Water Resources Management manuscript No.
(will be inserted by the editor)

Dynamic regression model for hourly river level forecasting under risk situations: An application to the Ebro River.

Cebrián A.C. · Abaurrea J. · Asín J. · Segarra E.

Received: date / Accepted: date

Abstract This work proposes a new statistical modelling approach to forecast the hourly river level at a gauging station, under potential flood risk situations and over a medium-term prediction horizon (around three days). For that aim we introduce a new model, the switching regression model with ARMA errors, which takes into account the serial correlation structure of the hourly level series, and the changing time delay between them. A whole modelling approach is developed, including a two-step estimation, which improves the medium-term prediction performance of the model, and uncertainty measures of the predictions. The proposed model not only provides predictions for longer periods than other statistical models, but also helps to understand the physics of the river, by characterizing the relationship between the river level in a gauging station and its influential factors. This approach is applied to forecast the Ebro River level at Zaragoza (Spain), using as input the series at Tudela. The approach has shown to be useful and the resulting model provides satisfactory hourly predictions, which can be fast and easily updated, together with their confidence intervals. The fitted model outperforms the predictions from other statistical and numerical models, specially in long prediction horizons.

Keywords River level forecast · Dynamic regression models · Correlated errors · Switching regimes · Ebro River

A. C. Cebrián
University of Zaragoza
Tel.: +34-976762584
E-mail: acebrian@unizar.es

J. Abaurrea
University of Zaragoza

J. Asín
University of Zaragoza

E. Segarra
University of Zaragoza

1 Introduction

River level forecasting is an important issue in flood control and water management since it allows the activation of warning systems to mitigate the flood effects in reasonable lead time. However, it is a difficult task due to the complexity of the catchment hydrological systems and to the accuracy required for successful flood management strategies.

Physically based models are successfully used to obtain flow forecasts, but they are computationally expensive, and require spatially and temporally resolved meteorological data and detailed physical descriptions of the catchment (Leahy et al. 2008). Moreover, these large-scale systems may be inaccurate at some points. Many efforts are being made to develop statistical models to forecast river levels and flows. Artificial neural networks (ANNs) have been widely applied for river level forecasting (Yadav et al. 2016; Wei 2016). Nevertheless, the applicability of the ANN models may be limited by the fact that each network has to be specifically optimised and trained for each particular problem. Other types of models have been applied to model river levels and flows: fuzzy logic approaches (Alvisi et al. 2006; Keskin et al. 2006; Sen 2017), wavelet regression (Kisi 2011), bayesian estimation (Xu et al. 2018), support vector regression (Matos et al. 2018), and linear and nonlinear time series models, such as threshold autoregressive processes (Tong et al. 1985; Pedregal et al. 2009; Amiri 2015; Pulukuri et al. 2018). The main limitation of all them is that they provide short-term predictions, only a few time units ahead. Another disadvantages are that some of them, such as ANN, do not allow to find an explicit expression of the relationship between the response and the causative factors (Kisi 2011), and that inference tools and confidence intervals are not easy to obtain.

This work steams from a problem put forward by the *Confederación Hidrográfica del Ebro* (CHE), the water management office of the Ebro River basin (Spain). The CHE needs a statistical model to complement and improve the prediction of the river level provided by the numerical model currently used (CHE 2015). The model leads to an overestimation of the level in some flood risk situations at Zaragoza, the biggest city in the basin. The objective of this work is to develop a statistical model to forecast the hourly level at a gauging station, using the information from an upstream station, in potential flood risk situations. The model should be able to provide predictions up to around three days ahead (the time needed to activate warning systems) which can be easily and frequently updated, and uncertainty measures of those predictions. It is also of interest to characterize the relationship between the variation in the river level and its causative factors, in this case, the level in the upstream station.

To this aim, a new statistical modelling approach in the framework of dynamic regression models is proposed. Regression models with ARMA (autoregressive moving-average) errors allow us to find an explicit relationship between the response and the covariates, taking into account the correlation structure of the series (Brockwell and Davis 2016; Abaurrea et al. 2011). On

the other hand, switching regression models (Hubrich and Terasvirta 2013) are able to represent a nonlinear relationship between the variables. Hence, we introduce the switching regression model with ARMA errors which joins the advantages of both models: it takes into account the serial correlation of the level series, and the changing time delay between two gauging stations, which depends on the river flow. A whole statistical modelling approach is developed, from the definition and selection of adequate covariates to the validation. Concerning the estimation, a two-step approach, which improves the performance of the medium-term predictions is developed.

This model is applied to predict the Ebro River level at Zaragoza from one up to 64 hours ahead, and it is considered that risk situations occur when the river level rises above a certain threshold. The alert system for Zaragoza used to be based on Castejón but, in 2015, a large flood provoked a bypass of the river, leading to partial flow measures at that station (CHE 2015, pp 21). The gauging station of Tudela, with a more reliable record, is recommended now by the CHE as input station. The level at Tudela summarizes all the upstream information, and the tributaries between Tudela and Zaragoza can be considered negligible. Precipitation is neither relevant, given the arid climate in this area.

To sum up, the novelty of the proposed model is that it allows to include complex serial and cross-correlation structures, and a changing relationship between the response and the covariates, what is useful to model the changing time delay between the river level at two locations. It can be easily generalized to include other type of covariates, such as atmospheric factors or the contribution of tributaries of the river. Inference tools to test the influence of those covariates are also available. Finally, the model provides medium-term predictions (up to around three days) and confidence intervals for those predictions.

The paper is organized as follows. Section 2 presents the dynamic regression models and the methodology. Data are described in Section 3. In Section 4, models are fitted and compared, and predictions for different time horizons are obtained. A comparison of the performance of our models and other approaches is shown in Section 5. Finally, Section 6 is devoted to the conclusions.

