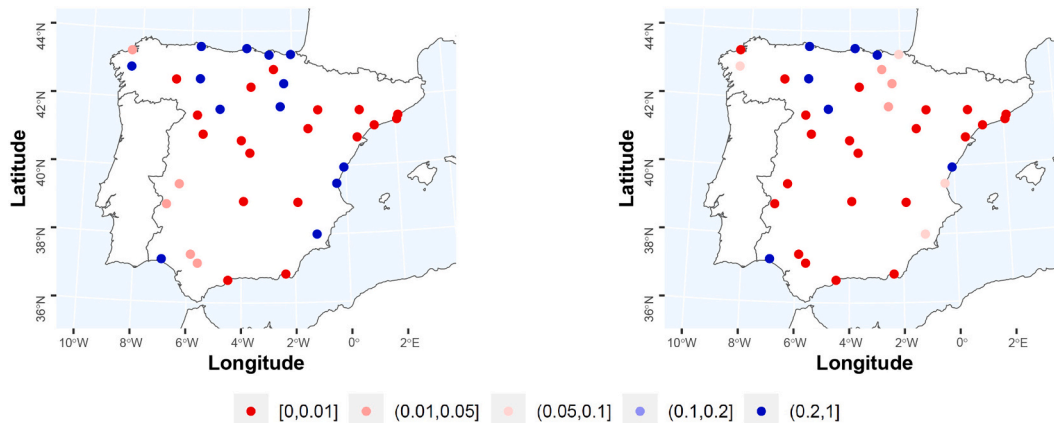
Fig. 7. P-values from the analysis of the effect of a trend in the upper records (left) and in the upper tail (right) in daily T_x series during the entire year in Spanish locations.

level are found only in some areas in the Mediterranean coast, see right plot in the last row in Fig. 8.

Even within seasons, the behavior may not be homogeneous, so that the same tests are applied by month (maps not shown). Concerning the upper extremes, no significant effect of a positive trend is found in any location, in February, March and November, and in September only in two Mediterranean locations (Barcelona-Fabra and Tortosa). On the other hand, June, July and August are the months with largest areas of Spain with evidences of the effects of a trend, both in the upper tail and the upper records. As an example, the p-values from the upper tail in June are shown in Fig. 8.

Regional analysis. Since global warming is a spatial phenomenon, it is of interest to analyze its effects at a regional scale, by joining information from all the series available in a region. The proposed tests can be used to study this type of hypothesis, using the permutation approach if it cannot be assumed that the series available in a region are independent, as it is often the case. Note that the regional analysis will be more powerful since joining series from different locations, more information is provided to the tests.

Here, we consider three regions with different climate characteristics: the North Atlantic, the Continental and the Mediterranean areas. See in Fig. 1 the locations by color in each region. The effect of a positive trend in the upper tail and in the upper records is significant at a significance level $\alpha = 0.05$ in all the regions in the four seasons. Table 1 shows the years identified by the change-point test to detect when the upper record occurrence is significantly different from the occurrence in i.i.d. series, in each season and in the whole year. The behavior is quite homogeneous all over Spain, with slight differences between the three regions, but relevant differences appear between seasons. Winter is the season where the evidences of non-stationarity in the upper records appear first, around 1975, while in autumn the change-points occur 30 year later. In spring and summer the non-stationarity is detected from around 1985.

The regional analysis is also developed by month, and the resulting p-values are summarized in Fig. 9. Again, the different p-values are shown in a red–blue scale according to their significance. In this case, in each region, the 12 p-values corresponding to one month are corrected using the Benjamini-Yekutieli approach. The results show that February, September and November are the months where the upper extremes and upper records are less affected by global warming, with no significant effect in any of the climate regions. In contrast, April and October are the months where a significant behavior is observed all over Spain. The Mediterranean region is the area where the number of months with strong evidences is higher, in all the year except the three months previously mentioned, while the North Atlantic region is the area where less evidences of non-stationarity occur; only in April and October a

significant behavior is detected. Concerning the upper records, the same pattern is observed, although the effects are slightly weaker.

3.2.2. Analysis of annual maximum and annual minimum of T_x

In this section, the behavior of the records and both tails of the annual maxima and the annual minima of T_x, the daily maximum temperature, are analyzed. Unlike the case of daily series, only one series per location is available for this analysis. Given that with $M = 1$, the power of the test is low for weak trends, only the regional analysis is carried out. Table 2 summarizes the p-values of the tests to assess the effect of an increasing trend in the upper tail, the upper records, the lower tail and the lower records in the annual maximum and in the annual minimum of T_x, in the three climate regions. The effect in the tails, both in upper and lower, is significant at a significance level $\alpha = 0.05$ in all the regions. In the records, the effect is slightly weaker, specially in the lower records of the annual maxima, where the p-values are not significant at a $\alpha = 0.05$ level in any region. This implies that there are not enough evidence to state that the occurrence of lower records is not lower than $1/t$ in that signal, and could suggest an increase of variability. However, the tests to detect variability, based on statistic $\mathcal{N}_{var}$, are not significant at any usual significance level, with p-values 0.16, 0.53 and 0.74; this means that there is no significant evidence of an increase of variability in annual maximum temperature in any of the three considered regions.

3.3. Records in T_x and daily temperatures at geopotential levels

The aim of this section is to analyze the occurrence of records in temperatures at geopotential levels, 850, 700, 500 and 300 hPa, to identify whether it is related to the occurrence of records in the surface temperature.

3.3.1. Data and exploratory analysis

The considered database includes the daily temperature series at 850, 700, 500 and 300 hPa levels, measured at 12:00 from 1960 to 2021, in the 176 points from the $1° \times 1°$ grid 35°N–45°N and 10°W–5°E that covers the Iberian Peninsula. This database is obtained from the 5th generation ECMWF reanalysis database ERA5 available in Climate Data Store (CDS) of Copernicus Climate Change Service (C3S) (Hersbach et al., 2023), see map in Fig. S6 from Supplementary material.

As in surface temperature signals, the effects of global warming are clear in the mean evolution of air temperature at 850, 700, 500 and 300 hPa geopotential levels. The mean temperature at the four levels shows an increase in the period 1960–2021, but the magnitude of that increase and the spatial variability are different. The effect of relief on the spatial behavior is more clear on temperature at 850 hPa level, where the
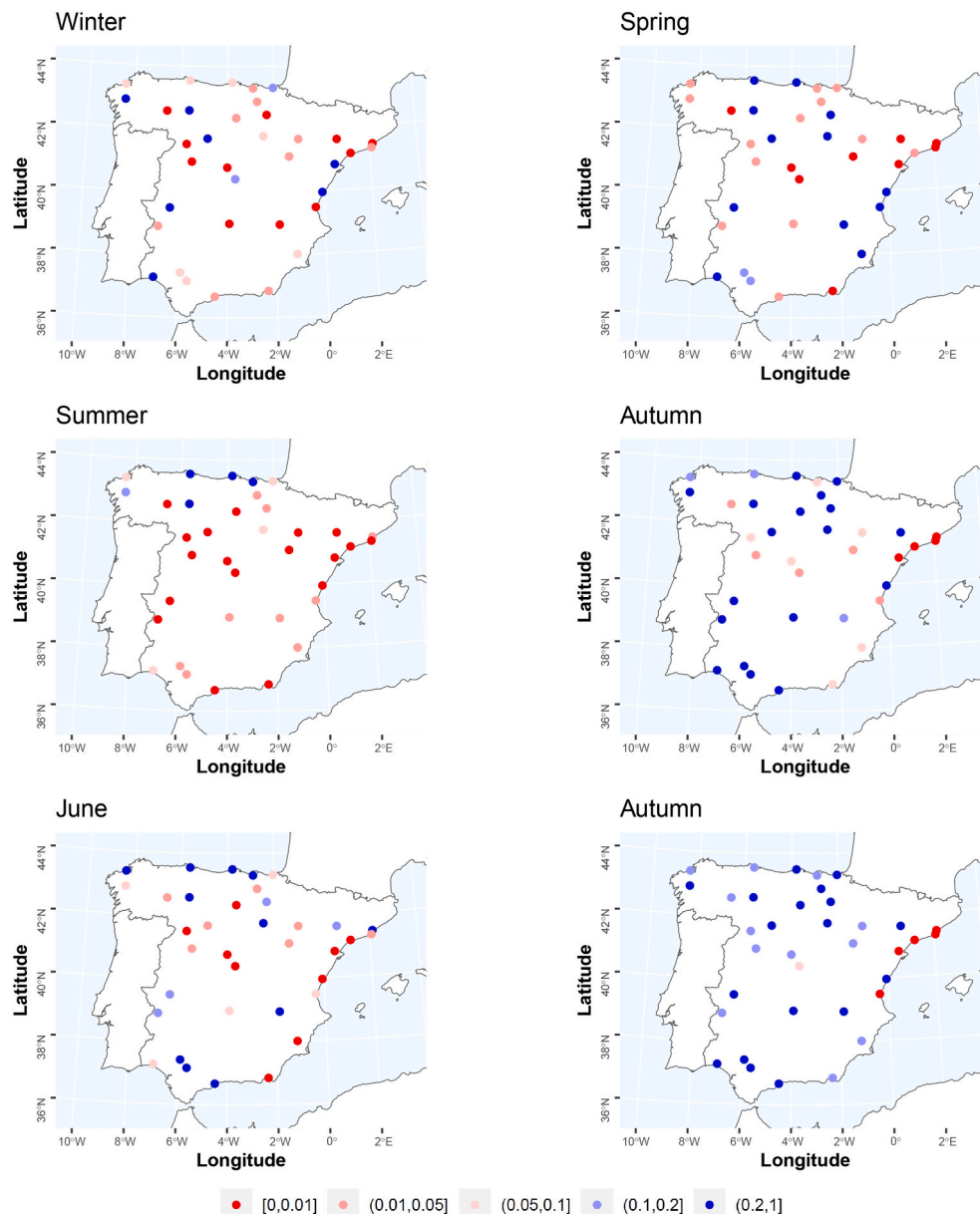
Fig. 8. P-values from the analysis of the effect of an increasing trend in the upper extremes in T_x by season in 36 Spanish locations (first and second row). Same map only for June (left plot last row) and for the upper records in autumn (right plot last row).

Table 1
Time points identified by the change-point test to detect when the upper record occurrence is significantly different from the occurrence in i.i.d. series, in each season and in the whole year.

Region	Series				
	Year	winter	spring	summer	autumn
North Atlantic	1976	1976	1980	1988	2010
Continental	1986	1973	1988	1986	2003
Mediterranean	1979	1973	1986	1986	2008

increasing trend over time of temperature varies from 0.20 in the North-West to 0.35°C/decade in the center of the Iberian Peninsula. At upper levels, the spatial variability is lower with trends from around 0.20 to 0.26°C/decade, with the highest increases in the South-West. Maps of the estimated trend at each geopotential level are shown in Fig. S7 from Supplementary material.

Working with the original time series, an absolute record is defined

to be a value that exceeds all previous values in the series. Considering the 704 daily series of temperatures, one associated to each of the four geopotential levels in each of the 176 grid-points, the median number of absolute records since 1962 is 6. Given that four air temperature series are available at each grid-point we identify simultaneous records, based on the idea of compound events (Zscheischler et al., 2020), that represent the occurrence of a highly extreme situation in the geopotential temperature series. A "compound record" at a grid-point in a day is defined when simultaneous absolute records occur in at least 2 geopotential level series. Table 3 on the left summarizes dates since 1971 with compound records in at least 6 points of the grid. It shows on the right dates when the absolute record occurs simultaneously in more than two T_x series. Simultaneous compound records are temporally close to simultaneous absolute records in T_x series, for example, the episode 13–16 August 1987, or the unprecedented heatwave in 26–29 June 2019, that caused record-breaking high temperatures in about one-third areas of Europe (Sousa et al., 2019; Xu et al., 2021). An exploratory analysis of these compound records is shown in Tables S2 and S3 in the
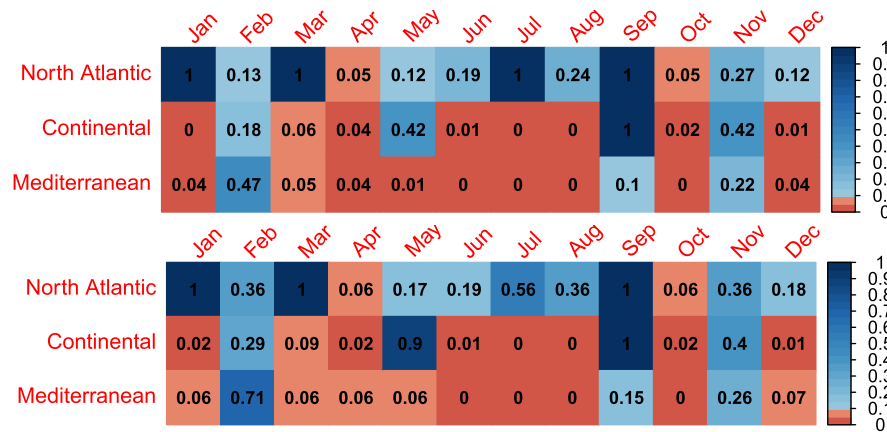
Fig. 9. P-values from the analysis of the effect of a positive trend in the upper tails (top) and upper records (bottom) in T_x by month in the three climate regions.

Table 2

P-values of the effect of an increasing trend in the upper tails (UT), the upper records (UR), the lower tail (LT) and the lower records (LR) in annual maximum and minimum of T_x.

Region	Annual max.				Annual min.			
	UT	UR	LT	LR	UT	UR	LT	LR
North Atlantic	0.000	0.002	0.023	0.104	0.007	0.009	0.013	0.013
Continental	0.000	0.000	0.021	0.121	0.041	0.067	0.005	0.002
Mediterranean	0.011	0.007	0.012	0.056	0.008	0.013	0.039	0.013

Table 3

Dates from 1971 with compound records in at least 6 points of the grid. Right column shows next dates with absolute record simultaneously in more than two T_x series.

Compound event	Absolute record in more than 2 stations
1980–8–2	
1983–7–30	
1987–8–13, 1987–8–16	1987–8–12, 1987–8–13
2019–6–26, 2019–6–27	2019–6–29
2021–7–10	

Supplementary material.

3.3.2. Time evolution of records in daily temperature at geopotential levels

Our aim in this section is to compare the time evolution of the number of records and the time where non-stationary behavior is detected in surface and geopotential level temperature series. To that end, the weighted cumulative number of records, $\overline{\mathcal{N}}_t$, in the observed temperature series T_x in a location and in temperature series at a given geopotential level in the four grid-points around that location, are plotted in the same graph. In the same plot, we show the years where the estimated change-point occurs in each of the five series. Fig. 10 shows, as an illustration, the graph for Barcelona-Fabra in July for the four geopotential levels. At 850 and 500 levels, the time evolution of the records and the estimated change-points are very similar in T_x and in the four grid-points and, at 700 hPa, in two of the grid-points. However, the estimated change-point at 300 hPa occurs later, around 1998. The same plots for Barcelona-Fabra in January and Sevilla in July are shown in Figs. S8 and S9, respectively; and a similar plots based on the 36 series is shown in Fig. S10 in the Supplementary material.

