scientific data

OPEN

DATA DESCRIPTOR

SC-PREC4SA: A serially complete daily precipitation dataset for South America

Adrian Huerta[1,2] ✉, Roberto Serrano-Notivoli[3] & Stefan Brönnimann[1,2]

This study introduces the serially complete precipitation dataset for South America (SC-PREC4SA), a daily precipitation dataset (1960-2015) designed to address observational gaps and ensure temporal consistency across diverse climates. The raw dataset underwent quality control, gap-filling, and homogenization procedures. Applied robust quality control highlighted common but also overlooked issues, enhancing data reliability. Gap-filling achieved a mean accuracy of 70 % (60 %) in the prediction on wet/dry days (wet-day magnitude). These metrics highlight the reliability of the gap-filling process, particularly in mixed climates, where station networks are sparse. The homogenization algorithm, focused primarily on wet days, effectively reduced inhomogeneities while preserving precipitation variability across South America. By integrating a unified framework and multiple outputs from 7794 stations, SC-PREC4SA provides a robust dataset that captures daily precipitation patterns with high to moderate accuracy and consistency. It offers a valuable resource for climate research, hydrological modeling, and water resource management, addressing longstanding challenges in precipitation data availability and quality for South America.

Background & Summary

Precipitation is a fundamental climate parameter integral to the global water and energy cycles[1,2], with applications across fields such as hydrology, agriculture, climate science, and water resource management. Access to long-term, high-quality precipitation data is essential for analysis and modeling in these areas. While some regions (e.g., Europe and North America) are covered by dense weather station networks, others (e.g., Africa and South America) have sparse and uneven coverage due to economic and geographical constraints. Gridded precipitation data, such as radar, satellite, reanalysis, and merged products[3], are often preferred as they offer spatial continuity, particularly in data-sparse regions. However, these gridded datasets rely on precipitation station data for assimilation and bias correction to enhance their quality, as station data remains the most reliable source for precipitation measurement. Developing serially complete station datasets is therefore crucial, both to improve gridded datasets and to ensure long-term data quality and continuity[4–7].

South America, a vast region with significant meridional reach and prominent orography, encompasses diverse climate and weather patterns, ranging from tropical to extra-tropical zones[8,9]. Its climate is shaped by the Andes Cordillera (the longest continental mountain with an average height ≈ 4000 m) and the Amazon rainforest (the largest rainforest on earth), which play crucial roles in humidity supply and vapor transport across the continent[10]. This diversity is exemplified by the extreme contrast between Colombia's equatorial regions-among the wettest globally[11]-and Chile's Atacama Desert, the driest place on Earth[12]. Such complexity leads to strong spatial-temporal precipitation variability, with marked seasonal patterns in some areas and little to none in others[13,14]. South America is also highly susceptible to extreme events that profoundly impact socio-economic activities, energy demand, and public health. Intense rainfall can lead to flooding and landslides, especially in Andean regions, where lower-income communities in informal housing are particularly vulnerable[15,16]. Given these dynamics, it is essential to understand South America's climate variability, yet South America's precipitation weather station data coverage is limited and inconsistent[17–19]. Differences in data density and management practices across National Meteorological and Hydrological Services (NMHSs), combined with challenging

[1]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland. [2]Institute of Geography, University of Bern, Bern, Switzerland. [3]Departamento de Geografia y Ordenación del Territorio, Instituto Universitario de Ciencias Ambientales (IUCA), Universidad de Zaragoza, Zaragoza, Spain. ✉e-mail: adrhuerta@gmail.com

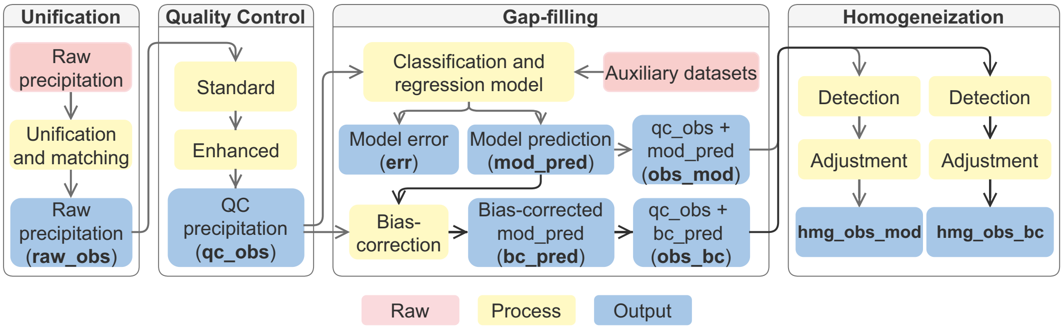

Fig. 1 Schematic overview of the development of a serially complete dataset of daily precipitation for South America (SC-PREC4SA). Raw data, related processes, and main output files are specified.

terrain, highlight the need for comprehensive continental datasets to better address cross-boundary climate and extreme weather events.