2 Methodology

This section starts by introducing the switching regression models with ARMA errors. Then, a new two-step estimation procedure is proposed, and the prediction approach and diagnosis tools are described.

2.1 Switching regression models with ARMA errors

Ordinary regression models are a useful prediction tool, but they require independent observations. This limitation may often be avoided by using as a

covariate the lagged dependent variable. However, this approach has problems when the aim is to obtain medium or long-term predictions since the value of that covariate is unknown. An adequate approach in that situation is based on a Regression model with ARMA errors (denoted RARMA herein),

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \ldots + \beta_k X_{k,t} + W_t \tag{1}$$

where Y_t and $X_{1,t}, \ldots, X_{k,t}$ are the response and the covariate series, β_0, $\beta_1, \ldots \beta_k$ the regression coefficients, and the error series W_t is a causal, zero-mean, $ARMA(p, q)$ process, satisfying

$$\Phi_p(B) W_t = \Theta_q(B) Z_t \tag{2}$$

where $Z_t \sim N(0, \sigma^2)$ is an uncorrelated series of normal random variables with zero mean and constant variance, and $\Phi_p(B)$ and $\Theta_q(B)$ are polynomials of order p and q in the backshift operator B, $BZ_t = Z_{t-1}$.

One limitation of the RARMA model is that it assumes that there is a linear relationship between the response and the covariates, and that the time correlation structure does not change over the whole period. Some time series models have been developed to allow nonlinear relationships, for example the threshold moving-average models (Ling et al. 2007) or the dynamic switching regression model by Hubrich and Terasvirta (2013) which includes covariate terms,

$$Y_t = \sum_{i=1}^{R} (\Phi_{pi}(B) Y_t + \beta_{i0} + \beta_{i1} X_{1,t} + \ldots + \beta_{ik} X_{k,t}) I(c_{i-1} < S_t < c_i) + W_t \tag{3}$$

with S_t a transition variable, $c_0 = -\infty, c_1, \ldots, c_{R-1}, c_R = \infty$ the switching thresholds, and $I(A)$ a binary variable indicating the occurrence of A.

Following the previous models, we introduce the Switching Regression model with ARMA errors (SRARMA). This model distinguishes R different and independent regimes defined in terms of a transition variable S_t, and different error correlation structures and cross-correlation relationships between the response and the covariates are fitted in each regime:

$$Y_t = \sum_{i=1}^{R} (\beta_{i0} + \beta_{i1} X_{1,t} + \ldots + \beta_{ik} X_{k,t} + W_{i,t}) I(c_{i-1} < S_t < c_i) \tag{4}$$

where each $W_{i,t}$ is an $ARMA(p_i, q_i)$ process. S_t can be a covariate or any other variable, and different covariates can be included in each regime.

2.2 Model estimation

The estimation of the SRARMA model reduces to the estimation of R independent submodels. Each submodel is a RARMA model, fitted using the observations in a regime, $c_{i-1} < S_t < c_i$, while the other observations are transformed into missing values, to keep the time structure.

The first issue is to determine the number of regimes R, and the thresholds c_i. If there does not exist previous information about how to define the regimes, R and c_i can be selected using a preliminary analysis of the correlation structure. Some graphical tools for this analysis are described in Section 4.2.

2.2.1 ML estimation of RARMA regimes

The most common estimation method in RARMA models is maximum likelihood (ML) (Brockwell and Davis 2016, Chap. 6): given p and q, the coefficient vectors $\Phi_p = (\phi_1, \ldots, \phi_p)$, $\Theta_q = (\theta_1, \ldots, \theta_q)$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ and σ^2 are obtained simultaneously by maximizing the loglikelihood function expressed in terms of the one-step prediction errors. Since p and q must be given in advance, we suggest the modelling approach by Makridakis et al. (2008), where an initial estimation of the ARMA errors $\hat{w}_t$ (obtained using a proxy model AR(1)) is used to select adequate p and q values.

Under mild regularity assumptions and large samples, ML estimators are approximately Normal with variances which are, at least, as small as those of other asymptotically Normal estimators. Moreover, the regression and ARMA coefficients are asymptotically independent. Even if the series Z_t is not Normal, it makes sense to use the Normal loglikelihood as a goodness-of-fit measure, and the resulting estimators are still called ML estimators. The reason is that their large-sample distribution is the same for Z_t i.i.d., regardless of whether they are Normal or not.

Inference on ML models is easy since the covariance matrix of the estimators can be estimated as $(-\mathbf{H})^{-1}$, where $\mathbf{H}$ is the Hessian matrix of the log-likelihood evaluated at its maximum. Then, standard confidence intervals and t-tests to check $\beta_i = 0$ are obtained in the usual way. Applying ML properties, standard errors and confidence intervals for the predictions can also be calculated.

2.2.2 TS estimation of RARMA regimes

Despite the good properties of ML estimators, the resulting models may not be optimal for long-term predictions. In effect, ML estimation maximizes a likelihood based on one-step errors, which does not guarantee that the resulting model gives the best h-step predictions for $h > 1$. RARMA predictions are the sum of the regression and the ARMA predictions. ARMA predictions are based on the serial correlation, so its performance decreases when h increases, unlike regression predictions. Global ML estimation gives preference to the ARMA terms due to their good one-step predictions, while covariates which may be useful for long-term prediction are not significant in the presence of those ARMA terms. To avoid this effect, the following two-step estimation (TS) is proposed.

1. The regression part is estimated by ordinary least squares (OLS). If the matrix $\mathbf{X} = (X_1, \ldots, X_k)$ is non-random, or conditionally on it, even when

the model errors are non-Normal and dependent, the OLS estimators are unbiased with covariance matrix,

$$Cov(\hat{\beta}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Gamma_n}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{5}$$

with $\mathbf{\Gamma_n} = E(WW')$ the covariance matrix of the errors $W' = (W_1, \ldots, W_n)$ (Brockwell and Davis 2016).