4. Discussion and conclusions

This work aims to analyze and evaluate the effects of global warming in the upper tail (extreme and record-breaking events) of daily maximum temperature series over peninsular Spain. Since, an

exploratory analysis shows that the time evolution of the central part of the distribution of daily maximum temperature is not parallel to the evolution of the tails, it is necessary to perform a specific analysis of records and extreme temperatures. Here, we propose the use of statistical tools and record tests described in Cebrián et al. (2022) and Castillo-Mateo (2022) to assess the effect of global warming in the upper tail in different temperature signals in Spain, at different time and spatial scales. The analysis of short periods of time, for example months, and regional analysis imply the study of series measured in close days and/or in close locations, which are correlated. Given that the previous tests require independent series, we propose a modification based on a permutation approach to compute the p-values, which can be used with correlated series. Using these tests, we evaluate and compare the effects of climate change in the extremes and the record-breaking events all over Spain, in different Spanish regions, and in different periods of the year. All the record tests and statistical tools used in the analysis are freely available in the R package RecordTest. The proposed tests can be applied to analyze the evolution of extremes and records in observed series and also in gridded data obtained from reanalysis or Earth System Models, taking into account the spatial dependence between them.

As far as we know, this is the first study of temperature records developed in Spain that does not focus on a particular event. Sousa et al. (2019) described the intense heatwaves striking the Iberian Peninsula in early August 2018 and late June 2019. They found the Saharan air intrusions as a relevant mechanism for Iberian heatwaves. Their results are in agreement with the compound events and absolute records on late June 2019 observed in our work. Also there are some analyses about extremes, for example, Serrano-Notivoli et al. (2022) studied the heatwaves in Spain with a high-resolution gridded daily temperature dataset 1940–2014. They found a tipping point in the early 1980s from which heatwaves became more frequent, this is compatible with our results in Table 1 for summer. In other regions of the world, some authors found similar results about temperature records. McBride et al. (2022) studied the record-breaking frequencies of the highest daily maximum temperature in South Africa for the 1951–2019 period. As in this work, they found that the number of records is higher than the theoretically expected in a stationary climate. They also found that mainland stations
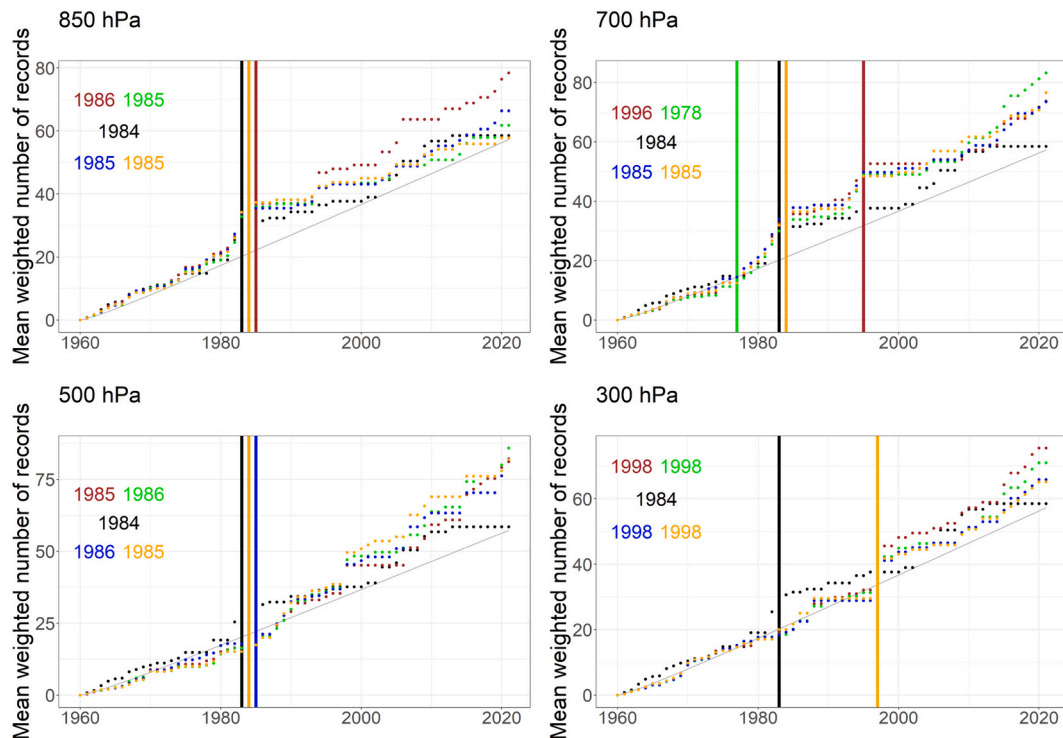
Fig. 10. Mean weighted number of records up to time t in July versus t, in T_x in Barcelona-Fabra (black) and in air temperature series in the four grid-points around Barcelona (NW: brown, NE: green, SW: blue and SE: orange) at 850, 700, 500 and 300 hPa. The years where the estimated change points occur in each of the four grid-points are shown, and plotted as vertical lines.

are more affected by the increasing number of records than stations close to the coast.

Our results yield the following conclusions about the evolution of extreme temperatures over Spain:

- Significant evidences of the effect of an increasing trend in the occurrence of upper extreme and record-breaking events in daily maximum temperature have been found in most of Spain, when the whole year is studied. Only in the North and the Cantabrian Coast areas the effects are not significant.
- The effects are heterogeneous within the year. Autumn is the season where the effects are weaker; only on the coast of Catalonia, the evidences are significant. On the other hand, summer is the season where the effects are stronger: in all the locations, except 5 locations in the North coast area, significant effects are found.
- Concerning the spatial variability, the Mediterranean region is the area where the number of months with strong evidences is higher, while the North Atlantic region is the area where less evidences of non-stationarity occur.
- No evidence of an increase of variability in daily maximum temperature has been found.

Many of these results are in agreement with those found in other areas. In Romania, Busuioc et al. (2015) detected significant increasing trends for the temperature extremes in all seasons, except for autumn; they found the highest increasing rate in summer, and that it is possibly associated with the the Atlantic multidecadal oscillation. Tošić et al. (2023) also found that, in Serbia, there are significant changes in temperature extremes consistent with warming, specially in summer, and that they have a highly positive correlation with the East Atlantic pattern. Sánchez-Benítez et al. (2020) summarized the connection between Iberian heat waves, and atmospheric circulation patterns in four weather regimes, and found that those heatwaves are associated with ridge conditions in western Europe.

The results from this work suggest that more analyses in a spatio-temporal framework about the occurrence of records and extreme temperature events in Spain are needed, and they can open new research lines. In particular, more research to study the relation between records and the atmospheric situation expressed by air temperature at different geopotential levels is of interest, and also with teleconnection patterns to find plausible physical mechanisms. It is also of interest to develop space–time models for the occurrence of records. These models should include long-term trends, space–time dependence in the tails and random effects, so that Bayesian models can provide an adequate framework. Models including atmospheric variables as covariates would be also useful as statistical downscaling models, and they could be used to obtain future projections of the behavior of records under different climate scenarios.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

To enhance reproducibility, the methodological tools used in this paper are provided in the R package RecordTest (Castillo-Mateo et al., 2023). Specific data and code scripts in R are available upon request to the authors.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.atmosres.2023.106934.

References

Anderson, A., Kostinski, A., 2011. Evolution and Distribution of Record-Breaking High and Low Monthly Mean Temperatures. J. Appl. Meteorol. Climatol. 50 (9), 1859–1871. https://doi.org/10.1175/JAMC-D-10-05025.1.

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 1998. Records, Wiley Series in Probability and Statistics. John Wiley & Sons, New York. https://doi.org/10.1002/9781118150412.

Benestad, R.E., 2003. How Often Can We Expect a Record Event? Clim. Res. 25 (1), 3–13. https://doi.org/10.3354/cr025003.

Benestad, R.E., 2004. Record-Values, Nonstationarity Tests and Extreme Value Distributions. Global Planet. Change 44 (1–4), 11–26. https://doi.org/10.1016/j.gloplacha.2004.06.002.

Benestad, R.E., 2008. A simple test for changes in statistical distributions. Eos 89 (41), 389–390. https://doi.org/10.1029/2008EO410002.

Benjamini, Y., Yekutieli, D., 2005. False discovery rate–adjusted multiple confidence intervals for selected parameters. J. Am. Stat. Assoc. 100 (469), 71–81. https://doi.org/10.1198/016214504000001907.

Bunge, J., Goldie, C.M., 2001. Record sequences and their applications. Handb. Stat. 19, 277–308. https://doi.org/10.1016/S0169-7161(01)19012-7.

Busuioc, A., Dobrinescu, A., Birsan, M.-V., Dumitrescu, A., Orzan, A., 2015. Spatial and temporal variability of climate extremes in Romania and associated large-scale mechanisms. Int. J. Climatol. 35 (7), 1278–1300. https://doi.org/10.1002/joc.4054.

Castillo-Mateo, J., 2022. Distribution-free changepoint detection tests based on the breaking of records. Environ. Ecol. Stat. 29 (3), 655–676. https://doi.org/10.1007/s10651-022-00539-2.

Castillo-Mateo, J., 2023. RecordTest: Inference Tools in Time Series Based on Record Statistics, URL:https://CRAN.R-project.org/package=RecordTest, R package version 2.2.0.

Castillo-Mateo, J., Cebrián, A.C., Asín, J., 2023. RecordTest: An R package to analyse non-stationarity in the extremes based on record-breaking events. J. Stat. Softw. 106 (5), 1–28. https://doi.org/10.18637/jss.v106.i05.

Cebrián, A.C., Castillo-Mateo, J., Asín, J., 2022. Record tests to detect non-stationarity in the tails with an application to climate change. Stoch. Env. Res. Risk Assess. 36, 313–330. https://doi.org/10.1007/s00477-021-02122-w.

Chazarra-Bernabé, A., Lorenzo Mariño, B., Romero Fresneda, R., Moreno García, J.V., 2022. Evolución de los climas de Köppen en España en el periodo 1951–2020, Notas técnicas de AEMET 37, doi: 10.31978/666-22-011-4.

Cos, J., Doblas-Reyes, F., Jury, M., Marcos, R., Bretonnière, P.-A., Samsó, M., 2022. The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections. Earth Syst. Dyn. 13 (1), 321–340. https://doi.org/10.5194/esd-13-321-2022.

Coumou, D., Rahmstorf, S., 2012. A Decade of Weather Extremes. Nat. Clim. Change 2, 491–496. https://doi.org/10.1038/nclimate1452.

Coumou, D., Robinson, A., Rahmstorf, S., 2013. Global Increase in Record-Breaking Monthly-Mean Temperatures. Clim. Change 118 (3–4), 771–782. https://doi.org/10.1007/s10584-012-0668-1.

Diersen, J., Trenkler, G., 1996. Records Tests for Trend in Location. Statistics 28 (1), 1–12. https://doi.org/10.1080/02331889708802543.

Diffenbaugh, N.S., Giorgi, F., 2012. Climate change hotspots in the CMIP5 global climate model ensemble. Clim. Change 114, 813–822. https://doi.org/10.1007/s10584-012-0570-x.

Feller, W., 1991. An Introduction to Probability Theory and its Applications, second ed., vol. 2. John Wiley & Sons, New York.

Fischer, E.M., Sippel, S., Knutti, R., 2021. Increasing probability of record-shattering climate extremes. Nat. Clim. Change 11 (8), 689–695. https://doi.org/10.1038/s41558-021-01092-9.

Foster, F.G., Stuart, A., 1954. Distribution-Free Tests in Time-Series Based on the Breaking of Records. J. R. Stat. Soc. B 16 (1), 1–22.

Gouet, R., Lafuente, M., López, F.J., Sanz, G., 2020. Exact and asymptotic properties of δ) records in the linear drift model. J. Stat. Mech: Theory Exp. 2020 (10), 103201 https://doi.org/10.1088/1742-5468/abb4dc.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J.-N., 2023. ERA5 hourly data on pressure levels from 1940 to present, Tech. Rep. Copernicus Climate Change Service (C3S), Climate Data Store (CDS). https://doi.org/10.24381/cds.bd0915c6.

Kendall, M., Gibbons, J.D., 1990. Rank Correlation Methods, A Charles Griffin Title, fifth ed. Oxford University Press, New York.

Klein Tank, A.M.G., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van

Engelen, A.F.V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L.V., Petrovic, P., 2002. Daily Dataset of 20th-Century Surface Air Temperature and Precipitation Series for the European Climate Assessment. Int. J. Climatol. 22 (12), 1441–1453. https://doi.org/10.1002/joc.773.

Kysely, J., 2010. Recent severe heat waves in central Europe: how to view them in a long-term prospect? Int. J. Climatol. 30 (1), 89–109. https://doi.org/10.1002/joc.1874.

Lionello, P., Scarascia, L., 2018. The relation between climate change in the Mediterranean region and global warming. Reg. Environ. Change 18, 1481–1493. https://doi.org/10.1007/s10113-018-1290-1.

Mann, H.B., 1945. Nonparametric Tests Against Trend. Econometrica 13 (3), 245–259. https://doi.org/10.2307/1907187.

McBride, C.M., Kruger, A.C., Dyson, L., 2022. Trends in probabilities of temperature records in the non-stationary climate of South Africa. Int. J. Climatol. 42 (3), 1692–1705. https://doi.org/10.1002/joc.7329.

MedECC, 2020. Climate and Environmental Change in the Mediterranean Basin – Current Situation and Risks for the Future. First Mediterranean assessment report, Tech. Rep. Union for the Mediterranean, Plan Bleu, UNEP/MAP, Marseille, France. https://doi.org/10.5281/zenodo.4768833.

Meehl, G.A., Tebaldi, C., Walton, G., Easterling, D., McDaniel, L., 2009. Relative increase of record high maximum temperatures compared to record low minimum temperatures in the US. Geophys. Res. Lett. 36 (23) https://doi.org/10.1029/2009GL040736.

Om, K.-C., Ren, G., Kim, K.-H., Pak, Y.-I., Jong, S.-I., Kil, H.-N., 2022. Observed trends in extreme temperature events over northern part of the Korean Peninsula during 1960–2019 and a comparative overview. Atmos. Res. 270, 106061 https://doi.org/10.1016/j.atmosres.2022.106061.

Peña-Angulo, D., Gonzalez-Hidalgo, J.C., Sandonís, L., Beguería, S., Tomas-Burguera, M., López-Bustins, J.A., Lemus-Canovas, M., Martin-Vide, J., 2021. Seasonal temperature trends on the Spanish mainland: A secular study (1916–2015). Int. J. Climatol. 41 (5), 3071–3084. https://doi.org/10.1002/joc.7006.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL:https://www.R-project.org/.

Redner, S., Petersen, M.R., 2006. Role of global warming on the statistics of record-breaking temperatures. Phys. Rev. E 74 (6), 061114. https://doi.org/10.1103/PhysRevE.74.061114.

Saddique, N., Khaliq, A., Bernhofer, C., 2020. Trends in temperature and precipitation extremes in historical (1961–1990) and projected. Stoch. Env. Res. Risk Assess. 34 (10), 1441–1455. https://doi.org/10.1007/s00477-020-01829-6.

Salameh, A.A., Gámiz-Fortis, S.R., Castro-Díez, Y., Abu Hammad, A., Esteban-Parra, M.J., 2019. Spatio-temporal analysis for extreme temperature indices over the Levant region. Int. J. Climatol. 39 (15), 5556–5582. https://doi.org/10.1002/joc.6171.

Sánchez-Benítez, A., Barriopedro, D., García-Herrera, R., 2020. Tracking Iberian heatwaves from a new perspective. Weather Clim. Extrem. 28, 100238 https://doi.org/10.1016/j.wace.2019.100238.