Developing serially complete datasets of daily precipitation for South America is a relatively under-researched area. Existing serial datasets are mostly limited to specific regions, such as the Central Andes[20] (Bolivia and Peru) and Patagonia[21] (Argentina and Chile), or have been developed as intermediate products for continental-[22] and country-scale gridded precipitation datasets, such as in Brazil[23], Chile[24], Ecuador[25], and Peru[25–27]. Globally, while more serial datasets are available[28–30], these are typically gridded products that often lack access to the underlying station time series data. Furthermore, few global datasets include serial station data alongside the gridded outputs[7,31]. While global initiatives exist, they frequently face limitations, including incomplete data collection and variable processing approaches. Built largely from international sources, global datasets may lack detailed local information, especially in countries with restricted data-sharing policies[32,33]. Additionally, most global and regional datasets apply only one or two of the essential steps, such as quality control, gap-filling, or homogeneity adjustments, rather than an integrated approach. Creating a unified framework that incorporates all these procedures is essential for fully supporting researchers and users with different needs. Thus, developing a high-quality, serially complete dataset with a comprehensive framework for South America is a critical challenge that remains to be addressed.

In this study, we present a serially complete dataset of daily precipitation for South America (SC-PREC4SA), spanning from 1960 to 2015. This dataset was developed through a consolidated framework that integrates four key processes: unification, quality control, gap-filling, and homogenization. This approach provides researchers and users with long-term, quality-controlled data, along with gap-filled and homogenized time series outputs (each process with its output). As the first dataset of its kind in the region, SC-PREC4SA offers an unprecedented opportunity to comprehensively study precipitation patterns in South America based on observational data.

Methods

Overview. The methodology to produce SC-PREC4SA consists of four major procedures (Fig. 1): unification, quality control, gap-filling, and homogeneity. First, we unified the obtained raw database because multiple stations may share the same data or location (raw_obs). We then used quality control processes to guarantee that the precipitation data was of the highest possible quality (qc_obs). This process employs both a regular process and an enhanced protocol. Next, the gap-filling process was carried out, which involved using auxiliary datasets in conjunction with statistical learning models to provide precipitation predictions (mod_pred and bc_pred) and associated errors (err) for each day and station location of qc_obs. In this step, we created two databases based on the type of prediction involved to fill the gaps (obs_mod and obs_bc). Following that, obs_mod and obs_bc are homogenized independently (detection and adjustment) to ensure temporal variability and reduce potential inhomogeneities of the previously applied process (hmg_obs_mod and hmg_obs_bc). At last, we generated multiple databases as outputs, each of which reflected the results of the key procedures. In the following sections, we present further information and specifics concerning these outputs, as well as the processes used.

Finally, due to the complex climate variability and topography in South America, instead of dividing the continent by countries (Fig. 2a), we used ecological regions (ecoregions) that depict a more accurate regionalization. The ecoregions were as follows (Fig. 2b): Northern Andes (NAS), Peruvian/Atacaman Deserts (PAD), Central Andes (CAS), Southern Andes (SAS), Amazonian-Orinocan Lowland (AOL), Eastern Highlands (EHL), Gran Chaco (GCH), Pampas (PPS) and Monte-Patagonian (MPN). We obtained this ecoregion classification from Griffith *et al.*[34], although most of them belong to classification level I, we also decided to add one that belongs to classification level II (PAD). This was done to differentiate how extremely arid that region was[12]. In addition, we reclassified a small part of EHL within AOL to ensure better spatial discretization.

Data. *Precipitation raw data.* The raw precipitation database used in this study belongs to different sources that come from NMHSs of South America (Fig. 2a) and global databases such as Latin America as the Climate Assessment & Dataset (LACA&D)[35] and Global Historical Climatology Network Daily (GHCNd)[32]. In particular, we collected stations from seven NMHSs (Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, and Peru). To fill the gap for other countries (Curacao, French Guiana, Guyana, Paraguay, Surinam, Trinidad and Tobago, Uruguay, and Venezuela), LACA&D and GHCNd were also used although some stations were also present for the other