2. An ARMA process is fitted to the regression residuals, $\hat{w}_t$, using ML. More details are shown in Section OR.1 (Online Resource 1).
3. $\mathbf{\Gamma_n}$ is estimated using the fitted ARMA model. Then, substituting $\widehat{\mathbf{\Gamma}}_\mathbf{n}$ in (5), an estimation of $Cov(\hat{\beta})$ is obtained.

Using $\widehat{Cov}(\hat{\beta})$ from step 3, a kind of confidence interval for β_i, $\hat{\beta}_i \pm 2s.e.(\hat{\beta}_i)$, can be used to analyse the influence of X_i on the response: X_i is considered relevant if the confidence interval does not contain the value 0. This is not a real test, since the Normal behaviour of the TS estimators is not guaranteed, but it is a useful tool.

2.3 Prediction

The prediction of a SRARMA model reduces to identify the regime where the prediction has to be obtained, and apply the corresponding RARMA model. To calculate a RARMA h-step prediction, the regression and the ARMA prediction $\hat{W}_{t+h}$ are obtained independently, and then combined,

$$\hat{Y}_{t+h} = \hat{\beta}_0 + \hat{\beta}_1 X_{1,t+h} + \ldots + \hat{\beta}_k X_{k,t+h} + \hat{W}_{t+h}. \tag{6}$$

To calculate the regression prediction, only the values $X_{1,t+h}, \ldots X_{k,t+h}$ are needed. If they are unknown, they can be predicted using ARMA processes fitted to the series $(X_{i,t})$ or other models. The error prediction $\hat{W}_{t+h}$ is obtained using standard ARMA techniques.

2.3.1 Confidence intervals of the predictions

With ML estimation, standard errors and confidence intervals of the predictions can be obtained from $\widehat{Cov}(\hat{\beta}) = (-\mathbf{H})^{-1}$. However, the hessian matrix of complex likelihood functions cannot always be computed due to convergence problems.

In those cases, or with TS estimators, confidence intervals of the predictions can be obtained using a resampling bootstrap approach. First, a random resampling with reposition of the residuals (z_t) yields a new series $(\tilde{z}_t^r)$. Then, new predictions are calculated as,

$$\tilde{y}_{t+h} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t+h} + \ldots + \hat{\beta}_k x_{k,t+h} + \tilde{w}_{t+h}^r \tag{7}$$

where $\tilde{w}_{t+h}^r = \hat{\Phi}_p^{-1}(B)\hat{\Theta}_q(B)\tilde{z}_{t+h}^r$.

This step is repeated up to obtain a sample of s predictions at each time t. Percentiles 2.5 y 97.5 of that sample define the lower and upper limits of the 95% prediction interval.

These confidence intervals only take into account the model error variability, not the variability of the parameter estimators, but this effect is low (Harvey 1993).

2.3.2 Prediction updating

Monitoring problems usually require real time h-step predictions, not for one step h but for the whole evolution up to a time horizon H, that is for $h = 1, ..., H$. To reproduce the way the models are used, predictions for different values H are calculated in the following way: first, predictions at the following H times, $t_{n+1}, t_{n+2}, \ldots, t_{n+H}$ are obtained. Then, it is assumed that the response is known up to t_{n+H}, and predictions at $t_{n+H+1}, t_{n+H+2}, \ldots, t_{n+2H}$ are obtained. This process is repeated up to cover the prediction period.

2.4 Diagnosis and goodness measures

Model diagnosis is based on the usual regression and time series diagnostic tools, such as Breusch-Godfrey (BG) test to analyse the serial correlation between the residuals. As goodness-of-fit (GOF) measures, we calculate the square of the correlation coefficient between the observed and the fitted values (R^2), and the mean absolute error (MAE). As goodness-of-prediction (GOP) measures, we calculate the following measures in a testing period: R^2_T, MAE_T, the Nash-Sutcliffe efficiency coefficient (EC_T), the median of the relative absolute errors ($PAE_{T,50}$), and the MAE_T restricted to observations over threshold c ($MAE_{c,T}$). More details are shown in Section OR.2 (Online Resource 1) where, in addition, all the steps of the modelling process are summarized in Figure OR.1.

3 Data

The Ebro basin, located in the North-East of Spain and covering 85.550 km^2, is the largest Spanish hydrological system discharging into the Mediterranean sea, see Figure 1. The Ebro River is the largest and longest (928 km) among those flowing completely in Spain, and it is included in the European Flood Awareness System (Thielen et al. 2009). It is the most diverse basin in the Iberian Peninsula concerning climate and orography. In its middle course, the river flows through the Ebro Depression with a semi-arid climate and mean annual rainfall around 450 mm.

The Ebro is an irregular river with a high annual flow variability: the rises occur from October to May, with a mean flow in February of 418 m^3/s in Zaragoza, and the low water levels in August, with a mean of 43 m^3/s. These

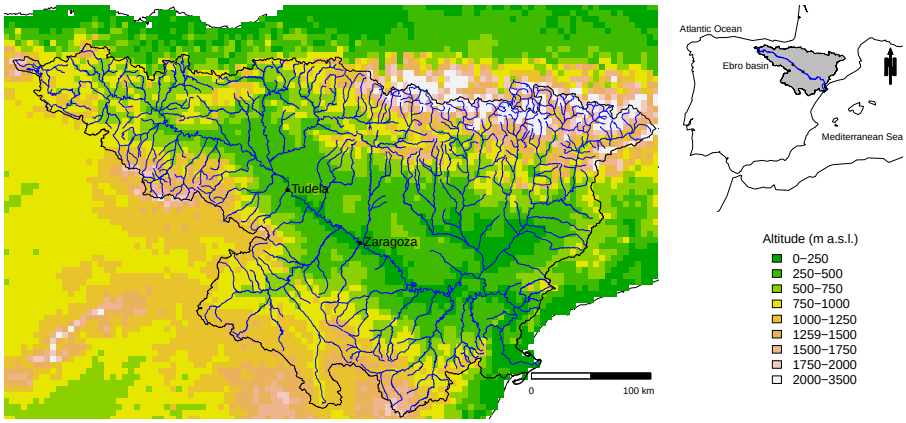

Fig. 1 Ebro basin and location of the gauging stations, Tudela and Zaragoza.

mean values are often exceeded, and the maximum flow in 2015 in Zaragoza was 2610 m^3/s. Floods are recurrent phenomena, and they lead to serious economical and human losses. The implementation of alert and protection measures in Zaragoza requires predictions up to 64 hours ahead. This period, suggested by the CHE, is the maximum time that a flood takes to arrive from Tudela to Zaragoza.