Sánchez-Lugo, A., Berrisford, P., Morice, C., Nicolas, J.P., 2019. Global surface temperature [in "State of the climate in 2018"]. Bull. Amer. Meteor. Soc. 100 (9), 11–14. https://doi.org/10.1175/2019BAMSStateoftheClimate.1.

Schär, C., Vidale, P.L., Lüthi, D., Frei, C., Häberli, C., Liniger, M.A., Appenzeller, C., 2004. The role of increasing temperature variability in European summer heatwaves. Nature 427 (6972), 332–336. https://doi.org/10.1038/nature02300.

Serrano-Notivoli, R., Lemus-Canovas, M., Barrao, S., Sarricolea, P., Meseguer-Ruiz, O., Tejedor, E., 2022. Heat and cold waves in mainland Spain: Origins, characteristics, and trends. Weather Clim. Extrem. 37, 100471 https://doi.org/10.1016/j.wace.2022.100471.

Sousa, P.M., Barriopedro, D., Ramos, A.M., García-Herrera, R., Espírito-Santo, F., Trigo, R.M., 2019. Saharan air intrusions as a relevant mechanism for Iberian heatwaves: The record breaking events of August 2018 and June 2019. Weather Clim. Extrem. 26, 100224 https://doi.org/10.1016/j.wace.2019.100224.

Tan, J., Zheng, Y., Song, G., Kalkstein, L.S., Kalkstein, A.J., Tang, X., 2007. Heat wave impacts on mortality in Shanghai, 1998 and 2003. Int. J. Biometeorol. 51, 193–200. https://doi.org/10.1007/s00484-006-0058-3.

Tošić, I., Tošić, M., Lazić, I., Aleksandrov, N., Putniković, S., Djurdjević, V., 2023. Spatio-temporal changes in the mean and extreme temperature indices for Serbia. Int. J. Climatol. 43 (5), 2391–2410. https://doi.org/10.1002/joc.7981.

Tuel, A., Eltahir, E.A., 2020. Why is the Mediterranean a climate change hot spot? J. Clim. 33 (14), 5829–5843. https://doi.org/10.1175/JCLI-D-19-0910.1.

van der Wiel, K., Bintanja, R., 2021. Contribution of climatic changes in mean and variability to monthly temperature and precipitation extremes. Commun. Earth Environ. 2 (1), 1–11. https://doi.org/10.1038/s43247-020-00077-4.

Welch, W.J., 1990. Construction of permutation tests. J. Am. Stat. Assoc. 85 (411), 693–698. https://doi.org/10.2307/2290004.

Wergen, G., Hense, A., Krug, J., 2014. Record Occurrence and Record Values in Daily and Monthly Temperatures. Clim. Dyn. 42 (5), 1275–1289. https://doi.org/10.1007/s00382-013-1693-0.

Xu, K., Wu, C., 2019. Projected changes of temperature extremes over nine major basins in China based on the CMIP5 multimodel ensembles. Stoch. Env. Res. Risk Assess. 33 (1), 321–339. https://doi.org/10.1007/s00477-018-1569-2.

Xu, P., Wang, L., Huang, P., Chen, W., 2021. Disentangling dynamical and thermodynamical contributions to the record-breaking heatwave over Central Europe in June 2019. Atmos. Res. 252, 105446 https://doi.org/10.1016/j.atmosres.2020.105446.

Yosef, Y., Aguilar, E., Alpert, P., 2021. Is it possible to fit extreme climate change indices together seamlessly in the era of accelerated warming? Int. J. Climatol. 41, E952–E963.

Yu, Y., You, Q., Zuo, Z., Zhang, Y., Cai, Z., Li, W., Jiang, Z., Ullah, S., Tang, X., Zhang, R., et al., 2023. Compound climate extremes in China: Trends, causes, and projections. Atmos. Res. 286, 106675 https://doi.org/10.1016/j.atmosres.2023.106675.

Zhang, J., Ding, T., Gao, H., 2021. Record-breaking high temperature in Southern China in 2017 and influence from the middle-latitude trough over the East of Japan. Atmos. Res. 258, 105615 https://doi.org/10.1016/j.atmosres.2021.105615.

Zorita, E., Stocker, T., von Storch, H., 2008. How unusual is the recent series of warm years? Geophys. Res. Lett. 35 (24), L24706. https://doi.org/10.1029/2008GL036228.

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R.M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M.D., et al., 2020. A typology of compound weather and climate events. Nat. Rev. Earth Environ. 1 (7), 333–347. https://doi.org/10.1038/s43017-020-0060-z.

Supplementary material to "Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021"

Jorge Castillo-Mateo, Ana C. Cebrián, and Jesús Asín

S1 Exploratory data analysis

S1.1 Distribution of daily maximum temperature
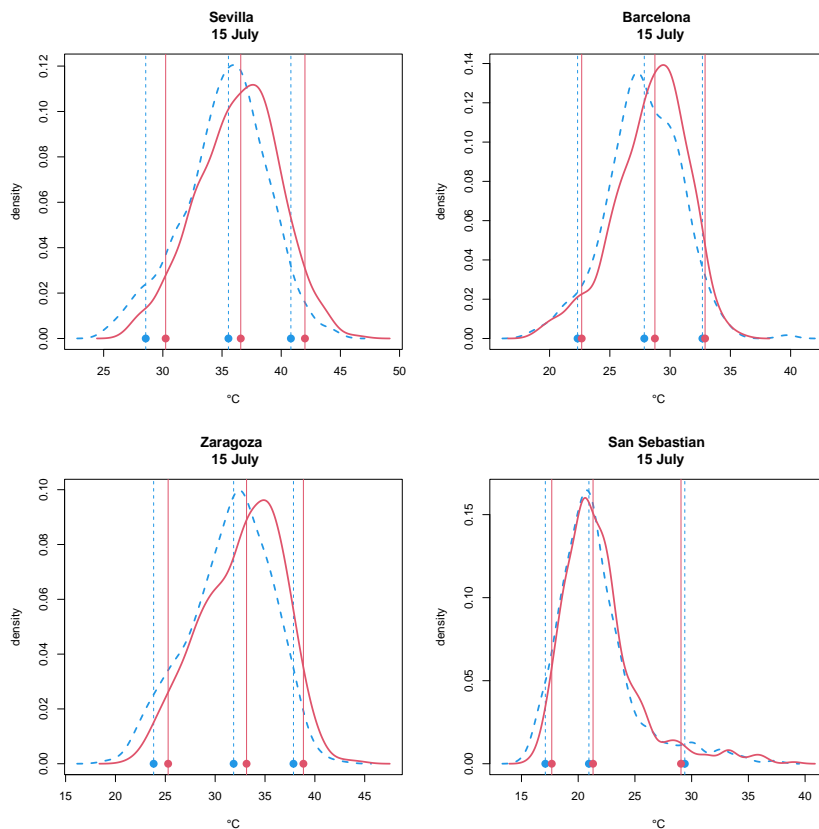
Figure S1: Kernel density estimation of T_x in Sevilla, Barcelona-Fabra, Zaragoza and San Sebastián on July 15 in the periods 1961–1990 (blue dashed) and 1991–2020 (red solid). Vertical lines show the 5th, 50th and 95th empirical percentiles.
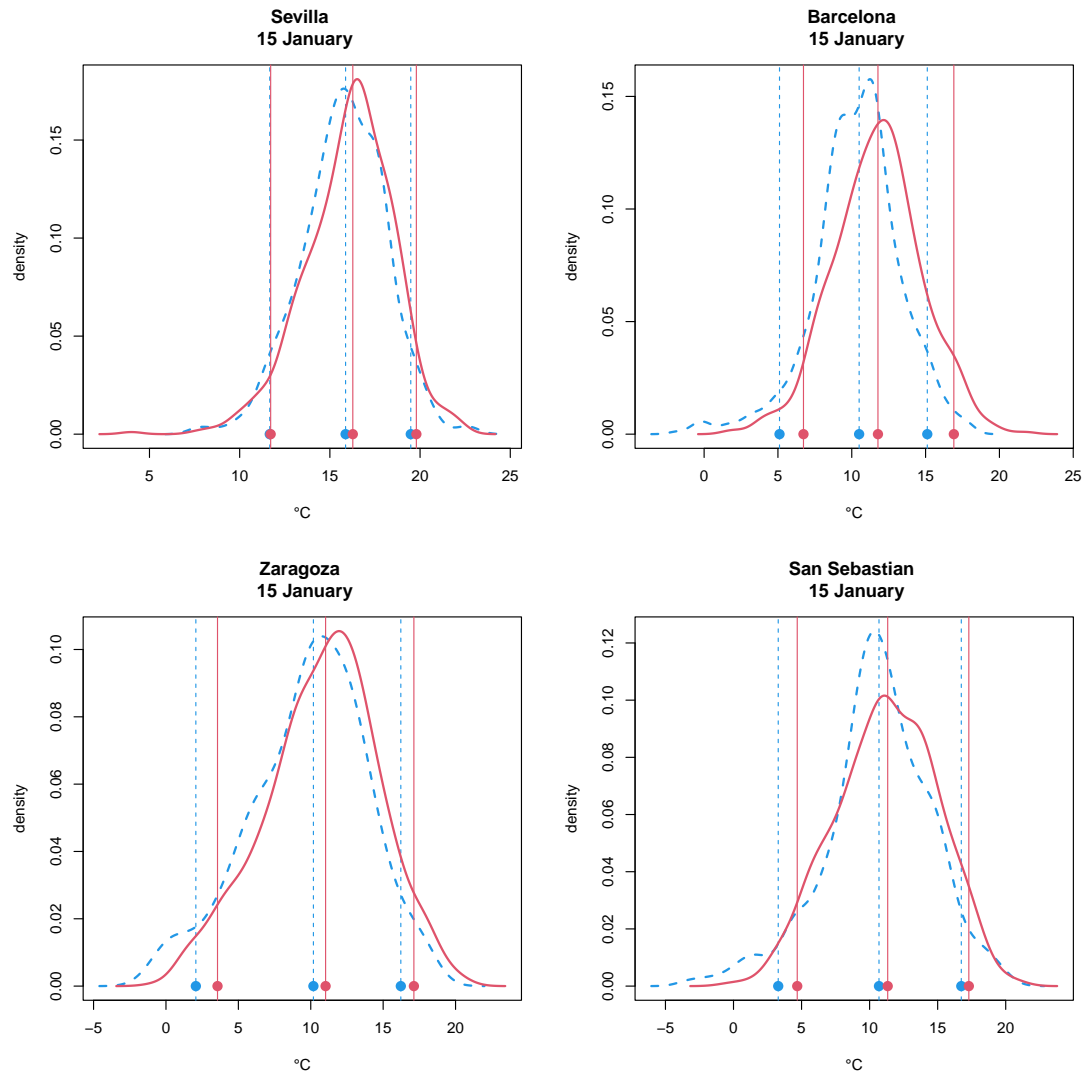
Figure S2: Kernel density estimation of T_x in Sevilla, Barcelona-Fabra, Zaragoza and San Sebastián on January 15 in the periods 1961–1990 (blue dashed) and 1991–2020 (red solid). Vertical lines show the 5th, 50th and 95th empirical percentiles.
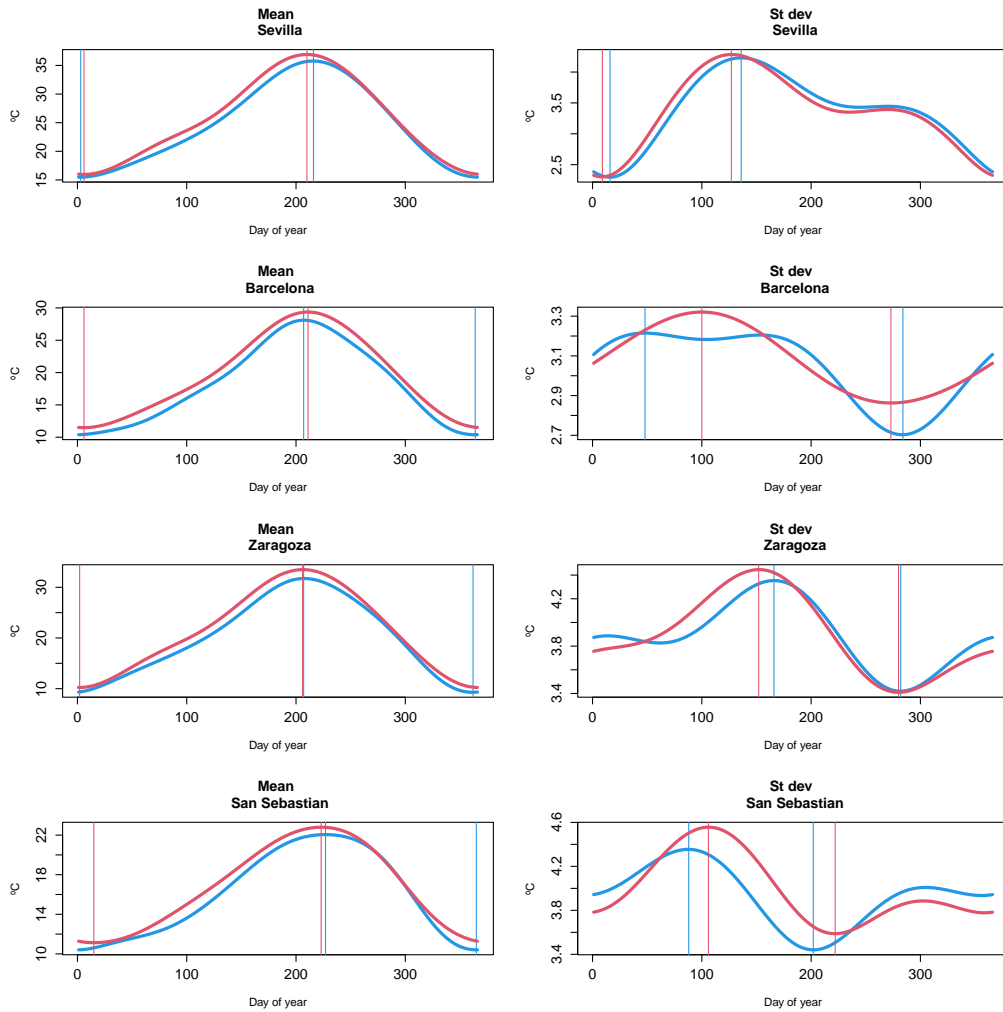
Figure S3: Mean value and standard deviation of T_x estimated by regression, using a linear model with harmonic terms for seasonality, in reference periods 1961–1990 (blue) and 1991–2020 (red), in Sevilla, Barcelona-Fabra, Zaragoza and San Sebastián.

Figure S3 shows the mean and standard deviation of T_x for every day within year using a moving window of 31 days. The period 1961–1990 in blue and 1991–2020 in red. The plots on the right express the change in the standard deviation between the two periods. The increase in the mean value varies spatially and depending on the time of the year. Particularly interesting is the change observed in the standard deviation in San Sebastián, where the variability has increased in spring, but decreased in winter.
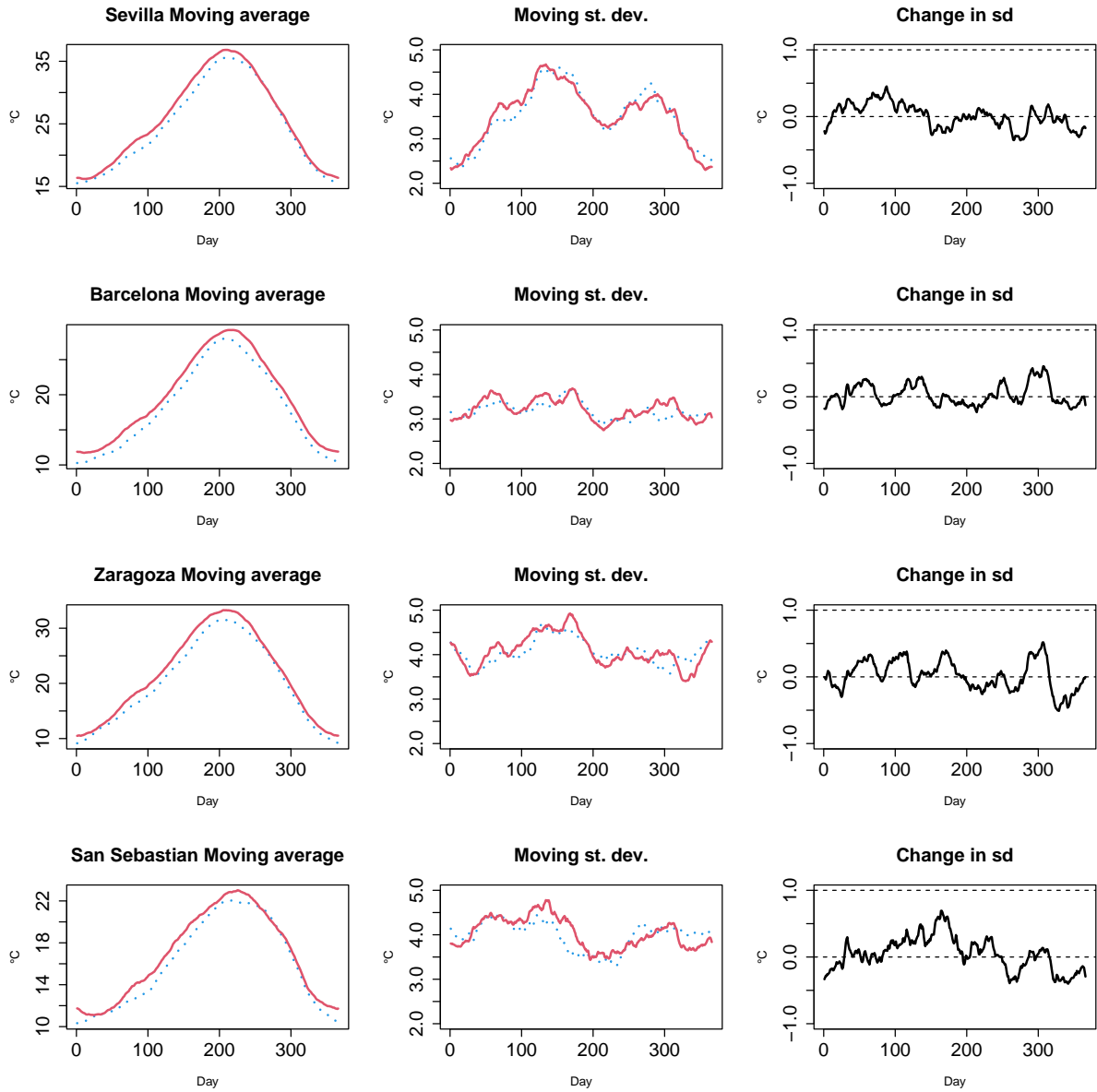
Figure S4: Mean value and standard deviation of T_x, obtained using a moving windows of 31 days, in reference periods 1961–1990 (blue) and 1991–2020 (red), and change between two periods. In Sevilla, Barcelona-Fabra, Zaragoza and San Sebastián.

4

S1.2 Trend in the mean and the quantiles

Figure S5: Median (solid) and 95th percentile (dashed) of T_x using a moving window of 31 days, in reference periods 1961–1990 (blue) and 1991–2020 (red), and change of standard deviation between two periods. In Sevilla, Barcelona-Fabra, Zaragoza and San Sebastián.

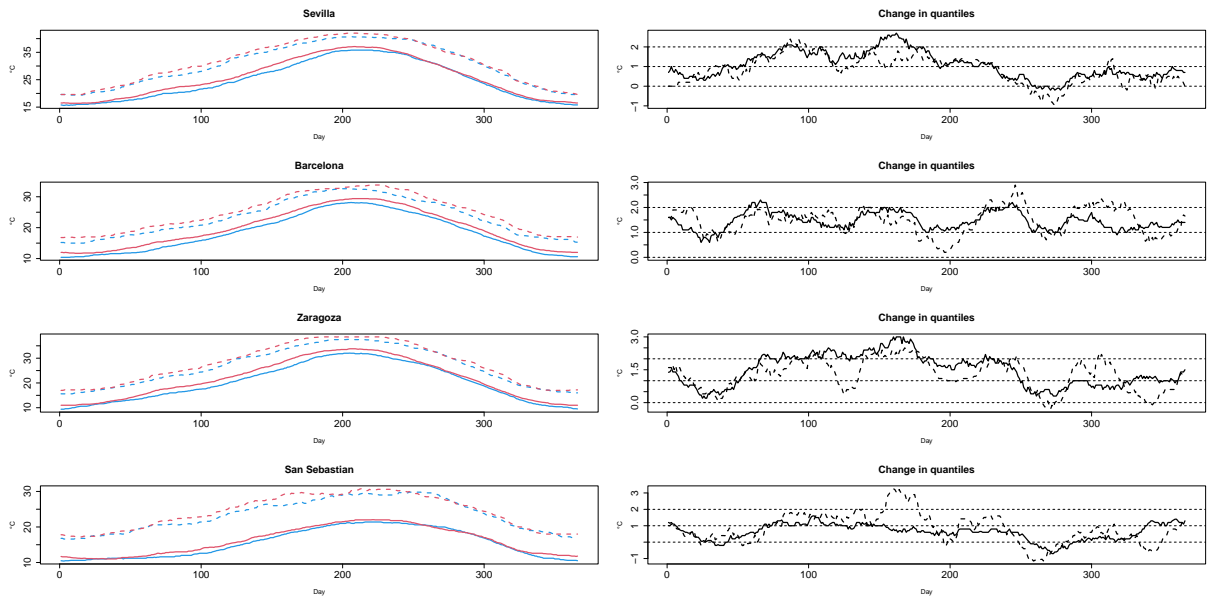|	January						July					
	Summary			Linear trend			Summary			Linear trend		
Station	$\bar{x}$	sd	q95	$\bar{x}$	sd	q95	$\bar{x}$	sd	q95	$\bar{x}$	sd	q95
TX.229	14.2	2.6	18	0.14	0.04	0.22	22.2	4.3	30	0.37	0.07	0.51
TX.230	10	2.9	14.6	0.17	0.08	0.22	18.9	4.5	26.4	0.35	0.06	0.49
TX.231	17.2	2.5	21.4	0.21	0.04	0.25	21.8	3	27	0.19	-0.06	0.11
TX.232	2.8	4.7	10.5	0.28	0.17	0.54	8	5.5	17.2	0.65	0.11	0.89
TX.233	9.1	3.5	14.5	0.33	0.05	0.53	17.3	4.6	25.3	0.36	0.05	0.49
TX.234	11.2	3.8	17.3	0.11	-0.08	0.05	15.1	4.3	23.2	0.42	0.1	0.57
TX.236	15	3.6	21	0.16	-0.01	0.16	22.2	3.7	27.9	0.36	0.07	0.42
TX.237	16.8	3.3	22.5	0.32	0.05	0.46	21.2	3.1	26.7	0.25	-0.03	0.21
TX.238	11	3.9	17.3	0.29	-0.01	0.39	20.4	4.4	27.5	0.49	0.05	0.57
TX.335	11.8	3	16.9	0.43	0.06	0.57	17.8	3.2	22.9	0.38	0.04	0.4
TX.336	11	3.4	16.2	0.27	0.08	0.36	19	4.5	26.5	0.32	0.03	0.38
TX.414	7.3	3.4	12.8	0.29	-0.02	0.28	15	4.7	23.2	0.54	-0.02	0.5
TX.416	11.1	3.2	16.2	0.2	0.11	0.42	20.3	4.7	28.2	0.41	0.05	0.47
TX.420	13.6	2.2	17	0.17	-0.04	0.09	16.6	2.9	22	0.44	0.09	0.6
TX.421	17.2	3.5	22.9	0.27	0.07	0.41	23.6	3.6	29.5	0.25	-0.02	0.2
TX.423	16.3	2.4	19.6	0.12	0.05	0.18	23.9	4	31	0.31	0.01	0.34
TX.424	8.2	3.9	15	0.29	0.08	0.41	15.2	5	23.8	0.35	-0.02	0.34
TX.1393	13.5	3.7	19.4	0.17	-0.03	0.21	18.1	4.4	26	0.52	0.08	0.6
TX.1394	11.3	2.5	15.3	0.1	-0.05	-0.01	16.7	4.9	26.4	0.46	0.15	0.73
TX.1396	9.5	3.4	15	0.34	0	0.36	18.5	5.3	27.4	0.49	0.1	0.76
TX.1397	7.3	3.5	13.4	0.14	0.02	0.14	15.3	4.7	23.6	0.19	0.01	0.34
TX.1398	10.4	3.8	16.6	0.34	0.04	0.43	18.6	4.6	26.2	0.55	-0.04	0.46
TX.1399	8.6	3.6	13.8	0.21	-0.03	0.23	18	4.5	26	0.49	0.08	0.62
TX.1401	14.5	3	19.7	0.28	0.04	0.42	19.7	2.9	24.6	0.3	-0.01	0.27
TX.2969	14	2.7	18.1	0.28	0.01	0.4	18.5	2.5	22.9	0.32	0.03	0.38
TX.3905	9.2	3.5	15	0.3	-0.05	0.3	16.2	4.8	24.6	0.6	-0.05	0.46
TX.3908	17	2	20	0.17	0.02	0.17	21.3	2.6	26.5	0.3	0.05	0.47
TX.3910	13.8	3.1	19.4	0.24	-0.04	0.18	16	2.5	20.4	0.2	-0.01	0.24
TX.3921	12.2	2.6	16.3	0.18	0.03	0.24	19.9	4.3	27.6	0.3	-0.03	0.28
TX.3922	13.3	3.1	18.8	0.22	-0.07	0.22	15.8	3.1	21.8	0.31	0.04	0.44
TX.3924	15.6	3.1	21.3	0.07	0.02	0.17	21	3	26	0.28	-0.02	0.24
TX.3937	16.5	2.4	20.1	0.05	0.05	0.09	22.4	3.5	29	0.1	-0.02	0.07
TX.3943	10.3	4.4	17.2	0.23	0.01	0.43	21.2	4	27.4	0.42	0.02	0.45
TX.3963	15.9	2.4	19.5	0.19	0	0.15	22.8	4	30	0.33	-0.01	0.26
TX.3970	7.6	3.2	12.5	0.07	-0.05	0.1	16	4.4	23.7	0.29	0.01	0.33
TX.3971	10.1	3.6	15.4	0.22	0.05	0.27	17.8	5	26.1	0.5	0.02	0.54

Table S1: Mean ($\bar{x}$), standard deviation (sd), and 95th percentile (q95) of T_x series (°C) for the reference period 1991–2020 in January and July. The least squares linear trend (°C/decade) for those summaries.

S2 Records in daily maximum temperature and daily temperatures at geopotential levels

Year	Month	Day	# records	# in contiguous days
Absolute records: Simultaneously more than 2 stations				
1975	7	16	3	3
1978	7	16	3	6
1981	7	29	4	6
1982	7	6	3	8
1982	7	7	5	9
1987	8	12	3	7
1987	8	13	4	7
1995	7	24	4	6
2003	8	4	3	3
2010	8	27	3	3
2012	8	10	4	4
2017	7	13	3	3
2019	6	29	3	4
Records in January: Simultaneously more than 5 stations				
1975	1	29	6	9
2003	1	27	6	8

Table S2: Dates when there are simultaneous records in several stations, number of stations and number of stations with absolute record from previous to following days. Absolute records in the series and absolute records in January subseries are summarized.

Year	Month	Day	# affected grid-points
Compound event based on absolute records			
1980	8	2	6
1983	7	30	8
1987	8	13	8
1987	8	16	7
2019	6	26	6
2019	6	27	7
2021	7	10	6
Compound event based on absolute records of January			
1981	1	21	6
1988	1	2	11
1998	1	3	41
2016	1	31	8
Compound event based on absolute records of July			
1981	7	30	13
1982	7	6	6
1983	7	30	10
1995	7	24	16
1998	7	28	10
2021	7	10	8

Table S3: Dates with compound events in at least 6 points of the grid. Compound events are based on absolute records in the series and absolute records in January subseries and July subseries. Only dates from 1971 are shown.
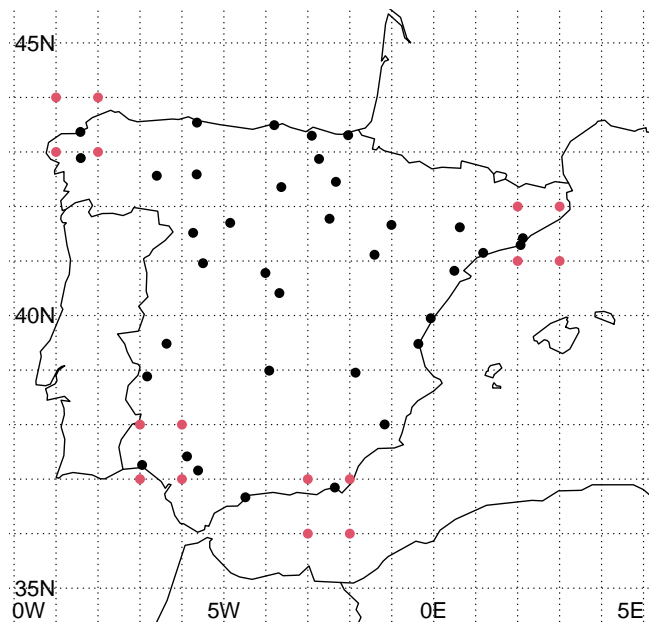
Figure S6: Grid of ERA5 reanalysis over Iberian peninsula.
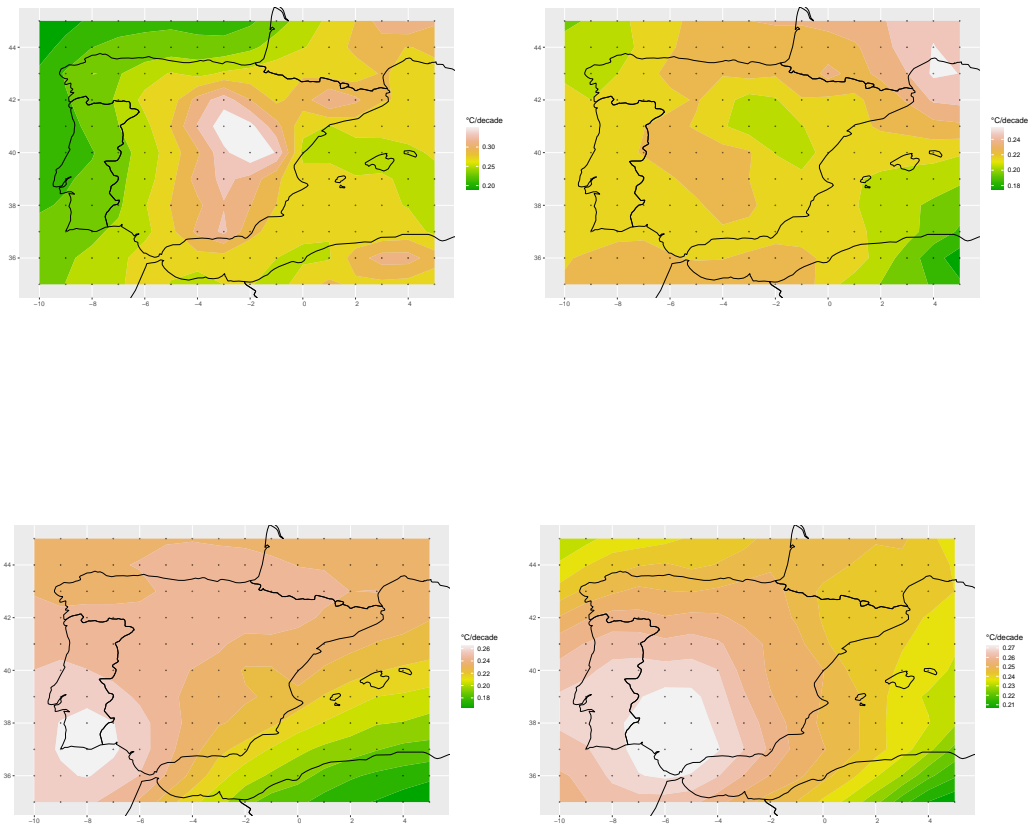
Figure S7: Map for the estimated linear trend (°C/decade) of the daily temperature series in every point of the $1° \times 1°$ ERA5 grid, at geopotential levels, upper 850 (left) and 700 hPa, bottom 500 (left) and 300 hPa.