Hourly series of the Ebro River level at Zaragoza and Tudela, approximately 90 km upstream, are measured by the automatic system of hydrological information of the Ebro basin. They are available from 17/11/2004 to 19/06/2016, so that the length of the series is 101591, with 1985 missing observations at Tudela and 15 at Zaragoza. Their mean levels are 1.09 and 1.27 m respectively. The training and testing periods go from 17/11/2004 to 31/12/2014 and from 01/01/2015 to 19/06/2016 respectively.

To model situations under potential flood risk, the series are filtered: times when the level at Tudela is greater or equal to 1.68 m are kept, while the others are changed to missing observations to keep the time structure. The filtered series at Tudela and Zaragoza, Tl_t and Zl_t, have 10571 and 12551 complete observations respectively. The threshold 1.68 m, suggested by the CHE, is the 90th percentile of Tudela, and corresponds to a flow of 500 m^3/s. The value 1.68 guarantees that relevant anthropic effects in the flow are removed. This threshold defines situations of potential risk since real flood danger appears with flows over 1000 m^3/s, that is levels above 2.60 m, at Tudela. A descriptive analysis of the series is shown in Section OR.3.1 (Online Resource 1).

4 Prediction models

A study to define the covariates and regimes of the model is described in this section, and different RARMA and SRARMA models are fitted and compared.

4.1 Defining potential covariates

4.1.1 Seasonal behaviour and trend

Although the level series show a strong seasonal behaviour, it disappears in the filtered series, which only include the values over the threshold. There is neither evidence of significant trends in the series. Moreover, given the hourly scale, if seasonal or trend components remain in Zl_t, they would be the same in Tl_t, so that if a Tl_t covariate is included, additional seasonal or trend terms are not needed. Nevertheless, the stationary assumption is checked in the modelling process, see Sections OR.1 and OR.2.3 (Online Resource 1).

4.1.2 Lagged Tl_t covariates

The level at time t at Zaragoza, Zl_t, is correlated with past values of Tl_t. To explore this relationship, the correlation between Zl_t and the level at Tudela h hours ago, Tl_{t-h}, is analysed for $h = 1, \ldots, 72$ (3 days); the results are shown in Section OR.3.2 (Online Resource 1). Since the maximum correlation, 0.97, occurs at lag 30, Tl_{t-30} is selected as the best lagged covariate. Nearby lags also have a strong influence but, given the high correlation between them, their inclusion would lead to collinearity problems. To deal with this issue, a new type of covariates is suggested.

4.1.3 Moving average covariates

The cross-correlation between Zl_t and Tl_t at different lags can be summarized, in a parsimonious way, by the moving-average of Tl_t over an adequate past period. To this aim, an adequate window length and lag period have to be selected.

A simple way to select the window length is to graphically compare Zl_t and moving-averages of Tl_t with different windows. Section OR.3.3.1 (Online Resource 1) summarizes the results with window lengths from 25 to 97 h. Since $Tlm_t^{0,24}$, the moving-average from time t to $t-24$, shows the most similar evolution to Zl_t, a 25-hour window is selected.

To select the lag period given a window length l, we identify the l lags with the highest cross-correlation. Section OR.3.3.2 (Online Resource 1) shows that in this case, $Tlm_t^{20,44}$ must be selected, with lags from 20 to 44 with a cross-correlation over 0.91. Similar results are obtained with slightly shifted windows.

4.2 Definition of the switching regimes

The mean time delay of the river between Tudela and Zaragoza is around one day but it may increases up to three days since, when the river starts to flood, the overflow runs more slowly. The consequence is that the correlation

structure between Tl_t and Zl_t depends on the river level. The delay between Zl_t and the moving-average variables is less changing, but there is still a nonlinear relationship between Zl_t and $Tlm_t^{20,44}$, as shown in Figure OR.5 (Online Resource 1). A SRARMA model allows to capture this nonlinearity, but the number of regimes and the thresholds $c_1, \ldots, c_R$ have to be previously fixed.

According to hydrological experts, only two regimes are needed in the Ebro River. To confirm this hypothesis, a correlation study is carried out in Section OR.3.4 (Online Resource 1). The results confirm that two regimes should be considered: the high regime (HR), which includes the times where the covariate is over percentile $p_{95} = 3.66$ m, and the low regime (LR), which includes the rest.

Different covariates can be included in each regime of SRARMA models and, as it can be seen in Figure OR.6 (Online Resource 1), a covariate with a higher lag period should be considered in HR. Since the time delay between Tudela and Zaragoza may increase up to around three days, the correlation between Zl_t and moving-average covariates with lag periods up to that time are compared. $Tlm_t^{36,60}$ and $Tlm_t^{20,44}$ are selected in HR and LR since they show the highest correlations, 0.90. and 0.97 respectively.

4.3 Fitted models

Three RARMA and one SRARMA models are summarized here, but more details are shown in Section OR.4 (Online Resource 1). In all the tests, decisions are taken at a significance level $\alpha = 0.05$. Concerning serial correlation, the diagnosis of a model is considered satisfactory, if all the BG p-values in lags $h = 1, ..., 72$ are non-significant.