S2.1 Records in the monthly maximum series

Figures S8 and S9 are explained in Section 3.3.2 from the Main text. Next, we consider the records in the series defined as the monthly maximum of T_x series in each station. We compare the evolution of extreme events in T_x, globally for the Iberian Peninsula, and in daily temperature in geopotential levels. In particular subseries of January and July are separately studied. In order to summarize the grid at each geopotential level, ERA5 statistics are obtained joining records of 4 corner points of ERA grid points, defined by 35N–45N and 10W–5E. Figure S10 represents evolution of $\bar{\mathcal{N}}_t$, the weighted cumulative number of records, joining the records of monthly maximum in each of Spanish T_x series (red). Results for January are shown in the left panel and for July in the right. The change-point is represented with a vertical red line. Equivalent statistics for records of monthly maximum of daily temperature series at 850 (orange), 700 (green), 500 (blue) and 300 (black) hPa. Overlapped vertical lines indicate the estimated change–point with the corresponding color. In July, results show an increasing for high levels more intense that the Spanish series and earlier in 850 hPa. In January, the change–point of Spanish station is estimated in 1998, coinciding with 700 hPa and near of 850 hPa result.
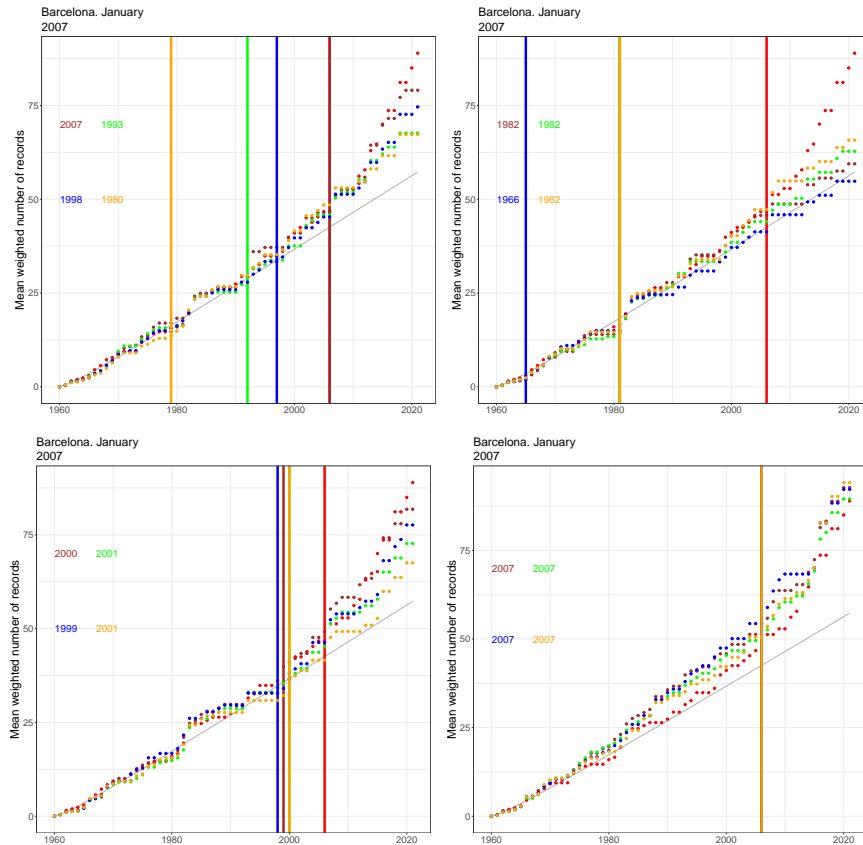
Figure S8: Mean weighted number of records up to time t versus t, in T_x Barcelona, January 1960–2021, and similar statistics in daily temperatures at 850, 700, 500 and 300 hPa, in 4 points of grid around Barcelona.

Figure S9: Mean weighted number of records up to time t versus t, in T_x Sevilla, July 1960–2021, and similar statistics in daily temperatures at 850, 700, 500 and 300 hPa, in 4 points of grid around Sevilla.

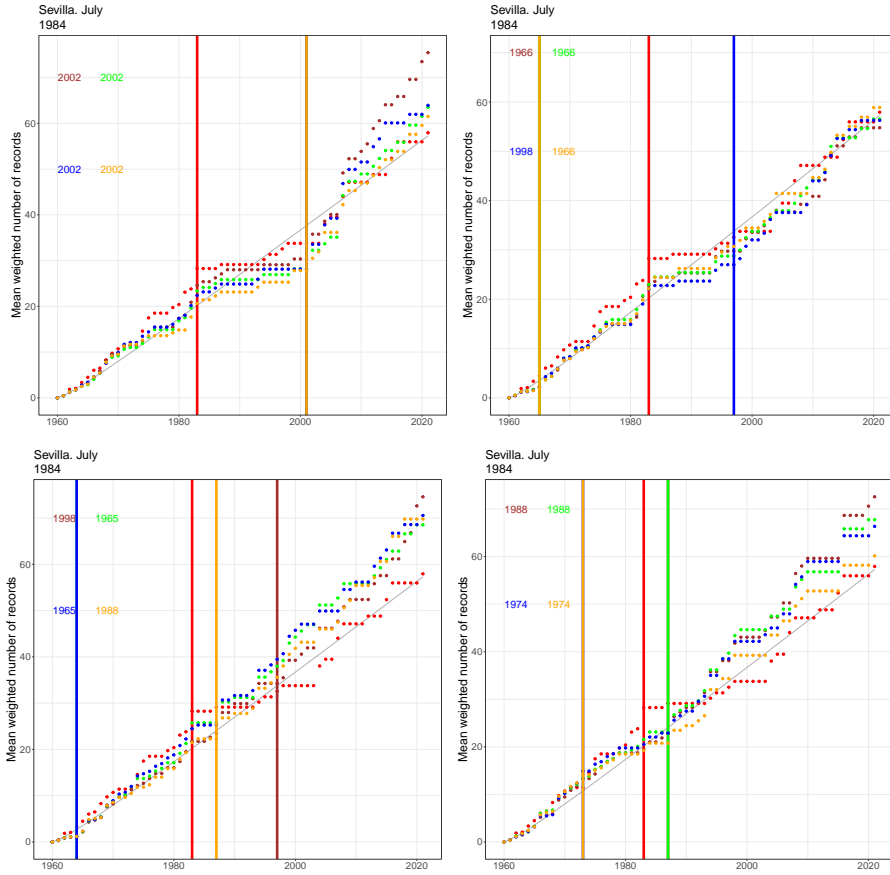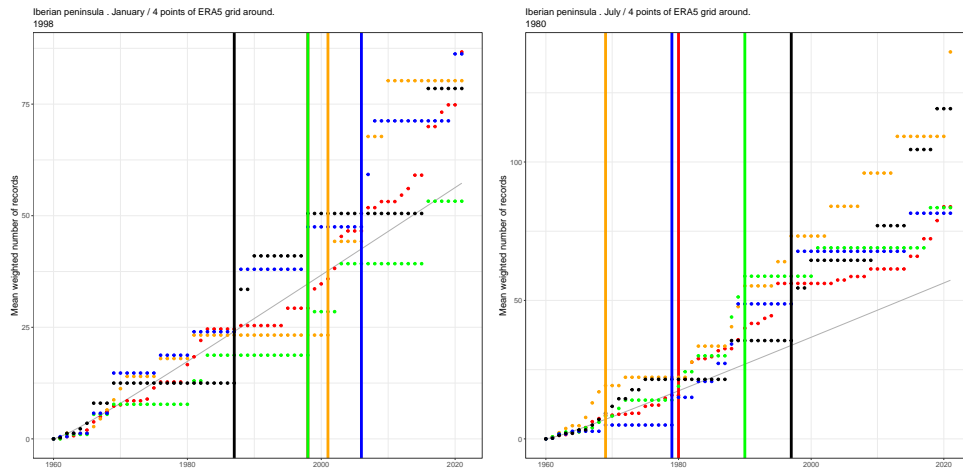Figure S10: Mean weighted number of records up to time t versus t, in Spanish T_x, and similar statistics of daily series of temperature at 850 (orange points), 700 (green), 500 (blue) and 300 hPa (black), in 4 corner points of grid. Left, January. Right, July.

Chapter 4

Discussion

This chapter will try to give a global discussion of the research papers that constituted the main body of this Thesis.

Firstly, the Thesis was focused on the development of a space-time model for the Aragón dataset in the basic geostatistical setting with Gaussian errors. The published paper Castillo-Mateo et al. (2022) aimed to learn about basic features of temperatures; it developed a Bayesian hierarchical mixed effects autoregressive model that characterized the mean of the response distribution; it incorporated some new features such as two temporal scales and four GPs to model spatially varying coefficients. A MCMC algorithm was designed to fit the model and Bayesian kriging was used for interpolation at unobserved locations, this was a recurrent approach in all subsequent objectives. Next, this model resulted especially useful for carrying out model-based inference. The unpublished work Cebrián et al. (2023) proposes model-based tools including predictive spatial probability surfaces and spatial extents for an event; it uses the output from the mean model to make inference around the center of the distribution. Particular cases of these tools were used to give a formal probabilistic definition for extents of EHEs. The published paper Cebrián, Asín, et al. (2022) used the output from the Schliep et al. (2021) model and the model-based tools to define concepts and learn about characteristics of EHEs.

Secondly, this Thesis was focused on more general space-time models, for all quantile levels of the response, particularly high quantiles. The published research paper Castillo-Mateo, Asín, Cebrián, Gelfand, et al. (2023) aimed to learn about the features of temperatures in high quantiles and compare them with the mean; it extended the mean model to the multiple QAR context using the AL distribution, and offered an attractive approach to obtain marginal quantiles from the conditional quantiles in the autoregression with the AL specification. From a theoretical point of view, joint QRs offers more opportunities for inference but a bigger challenge. The published research paper Castillo-Mateo et al. (*in press*) aimed to propose methodology

to advance in the development of joint QAR models; it proposed a novel specification through Kumaraswamy distributions and extended it to the spatial setting.

Thirdly and lastly, this Thesis was focused on the study of the occurrence of the most extreme observed values for the response—record-breaking events. The results from mean and QR gave strong evidence of the distinct behavior between the central part and the tails of the distribution, which led to the study of records in particular. The first aim was to make available an accessible framework for the statistical analysis of records. The published research papers Cebrián, Castillo-Mateo, et al. (2022) and Castillo-Mateo (2022) considered the framework of records across series for each calendar day. They developed statistical hypothesis tests to explore the null hypothesis of a stationary behavior against a trend or a change-point in the probabilities of record, respectively. Given the great importance of the study of records in several fields, the published research paper Castillo-Mateo, Cebrián, et al. (2023a) described the R package **RecordTest** (Castillo-Mateo, 2023b), which implements most of the statistical hypothesis tests and graphical tools based on the occurrence of records available in the literature. A limitation of the proposed tests was that they required independent series. The published research paper Castillo-Mateo, Cebrián, et al. (2023b) proposed a modification of the tests using permutation techniques; it implemented those tools to analyze the Spanish dataset. The second aim was to bring together the Bayesian hierarchical spate-time modeling framework and the analysis of records. The unpublished work Castillo-Mateo, Gelfand, et al. (2023) comes back to the Bayesian framework to propose a mixed effects logistic regression model with a very novel specification for the record indicators; it offers model-based full inference regarding any characteristic related to the occurrence of records.

Chapter 5

Conclusions and future work

This chapter will provide a concluding overview of the project by summarizing the main methodological contributions and research findings in connection with the initial objectives. It will also address the significance of the contributions and their limitations and suggest potential avenues for future research.

5.1 Summary and implication of the contributions

This project aimed to develop spatiotemporal statistical methodology for analyzing extreme temperature events, focusing on EHEs, high quantiles, and record-breaking events in Aragón and Spain. The objectives included assessing changes in EHEs duration and frequency, characterizing distribution behavior, comparing long-term trends between the mean and high quantiles, studying record occurrences against stationary behavior, and quantifying trends across time and space.

Contributions to Bayesian geostatistics. Acknowledging that climate change with regard to temperature is occurring both temporally and spatially, the contribution to the "basic" Bayesian geostatistical framework has included a very rich space-time mean model for daily maximum temperatures. The specification is continuous in space and autoregressive in time. With two time scales, autoregression was examined annually and also daily for the summer season within each year. Novel spatial structures have been found including spatially varying intercepts and trend coefficients as well as spatially varying autoregression coefficients and variances.

Posterior predictive samples of daily temperature time series on a fairly fine grid scale generated from particular choices of Bayesian space-time models have been used to asses space-time changes. Exceedance events for local reference values of the temperature distribution have been analyzed. Two basic ideas have been offered: probability surfaces which capture the spatial variation in the chance of an exceedance

event and provide climate risk maps; and spatial extents which, for a subregion of interest, capture the expected proportion of incidence of a given exceedance event over the region. For a specified region, for a given day, the definition of spatial extent takes the form of a block average over the region. It is an average of indicator variables which identify the exceedance of a local reference value by the daily maximum temperature for the day at each location within the region. It has been demonstrated that extents can be calculated through Monte Carlo integration and realizations can be obtained from the space-time models for daily maximum temperatures.

Exceedance events have been defined in terms of a local mean value or increments between two decades. Also, because notions of the spatial extent of EHEs have only been considered informally and descriptively in the climate community, a formal probabilistic definition for extents of EHEs has been introduced. These quantities are defined at daily scale and can be averaged to any other temporal scale of interest. Comparison of temperature evolution has been presented at daily and seasonal scale both temporally between decades and spatially between subregions.

Contributions to quantile regression. The contributions to QR include a modeling approach to predict a specific quantile in a spatiotemporal framework. A spatial conditional autoregressive model has been specified on a daily scale using the AL distribution for the errors. The considered specification enables spatial autoregression at a daily scale that captures serial correlation and facilitates assessment of persistence. The flexibility of the model is increased by considering two scales of time as well as seasonal behavior, time trend, and four GPs that represent the spatial dependence of the intercept, the trend, the serial dependence, and the scale of the AL errors. Bayesian model fitting enables full posterior inference for a given quantile. Although the model gives conditional quantiles, an attractive approach to obtain marginal quantiles at daily scale has been offered. These marginal quantiles enable interpolation. The approach can also provide marginal quantiles associated with averages of the response variable, both spatially and dynamically. Posterior inference to evaluate changes between marginal quantiles of spatial and time averages can also be implemented.