Model M1 with a lagged covariate. The RARMA model with covariate Tl_{t-30} fitted by ML gives $\hat{\beta}_1 = 0.003$ with $s.e.(\hat{\beta}_1) = 0.005$, so β_1 is not significantly different from 0. This model does not properly capture the correlation structure, with significant BG p-values from lag 12 onwards. The model estimated by TS gives $\hat{\beta}_1 = 1.04$ with $s.e.(\hat{\beta}_1) = 0.015$, and model diagnosis is satisfactory. This result show the advantages of the TS estimation in this type of models.

Model M2 with a moving-average covariate. The RARMA model estimated by ML with $Tlm_t^{20,44}$ gives $\hat{\beta}_1 = 1.04$ and $s.e.(\hat{\beta}_1) = 0.016$, which is significantly different from 0, and model diagnosis is satisfactory. An equivalent model is obtained using TS.

Model M3 with two covariates. The RARMA model with covariates Tl_{t-30} and $Tlm_t^{20,44}$ fitted by TS gives $\hat{\beta}_1 = 0.35$ and $\hat{\beta}_2 = 0.74$ with $s.e.(\hat{\beta}_1) = 0.030$ and $s.e.(\hat{\beta}_2) = 0.027$, suggesting that both covariates are relevant. Model diagnosis is satisfactory.

Model M4 with two regimes. The SRARMA model M4 with submodels M4-LR and M4-HR, including covariates $Tlm_t^{20,44}$ and $Tlm_t^{36,60}$ respectively, gives $\hat{\beta}_{1,LR} = 1.06$ and $\hat{\beta}_{1,HR} = 0.84$ with $s.e.(\hat{\beta}_{1,LR}) = 0.017$ and $s.e.(\hat{\beta}_{1,HR}) =$

0.390; this standard error is larger due to the drastic decrease of the sample size in HR, but the covariate is still relevant. The model diagnosis is satisfactory in both regimes.

4.4 Comparison of the models

Table 1 summarizes the models with details about values p and q, the estimated regression coefficients, their standard errors and GOF and GOP measures for three horizons, H=6, 24 and 64 hours. $MAE_{c,T}$ is calculated for $c = 3.5\ m$ and $c = 5.18\ m$, which separates the high values never observed in the training period.

Two types of covariates are included in the models: a lagged variable, Tl_{t-30}, in M1 and M3, and moving-averages in M2, M3 and M4. The window length is four days in M3, $Tlm_t^{0,96}$, and one day in M2 and M4. The one day moving-averages are also lagged: 20 h in M2, and M4-LR, $Tlm_t^{20,44}$, and 36 h in M4-HR, $Tlm_t^{36,60}$. It is noteworthy that the covariate coefficients and their standard errors in M1, M2 and M4-LR are quite similar, with $\hat{\beta}_1$ close to 1. If the coefficients of the two covariates in M3 are added, a similar value is obtained. This shows the stability of the models, since all the covariates give similar information.

In the 6-hour horizon, the predictions in the four models are very good, but M2 and M4 perform slightly better, with $MAE_T = 0.016\ m$. These predictions are shown in Figure OR.8 (Online Resource 1). Similar conclusions are obtained in the 24-hour horizon: M2 provides the best R_T^2 and EC_T, and M4 the best error measures, with $MAE_T = 0.06\ m$. M4 overestimates the highest peak (observations 550 to 650) what leads to a higher $MAE_{5.2,T}$.

In the 64-hour horizon, see Figure 2, the best predictions are provided by M4 with $MAE_T = 0.098$; this better performance is clearer in high levels, with a 22% reduction in $MAE_{5.2,T}$, with respect to the second best model. As in the 24-hour horizon, M1 provides the worst results, especially in the peaks, where it overestimates the observed values.

To sum up, the four models are satisfactory, although M1 shows the worst performance. M2 and M4 are the best options for H=6, and M4 for longer term predictions. For H=24 and H=64 h, only M4 predicts correctly the time of the peak in March 2015, higher than all the values in the training period, although it slightly overestimates the real value. The other models forecast better the peak value, but predict its occurrence around 30 hours before.

The bottom graphs of Figure 2 show the predictions of M2 and M4 for H=6 and H=64 h respectively, together with a 95% bootstrap confidence interval. The length of the intervals noticeably increases with H and h.

A limitation of these models is that they may require unknown values of the covariates for long-term predictions. For example, M2 and M4-LR, with covariate $Tlm_t^{20,44}$, require not yet observed values of Tudela levels for predictions with $h > 20$, and M4-HR, with $Tlm_t^{36,60}$, for $h > 36$ hours. In this case, the physically-based model used by the CHE provides the values in Tudela. A

Table 1 Fitted models and GOF and GOP measures. MAE measures are in metres.