Subsequently, modeling for joint (non-quantile crossing) QAR has been presented and consequentially expanded. In particular, the QAR(1) setting has been characterized in a way that allows for a more flexible autoregression structure than the one in the seminal paper by Koenker and Xiao (2006). This has been extended to the QAR(p) case. A novel multivariate time series version has been offered using a Gaussian copula. A spatial version has also been elaborated, using a GP copula based upon a GP in conjunction with four additional GPs. This model enables spatially

varying quantile functions. The modeling is entirely parametric through the use of the Kumaraswamy distributions.

Contributions to record-breaking analysis. This project has proposed several statistical tests based on the likelihood of the record indicators and complementary graphical tools to detect the existence of non-stationary behavior in the occurrence of records. It has also proposed three novel change-point detection tests whose statistic is of the CUSUM type based on the record indicators. The record tests use the properties of the occurrence of records in c.i.i.d. series, which makes them distribution-free, this is, the null hypothesis over the probabilities of record is $H_0 : p_t = 1/t$. Statistics to deal with seasonal series have also been considered. P-values are obtained through the exact distribution, the asymptotic distribution, Monte Carlo simulations, and permutations. Despite having a very small sample information compared to the total of the series, the Monte Carlo simulations have shown that the proposed record tests are powerful for detecting deviations from the null hypothesis and the change-point tests return a biased change-point estimate. The change-point estimate should be interpreted as the time since which an underlying trend has been impacting the observed record occurrences.

A substantial contribution is the R package **RecordTest**. As far as I know, this is the only available software package for the analysis of record-breaking events. It provides a useful general framework for a fully non-parametric analysis of non-stationary behavior in the extremes and the records in particular. The package offers functions that cover all the steps in this type of analysis. This includes functions to prepare the data, identify the variables for characterizing the record occurrence, and implement graphical tools for exploratory analysis. The main functionality of the package is the implementation of all the tests to detect non-stationarity based on records currently available in the literature, and complementary graphical tools.

Finally, this project has taken up the first fully developed modeling attempt to analyze the incidence of record-breaking temperatures. The modeling needed to effectively explain the incidence of record-breaking over this region during this period required careful specification of the probabilities of the indicator functions which define record-breaking sequences. Key features have included: explicit trend behavior, necessary autoregression, significance of distance to coast, useful interactions, spatial random effects, and very strong daily random effects.

Contributions to the analysis of climate change in Aragón and Spain. The above contributions were motivated through time series spanning over 60 years of daily temperatures from 18 and 40 sites in Aragón and Spain, respectively. Some interesting

results include a global increment of summer temperatures in Aragón in the period 1956–2015 of 0.21°C/decade in the mean and approximately 0.10°C/decade in the high quantiles. However, the spatial patterns for these trends in space vary a lot across quantile levels. A clear increase of the extent of EHEs across time is observed. For example, the posterior probability of the yearly average extent across summer months in the decade 2006–2015 being higher than in 1966–1975 is higher than 0.8. Also the first day of the summer where the extent of EHEs is higher than 10% has decreased around seven days. In a different direction, the joint QAR model has allowed to capture autoregression structure in daily temperature data, which is not strictly increasing in τ, but decreasing in both tails.

The record tests have found significant evidences of the effect of an increasing trend in the occurrence of record-breaking events in daily maximum temperature in most of Spain with the exception of the Cantabrian Coast. The stronger evidence in time is observed during summer. The model suggests that in 90% of the region there is a significantly higher number of records compared to the stationary case, and the total number of records in decade 2012–2021 has almost doubled compared to a stationary situation.

The proposed modeling and complementary tools can be adapted to other regions. The models are highly adaptable, allowing for different response variables in spatiotemporal contexts, consideration of other geographical covariates, and excluding or including specific spatial processes as needed. The flexible autoregression terms can express behavior in series where serial correlation is an important source of variation. Overall, the project provides valuable tools for understanding and addressing the effects of climate change on the central part of the distribution as well as EHEs, quantiles, and record-breaking events with a wide range of applications in climate studies.

5.2 Limitations, improvement opportunities and future work

Small versus big n data. A limitation of the present analysis is the relatively sparse spatial coverage with only 18 monitoring stations in Aragón and 40 in Spain. This limited spatial sampling can lead to challenges in accurately characterizing spatial surfaces in the modeling, potentially resulting in less accurate spatial predictions. Future work in this domain could explore different regions that offer a more extensive network of spatial monitoring locations. With a big number of spatial locations, reduced rank approximations to GPs may be used to address the computational bottleneck, e.g., the Gaussian predictive process (Banerjee et al., 2008) or the

nearest-neighbour GP (Datta et al., 2016).

Fusion techniques for downscaling and climate projections. Working within the same region, there is a growing body of literature on data *fusion* techniques, particularly within the context of data assimilation with computer model output. The proposed models for observed station data could be extended to incorporate information from environmental computer models which provide output—rather than real data—for grid cells at some spatial scale. Data fusion would aim to integrate these two data sources to provide additional insights into the spatiotemporal responses over the same region and time period to enhance spatial interpolation accuracy. However, aligning data layers that may differ in spatial and temporal resolutions can be challenging. In this regard, addressing this misalignment issue when *downscaling* numerical model output to point-level observations becomes fundamental. For example, a straightforward extension of the mean or quantile autoregression models could incorporate the downscaler model by Berrocal et al. (2010) that regresses the observed data on the numerical model output using spatially varying coefficients. Also, there can be measurement error concerns when downscaling. To address this, the two models proposed by Berrocal et al. (2012) could be considered: (i) a Gaussian Markov random field smoothed downscaler, and (ii) a smoothed downscaler with spatially varying random weights. Moreover, future research directions could involve projecting temperature changes under different climate scenarios downscaling atmospheric variables, encompassing both mean and multiple quantile models, as well as forecasting spatial extents under various climate projections.

Software, model fitting, and MCMC. The spatiotemporal models proposed provide extensive opportunities for full inference, although their proper implementation might require expertise in multi-level stochastic modeling. The development of the MCMC algorithms necessary to fit these models can be complicated and specific. There are currently numerous tools available for fitting Bayesian models, specific R packages like **spBayes** (Finley et al., 2015; Finley & Banerjee, 2020) or **spTimer** (Bakar & Sahu, 2015), and probabilistic programming languages like JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017). The proposed models cannot be fitted using these tools, in the former case because the models they provide are highly specific, and in the latter case due to the high computational cost. The R package **spTReg** is under development and aims an easy and general implementation of the mean and quantile models. The development of user-friendly statistical tools and software packages for these models will greatly benefit the broader scientific community. Simplifying the implementation

of these models would enable researchers with varying levels of statistical expertise to contribute to this important field.

Joint QAR. It is important to note that covariates have not been introduced into the joint QAR modeling in order to achieve simple conditions for non-crossing of quantiles. In order to consider coherent implementation of covariates, conditions have to be imposed on the support for the covariates. Current work is in progress to bridge our modeling with the work of Yang and Tokdar (2017). The univariate QAR with covariates case could be addressed considering a suitable bounded prior guess in their so-called constraint-free parameterization. The spatial version could build upon the Chen and Tokdar (2021) modeling work with dependent quantile levels, but including spatially varying quantiles is in progress.

An obvious critique of the *linear* QAR model comes from the observation that, when the QAR slope coefficient depends upon τ, there must be a subregion of the support of Y_t for which the ordering of quantiles is reversed. Thus, the linear in lagged Y_t's formulation must be regarded as a useful local approximation over a bounded support. One remedy for this assumption will be to resort to *nonlinear* expressions of the QAR model. For example, Chen et al. (2009) explore one approach to models of this type based on copula specifications, and Xu and Reich (2023) offer a nonlinear QR that expands the pdf in I-spline basis functions where the covariate-dependent coefficients are modeled using neural networks. Greater flexibility comes at a price, the nonlinear specification loses the nice linear interpretation of the lags or covariate effects.

Record-breaking analysis. Numerous opportunities have emerged in the analysis of record-breaking events. For example, a typical approach in change-point analysis is to execute the statistical investigation based on a fixed sample size. Alternatively, Chu et al. (1996) developed a sequential test procedure for linear regression models. Current joint work with Lajos Horváth from University of Utah consists of extending their monitoring scheme in the context of record occurrences.

A different direction consists on the exploration of δ-records (Gouet et al., 2020), pooling information with values that are close from the record should improve prediction accuracy and give a better picture of climate change. The main idea consists of introducing a multinomial probit or logit model for three categories: records, values that are $-\delta$ units below the record, and all other values. Another interesting question arising in the analysis of records is the joint evolution of upper and lower records. Similar models that take into account upper and lower records in daily maximum and minimum temperatures, which will lead to a bivariate multinomial probit or logit

model, are under consideration.

In a similar direction, more research to study the relation between records and the atmospheric situation expressed by air temperature at different geopotential levels is of interest, and also with teleconnection patterns to find plausible physical mechanisms. Again, models including atmospheric variables as covariates would be also useful as statistical downscaling models, and they could be used to obtain future projections of the behavior of records under different climate scenarios.

5.3   Concluding remark

In conclusion, this Thesis delved into the intricate relationship between climate change and extreme temperature events. The project developed various statistical models and methodologies to decipher the changing landscape of these events over time and space. The findings underscore the impact of climate change on temperature trends, highlighting the need for comprehensive strategies to address its repercussions on public health, agriculture, and the economy. Furthermore, this project has contributed valuable insights into the analysis of extreme events and the development of innovative statistical tools for their assessment. The hierarchical modeling approach fitted within a Bayesian framework offering a fully model-based perspective has proven key in achieving these research objectives. Also, simpler statistical tools accessible to climate scientists and practitioners may become popular as climate change continues to have a significant impact on humanity. As I conclude this chapter, I emphasize the urgency of continued research in this critical field to better understand, adapt to, and mitigate the effects of climate change on our world.

Conclusiones y trabajo futuro

Este capítulo proporcionará un resumen final del proyecto resumiendo las principales contribuciones metodológicas y hallazgos de la investigación en relación a los objetivos iniciales. También abordará la importancia de las contribuciones y sus limitaciones y sugerirá posibles vías para futuras investigaciones.

5.1 Resumen e implicación de las contribuciones

Este proyecto tuvo como objetivo desarrollar una metodología estadística espaciotemporal para analizar eventos de temperatura extrema, centrándose en eventos de calor extremo (EHEs, por su acrónimo en inglés), cuantiles altos y eventos récord en Aragón y España. Los objetivos incluían evaluar cambios en la duración y frecuencia de los EHEs, caracterizar el comportamiento de la distribución, comparar tendencias a largo plazo entre la media y los cuantiles altos, estudiar las ocurrencias de récord frente al comportamiento estacionario y cuantificar tendencias en el tiempo y el espacio.

Contribuciones a la geoestadística bayesiana. Reconociendo que el cambio climático con respecto a la temperatura está ocurriendo tanto temporal como espacialmente, la contribución al marco geoestadístico bayesiano "básico" ha incluido un modelo muy rico de medias espacio-tiempo para las temperaturas máximas diarias. La especificación es continua en el espacio y autorregresiva en el tiempo. Con dos escalas de tiempo, la autorregresión se examinó anualmente y también diariamente para la temporada de verano de cada año. Se han encontrado nuevas estructuras espaciales que incluyen interceptos y coeficientes de tendencia que varían espacialmente, así como coeficientes autorregresivos y varianzas que varían espacialmente.

Para evaluar los cambios espaciotemporales se han utilizado muestras predictivas a posteriori de series temporales de temperatura diaria en una resolución de malla bastante fina generadas a partir de elecciones particulares de modelos bayesianos espacio-tiempo. Se han analizado eventos de exceso para valores de referencia locales de la distribución de temperatura. Se han ofrecido dos ideas básicas: superficies de

probabilidad que capturan la variación espacial en la probabilidad de un evento de exceso y proporcionan mapas de riesgo climático; y extensiones espaciales que, para una subregión de interés, capturan la proporción esperada de incidencia de un evento de exceso dado en la región. Para una región específica, para un día dado, la definición de extensión espacial toma la forma de un promedio por bloques sobre la región. Se trata de un promedio de variables indicadoras que identifican la superación de un valor de referencia local por parte de la temperatura máxima diaria para el día en cada ubicación dentro de la región. Se ha demostrado que las extensiones se pueden calcular mediante la integración de Monte Carlo y se pueden obtener realizaciones a partir de modelos espacio-tiempo para las temperaturas máximas diarias.

Los eventos de exceso se han definido en términos de un valor promedio local o incrementos entre dos décadas. Además, dado que las nociones sobre la extensión espacial de EHEs solo se han considerado de manera informal y descriptiva en la comunidad climática, se ha introducido una definición probabilística formal para las extensiones de EHEs. Estas cantidades se definen a escala diaria y se pueden promediar a cualquier otra escala temporal de interés. La comparación de la evolución de la temperatura se ha presentado a escala diaria y estacional, tanto temporalmente entre décadas como espacialmente entre subregiones.

Contribuciones a la regresión cuantílica. Las contribuciones a la regresión cuantílica (QR, por su acrónimo en inglés) incluyen un enfoque de modelización para predecir un cuantil específico en un marco espaciotemporal. Se ha especificado un modelo autorregresivo condicional espacial a escala diaria utilizando la distribución asimétrica de Laplace (AL) para los errores. La especificación considerada permite la autorregresión espacial a escala diaria que captura la correlación serial y facilita la evaluación de la persistencia. La flexibilidad del modelo se incrementa al considerar dos escalas de tiempo, así como comportamiento estacional, tendencia temporal y cuatro procesos gaussianos (GPs, por su acrónimo en inglés) que representan la dependencia espacial del intercepto, la tendencia, la dependencia serial y la escala de los errores AL. El ajuste del modelo bayesiano permite una inferencia completa a posteriori para un cuantil dado. Aunque el modelo proporciona cuantiles condicionales, se ha ofrecido una aproximación atractiva para obtener cuantiles marginales a escala diaria. Estos cuantiles marginales permiten la interpolación. El enfoque también puede proporcionar cuantiles marginales asociados con promedios de la variable respuesta, tanto espacial como dinámicamente. La inferencia a posteriori para evaluar los cambios entre cuantiles marginales de promedios espaciales y temporales también se puede implementar.

Posteriormente, se ha presentado y consecuentemente ampliado la modelización de

autorregresión de cuantiles (QAR, por su acrónimo en inglés) conjunta (sin cruce de cuantiles). En particular, se ha caracterizado el caso QAR(1) de manera que permita una estructura de autorregresión más flexible que la que se presenta en el trabajo fundacional de Koenker y Xiao (2006). Esto se ha extendido al caso QAR(p). Se ha ofrecido una novedosa versión para series temporales multivariantes utilizando una cópula gaussiana. También se ha elaborado una versión espacial, utilizando un proceso de cópula gaussiana basada en un GP junto con cuatro GPs adicionales. Este modelo permite funciones de cuantiles que varían espacialmente. La modelización es completamente paramétrica mediante el uso de las distribuciones de Kumaraswamy.