	Models	M1	M2	M3		M4-LR	M4-HR
	Method	TSE	MLE	TSE		TSE	TSE
	(p, q)	(13,2)	(15,6)	(18,6)		(19,3)	(19,4)
	$\hat{\beta}_0$	0.49	0.46	0.44		0.46	1.11
	$s.e.(\hat{\beta}_0)$	0.033	0.032	0.031		0.037	1.596
	Cov.	Tl_{t-30}	$Tlm_t^{20,44}$	Tl_{t-30}	$Tlm_t^{0,96}$	$Tlm_t^{20,44}$	$Tlm^{36,60}$
	$\hat{\beta}_i$	1.04	1.04	0.35	0.74	1.06	0.84
	$s.e.(\hat{\beta}_i)$	0.015	0.016	0.030	0.027	0.017	0.390
GOF	R^2	0.9996	0.9999	0.9997		0.9997	
	MAE	0.011	0.011	0.009		0.005	
GOP 6 h	R^2_T	0.997	0.998	0.998		0.998	
	EC_T	0.9995	0.9997	0.9997		0.9997	
	$PAE_{T,50}$	0.005	0.003	0.004		0.003	
	MAE_T	0.029	0.016	0.022		0.016	
	$MAE_{3.5,T}$	0.034	0.019	0.026		0.016	
	$MAE_{5.2,T}$	0.058	0.060	0.050		0.062	
GOP 24 h	R^2_T	0.973	0.986	0.983		0.985	
	EC_T	0.9951	0.9976	0.9971		0.9974	
	$PAE_{T,50}$	0.017	0.012	0.014		0.010	
	MAE_T	0.093	0.064	0.074		0.060	
	$MAE_{3.5,T}$	0.140	0.094	0.104		0.088	
	$MAE_{5.2,T}$	0.163	0.124	0.179		0.340	
GOP 64 h	R^2_T	0.961	0.969	0.976		0.979	
	EC_T	0.9930	0.9946	0.9958		0.9965	
	$PAE_{T,50}$	0.030	0.024	0.025		0.025	
	MAE_T	0.128	0.111	0.100		0.098	
	$MAE_{3.5,T}$	0.207	0.169	0.135		0.129	
	$MAE_{5.2,T}$	0.192	0.181	0.194		0.142	

more time-consuming solution, is to use different models for each prediction step h, and include in each model only covariates with known lagged values of Tl_t ($\leq$h).

5 Comparison with other prediction approaches

5.1 Comparison with other models

5.1.1 A physically-based model

A simulation of the physically-based (PhB) model currently used by the CHE is compared with the predictions of M2 for H=6 and M4 for H=64 h, in the period from 1/1/2015 to 15/03/2015. Only this period is available since the PhB predictions are not usually stored. The PhB simulations use all the information provided by the upstream stations, and temperature and rainfall in relevant gauging stations up to 15/03/2015. The predictions are shown in Figure 3 and the GOP measures, calculated in the same evaluation period, in Table 2. The

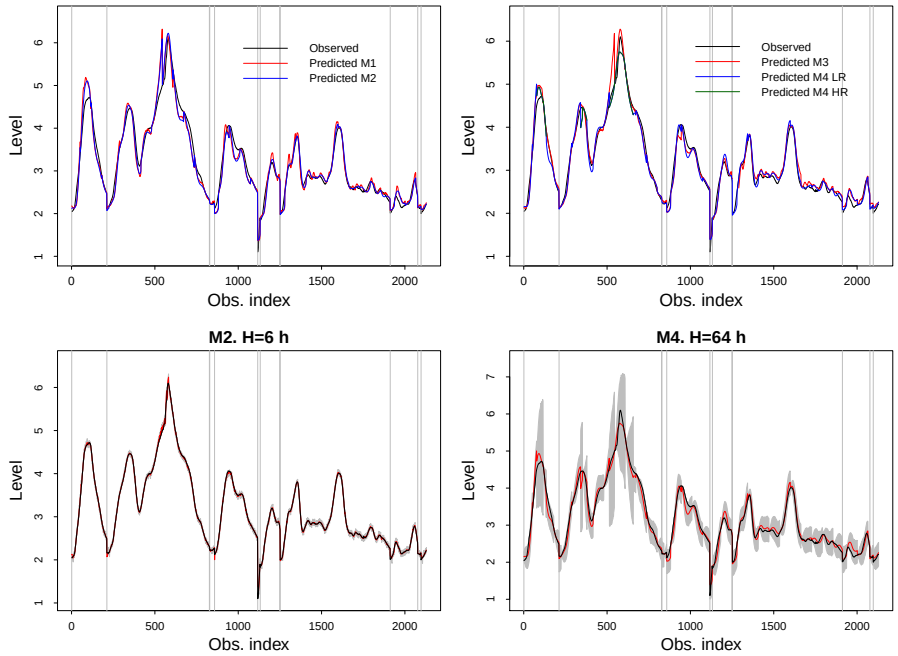

Fig. 2 Top: Observed and predicted levels (in metres) during the testing period for H=64 h, M1 and M2 (left), and M3 and M4 (right). Bottom: Observed (black) and predicted (red) levels and bootstrap 95% confidence intervals (grey) from M2 for H=6 (left), and M4 for H=64 h (right). Observation index gives the day position, with non-risk days removed. Vertical lines separate the consecutive risk periods.

Table 2 GOP measures of PhB and the best statistical models with H=6 and H=64 h.

Model	PhB Model	M2 (H=6 h)	M4 (H=64 h)
R^2_T	0.962	0.999	0.979
EC_T	0.915	0.999	0.977
$PAE_{T,50}$	0.027	0.003	0.025
MAE_T	0.176	0.017	0.112
$MAE_{3.5,T}$	0.242	0.021	0.136
$MAE_{5.2,T}$	0.200	0.060	0.142

measures of the statistical models are always better. MAE_T of the statistical models improves a 36% for H=64 and 90% for H=6 h. In observations over 3.5 m, the improvement increases to 44% and 91% respectively. These results confirm that the proposed models are a beneficial tool, and complement the physically-based model.