Contribuciones al análisis de récords. Este proyecto ha propuesto varios test estadísticos basados en la verosimilitud de los indicadores de récord y herramientas gráficas complementarias para detectar la existencia de comportamientos no estacionarios en la ocurrencia de récords. También ha propuesto tres novedosos test de detección de un punto de cambio cuyo estadístico del tipo CUSUM se basa en los indicadores de récord. Los test de récord utilizan las propiedades de la ocurrencia de récords en series continuas independientes e idénticamente distribuidas (c.i.i.d.), lo que los hace libres de distribución. Esto es, la hipótesis nula sobre las probabilidades de récord es $H_0 : p_t = 1/t$. También se han considerado estadísticos para tratar con series estacionales. Los p-valores se obtienen a través de la distribución exacta, la distribución asintótica, simulaciones Monte Carlo y permutaciones. A pesar de contar con una cantidad muy limitada de información muestral en comparación con el total de la serie, las simulaciones Monte Carlo han demostrado que los test de récord propuestos son potentes para detectar desviaciones de la hipótesis nula y que los test de punto de cambio devuelven una estimación del punto de cambio sesgada. La estimación del punto de cambio debe interpretarse como el tiempo desde el cual una tendencia subyacente ha estado afectando las ocurrencias de récords observados.

Una contribución importante es el paquete de R **RecordTest**. Hasta donde tengo conocimiento, este es el único paquete de software disponible para el análisis de eventos récord. Proporciona un marco general útil para un análisis completamente no paramétrico del comportamiento no estacionario en los extremos y en los récords en particular. El paquete ofrece funciones que abarcan todos los pasos en este tipo de análisis. Esto incluye funciones para preparar los datos, identificar las variables que caracterizan la ocurrencia de récords e implementar herramientas gráficas para análisis exploratorio. La funcionalidad principal del paquete es la implementación de todos los test actualmente disponibles en la literatura para detectar la no estacionariedad basada en récords, junto con herramientas gráficas complementarias.

Finalmente, este proyecto ha abordado el primer intento de modelización completamente desarrollado para analizar la incidencia de temperaturas récord. La modelización necesaria para explicar de manera efectiva la incidencia de récords en esta región durante este período requería una especificación cuidadosa de las probabilidades de las funciones indicadoras que definen secuencias de récords. Las características clave han incluido: comportamiento de tendencia explícito, autorregresión necesaria, significación de la distancia a la costa, interacciones útiles, efectos aleatorios espaciales y efectos aleatorios diarios muy fuertes.

Contribuciones al análisis de cambio climático en Aragón y España. Las contribuciones mencionadas fueron motivadas a través de series temporales que abarcan más de 60 años de temperaturas diarias de 18 y 40 sitios en Aragón y España, respectivamente. Algunos resultados interesantes incluyen un incremento global de las temperaturas de verano en Aragón en el período 1956–2015 de 0,21°C/década en la media y aproximadamente 0,10°C/década en los cuantiles altos. Sin embargo, los patrones espaciales de estas tendencias varían considerablemente a través de los cuantiles a distintos niveles. Se observa un claro aumento en la extensión de los EHEs a lo largo del tiempo. Por ejemplo, la probabilidad a posteriori de que la extensión promedio anual durante los meses de verano en la década 2006–2015 sea mayor que en 1966–1975 es superior a 0,8. Además, el primer día del verano en el que la extensión de los EHEs es superior al 10 % ha disminuido alrededor de siete días. En otra dirección, el modelo QAR conjunto ha permitido capturar la estructura de autorregresión en los datos diarios de temperatura, que no es estrictamente creciente en τ, sino que disminuye en ambas colas.

Los test de récord han encontrado evidencias significativas del efecto de una tendencia creciente en la ocurrencia de eventos que baten récords en la temperatura máxima diaria en la mayor parte de España, con la excepción de la costa cantábrica. Las evidencias más sólidas en términos temporales se observan durante verano. El modelo sugiere que en el 90 % de la región hay un número significativamente mayor de récords en comparación con la situación estacionaria, y el número total de récords en la década 2012–2021 casi se ha duplicado en comparación con una situación estacionaria.

La modelización propuesta y las herramientas complementarias pueden adaptarse a otras regiones. Los modelos son altamente adaptables, lo que permite diferentes variables de respuesta en contextos espaciotemporales, consideración de covariables geográficas adicionales y la inclusión o exclusión de procesos espaciales específicos según sea necesario. Los flexibles términos autorregresivos pueden expresar el comportamiento en series donde la correlación serial es una fuente importante de variación. En general, el

proyecto proporciona herramientas valiosas para comprender y abordar los efectos del cambio climático en la parte central de la distribución, así como en los EHEs, cuantiles y eventos récord, con una amplia gama de aplicaciones en estudios climáticos.

5.2 Limitaciones, oportunidades de mejora y trabajo futuro

Datos con n grande frente a pequeño. Una limitación del análisis actual es la cobertura espacial relativamente escasa, con solo 18 estaciones de monitorización en Aragón y 40 en España. Este muestreo espacial limitado puede plantear desafíos para caracterizar con precisión las superficies espaciales en la modelización, lo que posiblemente resulte en predicciones espaciales menos precisas. Trabajos futuros en este ámbito podrían explorar diferentes regiones que ofrezcan una red más extensa de ubicaciones de monitorización espacial. Con un gran número de ubicaciones espaciales, las aproximaciones para reducir el rango de los GPs pueden utilizarse para abordar el cuello de botella computacional, e.g., el proceso predictivo gaussiano (Banerjee et al., 2008) o el GP del vecino más cercano (Datta et al., 2016).

Técnicas de fusión para reducción de escala y proyecciones climáticas. Trabajando dentro de la misma región, hay un creciente cuerpo de literatura sobre técnicas de *fusión* de datos, especialmente en el contexto de la asimilación de datos con salidas de modelos por ordenador. Los modelos propuestos para los datos observados de las estaciones podrían ampliarse para incorporar información de los modelos por ordenador medioambientales que proporcionan salidas—en lugar de datos reales—para las celdas de una malla a alguna escala espacial. La fusión de datos tendría como objetivo integrar estas dos fuentes de datos para proporcionar información adicional sobre las respuestas espaciotemporales en la misma región y período de tiempo, con el fin de mejorar la precisión de la interpolación espacial. Sin embargo, alinear capas de datos que pueden diferir en resoluciones espaciales y temporales puede ser un desafío. En este sentido, abordar este problema de desalineación al *reducir la escala* de las salidas de modelos numéricos a observaciones a nivel de puntos se vuelve fundamental. Por ejemplo, una extensión sencilla de los modelos autorregresivos de medias o cuantiles podría incorporar el modelo de *reducción de escala* propuesto por Berrocal et al. (2010), que ajusta los datos observados a las salidas de modelos numéricos utilizando coeficientes que varían espacialmente. Además, puede haber preocupaciones sobre errores de medición al reducir la escala. Para abordar esto, se podrían considerar los dos modelos propuestos por Berrocal et al. (2012): (i) un modelo de reducción de escala

suavizado basado en un campo aleatorio de Markov gaussiano, y (ii) un modelo de reducción de escala suavizado con pesos aleatorios que varían espacialmente. Además, las futuras direcciones de investigación podrían involucrar la proyección de cambios de temperatura bajo diferentes escenarios climáticos, reduciendo la escala de variables atmosféricas, abarcando tanto el modelo de medias como el de cuantiles múltiples, así como la predicción de extensiones espaciales bajo diversas proyecciones climáticas.

Software, ajuste de modelos y MCMC. Los modelos espaciotemporales propuestos ofrecen amplias oportunidades para una inferencia completa, aunque su implementación adecuada podría requerir experiencia en modelado estocástico multinivel. El desarrollo de los algoritmos MCMC necesarios para ajustar estos modelos puede resultar complicado y específico. Actualmente existen numerosas herramientas disponibles para el ajuste de modelos bayesianos, paquetes de R específicos como **spBayes** (Finley et al., 2015; Finley & Banerjee, 2020) o **spTimer** (Bakar & Sahu, 2015), y lenguajes de programación probabilísticos como JAGS (Plummer, 2003) o Stan (Carpenter et al., 2017). Los modelos propuestos no pueden ajustarse con estas herramientas, en el primer caso porque los modelos que proporcionan son muy específicos, y en el segundo caso por el alto coste computacional. El paquete de R **spTReg** está en desarrollo y tiene como objetivo una implementación fácil y general de los modelos de media y cuantiles. El desarrollo de herramientas estadísticas y paquetes de software fáciles de usar para estos modelos beneficiará en gran medida a la comunidad científica en general. Simplificar la implementación de estos modelos permitiría a los investigadores con diversos niveles de experiencia estadística contribuir a este importante campo.

QAR conjunta. Es importante notar que no se han introducido covariables en la modelización QAR conjunta para lograr condiciones simples que eviten el cruce de cuantiles. Para considerar una implementación coherente de covariables, deben imponerse condiciones en el soporte de las covariables. Actualmente, se está trabajando en la integración de nuestro modelo con el trabajo de Yang y Tokdar (2017). El caso QAR univariante con covariables podría abordarse considerando una a priori acotada adecuada en su denominada parametrización sin restricciones. La versión espacial podría basarse en el trabajo de modelización de Chen y Tokdar (2021) con niveles de cuantiles dependientes, pero la inclusión de cuantiles que varían espacialmente está en proceso.

Una crítica evidente del modelo *lineal* de QAR viene de la observación de que, cuando el coeficiente de pendiente QAR depende de τ, debe existir una subregión en el

soporte de Y_t en la que se invierte el orden de los cuantiles. Por lo tanto, la formulación lineal en los retardos de Y_t debe considerarse como una útil aproximación local sobre un soporte acotado. Un remedio para esta suposición sería recurrir a expresiones *no lineales* del modelo QAR. Por ejemplo, Chen et al. (2009) exploran un enfoque de modelos de este tipo basados en especificaciones de cópulas, y Xu y Reich (2023) ofrecen una QR no lineal que expande la función de densidad de probabilidad (pdf, por su acrónimo en inglés) en funciones de base I-spline donde los coeficientes dependientes de covariables se modelan mediante redes neuronales. Mayor flexibilidad tiene un precio, la especificación no lineal pierde la agradable interpretación lineal de los retardos o los efectos de las covariables.

Análisis de récords.   Han surgido numerosas oportunidades en el análisis de eventos récord. Por ejemplo, un enfoque típico en el análisis de puntos de cambio es llevar a cabo la investigación estadística basada en un tamaño de muestra fijo. Alternativamente, Chu et al. (1996) desarrollaron un procedimiento de prueba secuencial para modelos de regresión lineal. El trabajo conjunto actual con Lajos Horváth de la University of Utah consiste en extender su esquema de monitoreo en el contexto de ocurrencias de récord.

Una dirección diferente consiste en la exploración de δ-récords (Gouet et al., 2020), donde la combinación de información con valores cercanos al récord debería mejorar la precisión de la predicción y ofrecer una mejor visión del cambio climático. La idea principal consiste en introducir un modelo probit o logit multinomial para tres categorías: récords, valores que están $-\delta$ unidades por debajo del récord y todos los demás valores. Otra pregunta interesante que surge en el análisis de los récords es la evolución conjunta de los récords superiores e inferiores. Se están considerando modelos similares que tienen en cuenta los récords superiores e inferiores en las temperaturas máximas y mínimas diarias, lo que llevará a un modelo probit o logit multinomial bivariante.

En una dirección similar, se necesita más investigación para estudiar la relación entre los récords y la situación atmosférica expresada por la temperatura del aire a diferentes niveles geo-potenciales, así como con patrones de teleconexión para encontrar mecanismos físicos plausibles. Una vez más, los modelos que incluyen variables atmosféricas como covariables también serían útiles como modelos estadísticos de reducción de la escala, y podrían utilizarse para obtener proyecciones futuras del comportamiento de los récords bajo diferentes escenarios climáticos.

5.3   Observación final

En conclusión, esta Tesis profundizó en la intrincada relación entre el cambio climático y los eventos de temperatura extrema. El proyecto desarrolló varios modelos y metodologías estadísticas para descifrar el cambiante panorama de estos eventos a lo largo del tiempo y el espacio. Los hallazgos subrayan el impacto del cambio climático en las tendencias de temperatura, destacando la necesidad de estrategias integrales para abordar sus repercusiones en la salud pública, la agricultura y la economía. Además, este proyecto ha aportado valiosas ideas en el análisis de eventos extremos y en el desarrollo de herramientas estadísticas innovadoras para su evaluación. El enfoque de modelización jerárquico dentro de un marco bayesiano, que ofrece una perspectiva totalmente basada en modelos, ha demostrado ser fundamental para alcanzar estos objetivos de investigación. Además, herramientas estadísticas más simples, accesibles para científicos y profesionales del clima, pueden volverse populares a medida que el cambio climático continúa teniendo un impacto significativo en la humanidad. Al concluir este capítulo, enfatizo la urgencia de seguir investigando en este campo crítico para comprender, adaptarse y mitigar los efectos del cambio climático en nuestro mundo.

Bibliography

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1998). *Records*. John Wiley & Sons. https://doi.org/10.1002/9781118150412

Bakar, K. S., & Sahu, S. K. (2015). **spTimer**: Spatio-temporal Bayesian modeling using R. *Journal of Statistical Software*, *63*(15), 1–32. https://doi.org/10.18637/jss.v063.i15

Ballerini, R., & Resnick, S. I. (1985). Records from improving populations. *Journal of Applied Probability*, *22*(3), 487–502. https://doi.org/10.2307/3213855

Ballerini, R., & Resnick, S. I. (1987). Records in the presence of a linear trend. *Advances in Applied Probability*, *19*(4), 801–828. https://doi.org/10.2307/1427103

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b17115

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825–848. https://doi.org/10.1111/j.1467-9868.2008.00663.x

Benestad, R. E. (2003). How often can we expect a record event? *Climate Research*, *25*(1), 3–13. https://doi.org/10.3354/cr025003

Benestad, R. E. (2004). Record-values, nonstationarity tests and extreme value distributions. *Global and Planetary Change*, *44*(1–4), 11–26. https://doi.org/10.1016/j.gloplacha.2004.06.002

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, *15*(2), 176–197. https://doi.org/10.1007/s13253-009-0004-z

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics*, *68*(3), 837–848. https://doi.org/10.1111/j.1541-0420.2011.01725.x

Bondell, H. D., Reich, B. J., & Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, *97*(4), 825–838. https://doi.org/10.1093/biomet/asq048

Carlin, B. P., & Herring, A. H. (2015). A conversation with Alan Gelfand. *Statistical Science*, *30*(3), 413–422. https://doi.org/10.1214/15-STS521

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Castillo-Mateo, J. (2022). Distribution-free changepoint detection tests based on the breaking of records. *Environmental and Ecological Statistics*, *29*(3), 655–676. https://doi.org/10.1007/s10651-022-00539-2

Castillo-Mateo, J., Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2023). Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics*, *17*(3), 2305–2325. https://doi.org/10.1214/22-AOAS1719

Castillo-Mateo, J., Asín, J., Cebrián, A. C., Mateo-Lázaro, J., & Abaurrea, J. (2023). Bayesian variable selection in generalized extreme value regression: Modeling annual maximum temperature. *Mathematics*, *11*(3), 759. https://doi.org/10.3390/math11030759

Castillo-Mateo, J., Cebrián, A. C., & Asín, J. (2023a). **RecordTest**: An R package to analyse non-stationarity in the extremes based on record-breaking events. *Journal of Statistical Software*, *106*(5), 1–28. https://doi.org/10.18637/jss.v106.i05

Castillo-Mateo, J., Cebrián, A. C., & Asín, J. (2023b). Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021. *Atmospheric Research*, *293*, 106934. https://doi.org/10.1016/j.atmosres.2023.106934

Castillo-Mateo, J., Gelfand, A. E., Asín, J., Cebrián, A. C., & Abaurrea, J. (*in press*). Bayesian joint quantile autoregression. *TEST*.