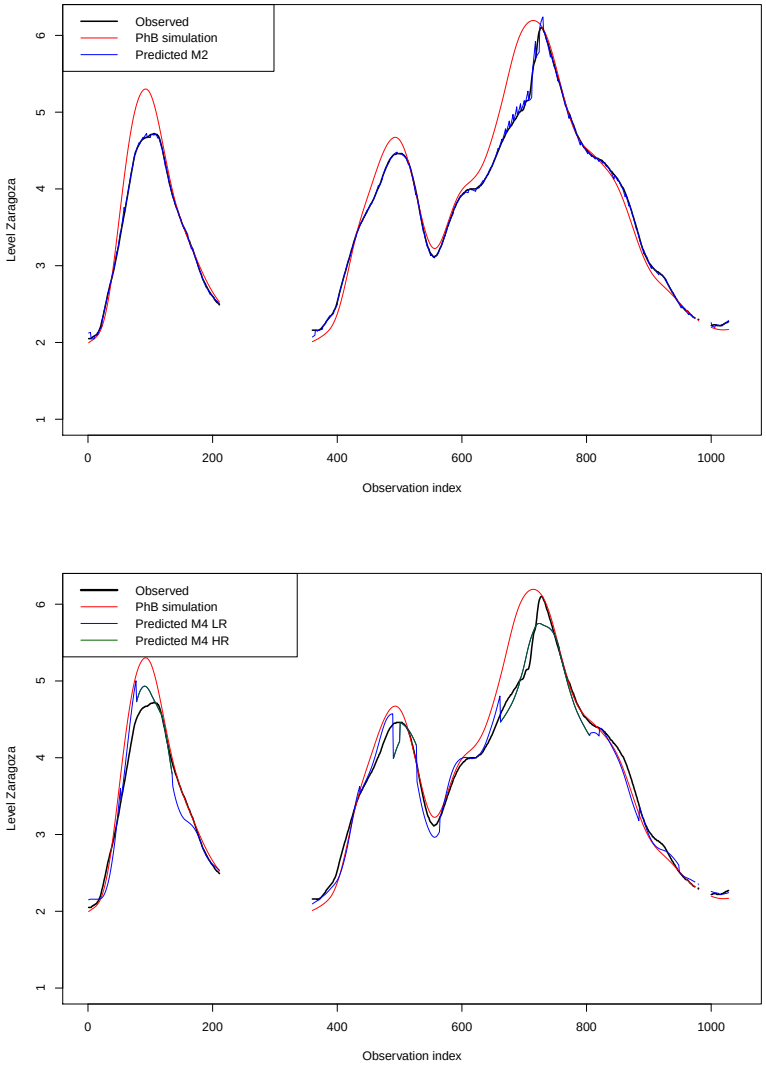

Fig. 3 Observed levels (in metres), PhB simulation and M2 predictions for H=6 (top) and M4 for H=64 h (bottom) during potential risk periods.

5.1.2 An artificial neural network approach

For comparison purposes, standard ANN models are fitted to Zl_t. The four sets of covariates in M1 to M4 are used, to compare predictions based on the same information. Since predictions from ANNs do not depend on the time-horizon, different H values are not considered. Models with different number of nodes

and layers (both between 1 and 2) and two algorithms, resilient backpropagation with weight backtracking, and its modified globally convergent version, are tried (Gunther and Fritsch 2010). The results in all the cases are quite similar, although convergence problems are found in the cases with two covariates and more layers and nodes. The best results are obtained with the covariates in M3, one node and one layer: $R^2_T = 0.969$, $EC_T = 0.969$, $PAE_{T,50} = 0.029$, $MAE_T = 0.123$, $MAE_{3.5,T} = 0.163$ and $MAE_{5.2,T} = 0.338$. The same model with two nodes gives better $MAE_T = 0.113$, but fits worse the higher values, $MAE_{5.2,T} = 0.643$. These values are worse than the results from M2 or M4, even for H=64 h, see Table 1.

5.2 Comparison with other studies

River forecasting usually aims to predict daily levels or flows, and most of the hourly models only provide predictions a few hours ahead. Different GOP measures are used in each work and, to establish a fairer comparison, all our measures are recalculated for a given h-step prediction instead of using a time-horizon H. Only dimensionless measures, such as R^2_T, EC_T or PAE_T, are used to compare level and flow models.

Pedregal et al. (2009) predict hourly flows and levels of the Ebro River in flood events at Zaragoza using the Castejón series, and a nonlinear system based on a State Dependent Parameter approach. It is a regression model with AR errors, where the coefficients are functions of the level. Their predictions give $PAE_{T,50} = 1\%$ for h=6 and 4% for h=24 h, while our corresponding values from M2 and M4 are $PAE_{T,50} = 0.3\%$ and 1%

Aguilar et al. (2013) model a river basin as a set of cascaded interconnected discrete-time linear adaptive models representing the different reaches. The parameters are updated applying an adaptation algorithm, and used to forecast the downstream level of each reach. The method is applied to the Ebro basin, and the hourly level at Zaragoza is forecast using Castejón and Tauste series. The model is tested for h=6, 24 and 41 h on a flood of 29 days in 2009 and gives $MAE_T = 0.024, 0.071$ and 0.122 m, while our models give $MAE_T = 0.009, 0.080$ and 0.103 m in the two floods in 2015 (75 days). Given that the maximum level in 2009 is 4.3, while in 2015 is 6.1 m, the relative error measures of our models would be even better. Their approach does not allow to quantify the prediction uncertainty, and confidence intervals cannot be obtained.

Alvisi et al. (2006) use fuzzy logic and ANN approaches to forecast the Reno River level. For h=6 and 12 h, their predictions give $RMSE_T = 0.16$ and $RMSE_T \approx 0.40$, while we obtain $RMSE_T = 0.05$ and $RMSE_T = 0.04$ m. Leahy et al. (2008) predict the level of an Irish river using ANNs. Their best predictions 5-hour ahead give $R^2_T = 0.98$, while we obtain $R^2_T = 0.999$.

Daily models provide less information, but only require one-step predictions for forecasting the following day while hourly models require 24-step predictions. Despite this disadvantage, our results are competitive. Kisi (2011)

applies wavelet regression to forecast daily river stage of the Schuylkill River (USA), and obtains $R_T^2 = 0.92$ and 0.94; our model M2 gives $R_T^2 = 0.99$ for h=24 h. Amiri (2015) uses nonlinear time series models to predict daily flows of the Colorado River. The EC_T for values over the 95th percentile are 0.993, 0.976 and 0.953 for 1, 2 and 3-day ahead predictions. For values exceeding the 90th percentile, we obtain EC_T =0.998 and 0.997 for h=24 and h=64 h.

To sum up, the SRARMA models lead to better GOP measures than other statistical approaches, such as ANN and fuzzy sets. Moreover, it provides satisfactory predictions for much longer time-horizons than other hourly models, and more detailed predictions than the daily ones.

6 Conclusions

This work develops a statistical procedure based on the switching regression model with ARMA errors to forecast the hourly river level at a station using the level at an upstream station, under potential flood risk situations. The approach is applied to predict the Ebro River level at Zaragoza (Spain) using the information at Tudela. The main conclusions of the work are,

- The hourly scale provides detailed information on the evolution of the level for flood warning systems, although it makes more complicated the modelling process.
- The SRARMA models are a flexible tool to model the river level, and provide satisfactory predictions for longer time-horizons than other models. They model the serial and cross-correlation of the series, and allow nonlinear relationships which represent better the dynamics of the river.
- The moving-average covariates are a parsimonious way of including information of many lags in the model. Tools to select the best window to define them are provided.
- The two-step estimation allows to include in the model covariates which improve long-term predictions, and to model the remaining serial correlation with ARMA errors.
- Model predictions can be easily obtained and updated.
- Confidence intervals of the predictions, and tests to select the significant covariates can be obtained.
- Concerning the results for the Ebro River, model M2, a RARMA model with covariate $Tl_t^{20,44}$, and M4, a SRARMA model with two regimes with covariates $Tl_t^{20,44}$ and $Tl_t^{36,60}$ are the best prediction models. Both give an equivalent performance in short-term predictions, but M4 outperforms M2 in longer term horizons (up to 64 hours). M4 also outperforms other models such as ANNs and the numerical model used by the CHE.

The proposed model can be implemented in other parts of the world, although the relevant covariates may be different, and contributions of large tributaries or atmospheric conditions may be required. However, the model is flexible enough to include them. When the level at an upstream station is

included, the election of the best moving-average may depend on the characteristics of the river, but tools to select the most adequate are provided.

Acknowledgements The authors are members of the research group Modelos Estocásticos, supported by the DGA, the European Social Fund and the project MTM2017-83812-P. The authors acknowledge CHE and especially G. Pérez and J.A. Álvarez for supplying the data and their advice. The authors thank the anonymous referees for their valuable comments.

References

Abaurrea J, Asín J, Cebrián AC, García Vera M (2011) Trend analysis of water quality series based on regression models with correlated errors. J Hydrol 400:341–352

Aguilar JV, Langarita P, Linares L, Gómez M, Rodellar J (2013) An adaptive predictive approach for river level forecasting. J Hydroinform 15(2):232–245

Alvisi S, Mascellani G, Franchini M, Bárdossy A (2006) Water level forecasting through fuzzy logic and artificial neural network approaches. Hydrol Earth Syst Sc 10:1–17

Amiri E (2015) Forecasting daily river flows using nonlinear time series models. J Hydrol 527:1054 – 1072

Brockwell PJ, Davis RA (2016) Introduction to time series and forecasting, 3rd edn. Springer

CHE (2015) Informe sobre las avenidas del primer trimestre de 2015 en la cuenca del Ebro. Tech. rep., Confederación Hidrográfica del Ebro, URL http://www.chebro.es

Gunther F, Fritsch S (2010) neuralnet: Training of neural networks. The R journal 2(1):30–38

Harvey AC (1993) Time Series Models. Harvester Wheatsheaf

Hubrich K, Terasvirta T (2013) Thresholds and smooth transitions in vector autoregressive models. Adv Econom 32:273–326

Keskin M, Taylan D, Terzi O (2006) Adaptive neural-based fuzzy inference system (ANFIS) approach for modelling hydrological time series. Hydrolog Sci J 51(4):588–598

Kisi O (2011) Wavelet regression model as an alternative to neural networks for river stage forecasting. Water Resour Manag 25(2):579–600

Leahy P, Kiely G, Corcoran G (2008) Structural optimisation and input selection of an artificial neural network for river level prediction. J Hydrol 355:192–201

Ling S, Tong H, Li D (2007) Ergodicity and invertibility of threshold moving-average models. Bernoulli 13(1):161–168

Makridakis S, Wheelwright SC, Hyndman RJ (2008) Adv Economing methods and applications. John Wiley & Sons

Matos J, Portela M, Schleiss A (2018) Towards safer data-driven forecasting of extreme streamflows. Water Resour Manag 32:701–720

Pedregal D, Rivas R, Feliu V, Snchez L, Linares A (2009) A non-linear forecasting system for the Ebro river at Zaragoza, Spain. Environ Modell Softw 24(4):502–509

Pulukuri S, Keesara V, Deva P (2018) Flow forecasting in a watershed using autoregressive updating model. Water Resour Manag 32(8):2701–2716

Sen Z (2017) Fuzzy Logic and Hydrological Modeling. CRC Press

Thielen J, Bartholmes J, Ramos MH, Roo AD (2009) The European flood alert system. Part 1: concept and development. Hydrology and Earth System Sciences 13(2):125–140

Tong H, Thanoon B, Gudmundsson G (1985) Threshold time series modeling of two icelandic riverflow systems. J Am Water Resour As 21(4):651–662

Wei CC, (2016) Comparing single-and two-segment statistical models with a conceptual rainfall-runoff model for river streamflow prediction during typhoons. Environ Modell Softw 85:112 – 128

Xu W, Jiang C, Yan L, Li L, Liu S (2018) An adaptive metropolis-hastings optimization algorithm of bayesian estimation in non-stationary flood frequency analysis. Water Resour Manag 32:1343–1366

Yadav B, Ch S, Mathur S, Adamowski J (2016) Discharge forecasting using an online sequential extreme learning machine (OS-ELM) model: A case study in Neckar river, Germany. Measurement 92:433 – 445