Castillo-Mateo, J., Gelfand, A. E., Gracia-Tabuenca, Z., Asín, J., & Cebrián, A. C. (2023). Spatio-temporal modeling for record-breaking temperature events in Spain. *Manuscript submitted for publication*.

Castillo-Mateo, J., Lafuente, M., Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2022). Spatial modeling of day-within-year temperature time series: An examination of daily maximum temperatures in Aragón, Spain. *Journal of Agricultural, Biological and Environmental Statistics*, *27*, 487–505. https://doi.org/10.1007/s13253-022-00493-3

Castillo-Mateo, J. (2023a). ***QAR****: Quantile autoRegression* [R package version 0.0.0-2]. https://github.com/JorgeCastilloMateo/QAR

Castillo-Mateo, J. (2023b). ***RecordTest****: Inference Tools in Time Series Based on Record Statistics* [R package version 2.2.0]. https://CRAN.R-project.org/package=RecordTest

Cebrián, A. C., Asín, J., Castillo-Mateo, J., Gelfand, A. E., & Abaurrea, J. (2023). Assessing space and time changes in daily maximum temperature in the Ebro basin (Spain) using model-based statistical tools [arXiv:2211.10784]. *Manuscript submitted for publication*.

Cebrián, A. C., Asín, J., Gelfand, A. E., Schliep, E. M., Castillo-Mateo, J., Beamonte, M. A., & Abaurrea, J. (2022). Spatio-temporal analysis of the extent of an extreme heat event. *Stochastic Environmental Research and Risk Assessment*, *36*, 2737–2751. https://doi.org/10.1007/s00477-021-02157-z

Cebrián, A. C., Castillo-Mateo, J., & Asín, J. (2022). Record tests to detect non-stationarity in the tails with an application to climate change. *Stochastic Environmental Research and Risk Assessment*, *36*, 313–330. https://doi.org/10.1007/s00477-021-02122-w

Chandler, K. N. (1952). The distribution and frequency of record values. *Journal of the Royal Statistical Society: Series B (Methodological)*, *14*(2), 220–228. https://doi.org/10.1111/j.2517-6161.1952.tb00115.x

Chen, X., Koenker, R., & Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *The Econometrics Journal*, *12*(s1), S50–S67. https://doi.org/10.1111/j.1368-423X.2008.00274.x

Chen, X., & Tokdar, S. T. (2021). Joint quantile regression for spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(4), 826–852. https://doi.org/10.1111/rssb.12467

Chu, C.-S. J., Stinchcombe, M., & White, H. (1996). Monitoring structural change. *Econometrica*, *64*(5), 1045–1065. https://doi.org/10.2307/2171955

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag. https://doi.org/10.1007/978-1-4471-3675-0

Coumou, D., & Rahmstorf, S. (2012). A decade of weather extremes. *Nature Climate Change*, *2*(7), 491–496. https://doi.org/10.1038/nclimate1452

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons. https://doi.org/10.1002/9781119115151

Cressie, N. A. C., & Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data* (1st ed.). John Wiley & Sons.

Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, *111*(514), 800–812. https://doi.org/10.1080/01621459.2015.1044091

Davison, A. C., Padoan, S. A., & Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, *27*(2), 161–186. https://doi.org/10.1214/11-STS376

de Haan, L., & Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer. https://doi.org/10.1007/0-387-34471-3

Diersen, J., & Trenkler, G. (1996). Records tests for trend in location. *Statistics*, *28*(1), 1–12. https://doi.org/10.1080/02331889708802543

Finley, A. O., & Banerjee, S. (2020). Bayesian spatially varying coefficient models in the **spBayes** R package. *Environmental Modelling & Software*, *125*, 104608. https://doi.org/10.1016/j.envsoft.2019.104608

Finley, A. O., Banerjee, S., & Gelfand, A. E. (2015). **spBayes** for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, *63*(13), 1–28. https://doi.org/10.18637/jss.v063.i13

Fischer, E. M., Sippel, S., & Knutti, R. (2021). Increasing probability of record-shattering climate extremes. *Nature Climate Change*, *11*(8), 689–695. https://doi.org/10.1038/s41558-021-01092-9

Foster, F. G., & Stuart, A. (1954). Distribution-free tests in time-series based on the breaking of records. *Journal of the Royal Statistical Society: Series B (Methodological)*, *16*(1), 1–22.

Franke, J., Wergen, G., & Krug, J. (2010). Records and sequences of records from random variables with a linear trend. *Journal of Statistical Mechanics: Theory and Experiment*, *2010*(10), P10013. https://doi.org/10.1088/1742-5468/2010/10/P10013

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b16018

Gouet, R., Lafuente, M., López, F. J., & Sanz, G. (2020). Exact and asymptotic properties of δ-records in the linear drift model. *Journal of Statistical Mechanics: Theory and Experiment*, *2020*(10), 103201. https://doi.org/10.1088/1742-5468/abb4dc

Held, L., & Holmes, C. C. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, *1*(1), 145–168. https://doi.org/10.1214/06-BA105

Huser, R., & Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes. *WIREs Computational Statistics*, *14*(1), e1537. https://doi.org/10.1002/wics.1537

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511754098

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, *9*(1), 155–176. https://doi.org/10.1146/annurev-economics-063016-103651

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50. https://doi.org/10.2307/1913643

Koenker, R., & Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, *101*(475), 980–990. https://doi.org/10.1198/016214506000000672

Kotz, S., Kozubowski, T. J., & Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhäuser. https://doi.org/10.1007/978-1-4612-0173-1

Lionello, P., & Scarascia, L. (2018). The relation between climate change in the Mediterranean region and global warming. *Regional Environmental Change*, *18*, 1481–1493. https://doi.org/10.1007/s10113-018-1290-1

Lum, K., & Gelfand, A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, *7*(2), 235–258. https://doi.org/10.1214/12-BA708

Mateo-Lázaro, J., Castillo-Mateo, J., García-Gil, A., Sánchez Navarro, J. Á., Santamarta, J. C., & Fuertes-Rodríguez, V. (2022). Impact of emergency drawdown in off-stream brackish reservoirs – the case of La Loteta dam in Spain. *Journal of Hydrology*, *611*, 128025. https://doi.org/10.1016/j.jhydrol.2022.128025

Mateo-Lázaro, J., Castillo-Mateo, J., García-Gil, A., Sánchez-Navarro, J. Á., Fuertes-Rodríguez, V., & Edo-Romero, V. (2020). Comparative hydrodynamic analysis by using two–dimensional models and application to a new bridge. *Water*, *12*(4), 997. https://doi.org/10.3390/w12040997

Mateo-Lázaro, J., Castillo-Mateo, J., Sánchez-Navarro, J. Á., Fuertes-Rodríguez, V., García-Gil, A., & Edo-Romero, V. (2018). New analysis method for continuous base-flow and availability of water resources based on parallel linear reservoir models. *Water*, *10*(4), 465. https://doi.org/10.3390/w10040465

Mateo-Lázaro, J., Castillo-Mateo, J., Sánchez-Navarro, J. Á., Fuertes-Rodríguez, V., García-Gil, A., & Edo-Romero, V. (2019). Assessment of the role of snowmelt

in a flood event in a gauged catchment. *Water*, *11*(3), 506. https://doi.org/10.3390/w11030506

Mateo-Lázaro, J., Sánchez-Navarro, J. Á., García-Gil, A., Edo-Romero, V., & Castillo-Mateo, J. (2016). Modelling and layout of drainage-levee devices in river sections. *Engineering Geology*, *214*, 11–19. https://doi.org/10.1016/j.enggeo.2016.09.011

McBride, C. M., Kruger, A. C., & Dyson, L. (2022). Trends in probabilities of temperature records in the non-stationary climate of South Africa. *International Journal of Climatology*, *42*(3), 1692–1705. https://doi.org/10.1002/joc.7329

Nevzorov, V. B. (2001). *Records: Mathematical Theory*. American Mathematical Society.

Plumer, B., & Shao, E. (2023). Heat records are broken around the globe as Earth warms, fast [Available at: https://www.nytimes.com/2023/07/06/climate/climate-change-record-heat.html (Accessed: July 9, 2023)]. *The New York Times*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, *124*, 1–10.

Rahmstorf, S., & Coumou, D. (2011). Increase of extreme events in a warming world. *Proceedings of the National Academy of Sciences*, *108*(44), 17905–17909. https://doi.org/10.1073/pnas.1101766108

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(4), 535–553. https://doi.org/10.1111/j.1467-9876.2011.01025.x

Reich, B. J., Fuentes, M., & Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association*, *106*(493), 6–20. https://doi.org/10.1198/jasa.2010.ap09237

Schliep, E. M., Gelfand, A. E., Abaurrea, J., Asín, J., Beamonte, M. A., & Cebrián, A. C. (2021). Long-term spatial modelling for characteristics of extreme heat events. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(3), 1070–1092. https://doi.org/10.1111/rssa.12710

Tokdar, S. T., & Kadane, J. B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis*, *7*(1), 51–72. https://doi.org/10.1214/12-BA702

Wergen, G., & Krug, J. (2010). Record-breaking temperatures reveal a warming climate. *EPL (Europhysics Letters)*, *92*(3), 30008. https://doi.org/10.1209/0295-5075/92/30008

WMO. (2022). *State of the Global Climate 2021* (tech. rep. WMO-No. 1290). WMO. Geneva, Switzerland.

Xu, S. G., & Reich, B. J. (2023). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, *79*(1), 151–164. https://doi.org/10.1111/biom.13576

Yang, Y., & Tokdar, S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, *112*(519), 1107–1120. https://doi.org/10.1080/01621459.2016.1192545

Yu, G. (2020). *__hexSticker__: Create hexagon sticker in R* [R package version 0.4.9]. https://CRAN.R-project.org/package=hexSticker

Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*(4), 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9

Appendices

Chapter A

Letter of acceptance of the work pending publication

TEST An Official Journal of the Spanish Society of Statistics and Operations Research

Date: October 10, 2023

TO WHOM IT MAY CONCERN

This is to certify that the article titled "Bayesian joint quantile autoregression", co-authored by Jorge Castillo-Mateo (corresponding author), Alan E. Gelfand, Jesús Asín, Ana C. Cebrián and Jesús Abaurrea, has been accepted for publication in the journal TEST (Springer). (Date of acceptance: October 6, 2023.)

ANGULO IBAÑEZ JOSE MIGUEL - 24150979J
Digitally signed by ANGULO IBAÑEZ JOSE MIGUEL - 24150979J
Date: 2023.10.10 14:49:31 +02'00'

José M. Angulo

(Co-)Editor-in-Chief of TEST

Chapter B

Details of the publications and doctoral student contributions

B.1   Journal impact factor and ranking by category

1. **Castillo-Mateo, J.**, Lafuente, M., Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2022). Spatial modeling of day-within-year temperature time series: an examination of daily maximum temperatures in Aragón, Spain. *Journal of Agricultural, Biological and Environmental Statistics, 27*(3), 487–505. https://doi.org/10.1007/s13253-022-00493-3 [arXiv:2201.01687]

    JIF 2022 (JCR): **1.4**. Rank: Q4 (72/92) Biology; Q4 (44/55) Mathematical & Computational Biology; **Q3 (66/125) Statistics & Probability**.

2. Cebrián, A. C., Asín, J., Gelfand, A. E., Schliep, E. M., **Castillo-Mateo, J.**, Beamonte, M. A., & Abaurrea, J. (2022). Spatio-temporal analysis of the extent of an extreme heat event. *Stochastic Environmental Research and Risk Assessment, 36*(9), 2737–2751. https://doi.org/10.1007/s00477-021-02157-z

    JIF 2022 (JCR): **4.2**. Rank: Q2 (37/139) Engineering, Civil; Q3 (28/55) Engineering, Environmental; Q2 (104/274) Environmental Sciences; **Q1 (10/125) Statistics & Probability**; Q2 (27/103) Water Resources.

3. **Castillo-Mateo, J.**, Asín, J., Cebrián, A. C., Gelfand, A. E., & Abaurrea, J. (2023). Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics, 17*(3), 2305–2325. https://doi.org/10.1214/22-AOAS1719

    JIF 2022 (JCR): **1.8**. Rank: **Q2 (44/125) Statistics & Probability**.

4. **Castillo-Mateo, J.**, Gelfand, A. E., Asín, J., Cebrián, A. C., & Abaurrea, J. (*in press*). Bayesian joint quantile autoregression. *TEST*. [arXiv:2305.19080]

JIF 2022 (JCR): **1.3**. Rank: **Q3 (72/125) Statistics & Probability**.

5. Cebrián, A. C., **Castillo-Mateo, J.**, & Asín, J. (2022). Record tests to detect non stationarity in the tails with an application to climate change. *Stochastic Environmental Research and Risk Assessment, 36*(2), 313–330. https://doi.org/10.1007/s00477-021-02122-w

   JIF 2022 (JCR): **4.2**. Rank: Q2 (37/139) Engineering, Civil; Q3 (28/55) Engineering, Environmental; Q2 (104/274) Environmental Sciences; **Q1 (10/125) Statistics & Probability**; Q2 (27/103) Water Resources.

6. **Castillo-Mateo, J.** (2022). Distribution-free changepoint detection tests based on the breaking of records. *Environmental and Ecological Statistics, 29*(3), 655–676. https://doi.org/10.1007/s10651-022-00539-2 [arXiv:2105.08186]

   JIF 2022 (JCR): **3.8**. Rank: Q2 (119/274) Environmental Sciences; Q1 (19/107) Mathematics, Interdisciplinary Applications; **Q1 (13/125) Statistics & Probability**.

7. **Castillo-Mateo, J.**, Cebrián, A. C., & Asín, J. (2023). **RecordTest**: An R package to analyse non-stationarity in the extremes based on record-breaking events. *Journal of Statistical Software, 106*(5), 1–28. https://doi.org/10.18637/jss.v106.i05

   JIF 2022 (JCR): **5.8**. Rank: Q1 (27/110) Computer Science, Interdisciplinary Applications; **Q1 (3/125) Statistics & Probability**.

8. **Castillo-Mateo, J.**, Cebrián, A. C., & Asín, J. (2023). Statistical analysis of extreme and record-breaking daily maximum temperatures in peninsular Spain during 1960–2021. *Atmospheric Research, 293*, 106934. https://doi.org/10.1016/j.atmosres.2023.106934

   JIF 2022 (JCR): **5.5**. Rank: **Q1 (18/94) Meteorology & Atmospheric Sciences**.

B.2   Doctoral student contributions

All the research papers presented reflect the research work developed by Jorge Castillo-Mateo during his time as a doctoral student. They were prepared with the collaboration of his Thesis supervisors, Ana C. Cebrián Guajardo and Alan E. Gelfand; his Thesis mentor, Jesús Asín Lafuente; and colleagues, Jesús M. Abaurrea León, María A. Beamonte San Agustín, Zeus Gracia Tabuenca, Miguel Lafuente Blasco, and Erin

M. Schliep. Jorge Castillo-Mateo's contribution to the publications is embodied in the following tasks: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. In particular, he developed and implemented all the MCMC algorithms and software for the analysis of records. When he was the lead author, he played a key role in conceptualization and methodology and was the main contributor responsible for writing the original draft.