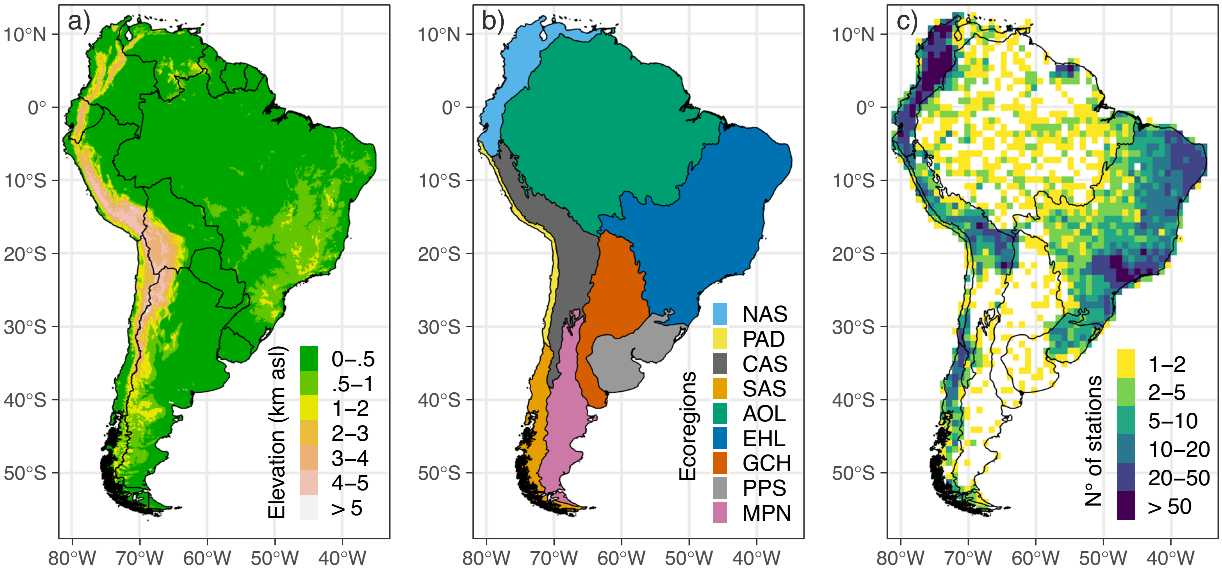

Fig. 2 (**a**) Study area of contiguous South America displaying the elevation and countries as black lines. (**b**) Ecoregions of contiguous South America: Northern Andes (NAS), Peruvian/Atacaman Deserts (PAD), Central Andes (CAS), Southern Andes (SAS), Amazonian-Orinocan Lowland (AOL), Eastern Highlands (EHL), Gran Chaco (GCH), Pampas (PPS), and Monte-Patagonian (MPN). (**c**) The number of raw collected stations in a grid size of 0.9° from 1960 to 2015; black lines represent the ecoregions.

countries (Supplementary Table 1). Therefore, we compiled a large set of precipitation data representing 15161 potential time series for the 1960 - 2015 period. Due to restrictions on South American NMHSs, the full raw data from some sources cannot be distributed with this publication. Readers who wish to obtain the primary data should contact each agency or institution previously mentioned. It is important to note that while a substantial portion of the raw data is openly accessible, several data series remain restricted and can only be accessed upon request (see Data Records and Acknowledgments section). Researchers are referred to revise the data provided by each institution via their official webpage, and for further data requests, contact each agency or institution individually.

South America has a very sparse and uneven spatial distribution of stations. Most (fewer) stations are found in the western and eastern (central) parts of the continent. From an ecoregion perspective (Fig. 2b,c), the three highest (lowest) dense ecoregions were EHL, NAS, and CAS (MPN, GCH, and PAD). Regarding the temporal availability of observations (Supplementary Fig. 1), it is noticed that at the South American scale, the amount of data increased from the 1960s to the 1980s, followed by a decrease until 2015. This increase and decrease pattern was also seen at the ecoregion scale, particularly in NAS, AOL, EHL, GCH, PPS, and GCH. Only PAD, CAS, SAS, and MPN presented a continuous increase from the 1960s.

Auxiliary datasets. In this study, we employed two auxiliary datasets: ERA5-Land precipitation and Digital Elevation Model (DEM)-derived topographic covariables. These were exclusively used for the gap-filling procedure and are detailed below.

The ERA5-Land[36] is an upgraded version of ERA5 designed specifically for land surface applications. It has a finer spatial resolution of 9 km compared to 31 km and 80 km for ERA5 and ERA-Interim, respectively. A triangular mesh-based linear approach is used to interpolate precipitation for ERA5-Land from ERA5. We downloaded the daily aggregated ERA5-Land precipitation (1960 - 2015) from https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR (accessed 25 October 2024).

The DEM-derived topographic covariables were provided by Amatulli *et al.*[37]. They established a suite of topographic covariables based on 7.5 arc-second Global Multi-resolution Terrain Elevation (GMTED2010[38]) data at different spatial resolutions on a global scale. Here, we used the 1 km spatial derived variables such as elevation, slope, aspect cosine, aspect sine, aspect eastness, aspect northness, roughness, topographic position index, terrain ruggedness index, vector ruggedness index, first-order partial derivative (E-W slope), second-order partial derivative (E-W slope), first-order partial derivative (N-S slope), second-order partial derivative (N-S slope), profile curvature and tangential curvature. DEM-derived topographic covariables were downloaded from https://www.earthenv.org/topography (accessed 25 October 2024). Furthermore, we determined other variables such as latitude, longitude, and distance to the ocean at the same spatial resolution. As a result, we used 19 topographic covariables.

Methodology.

Unification. The collected database is the result of merging many sources. As a consequence, duplicated or overlapped stations (similar locations or data) might be identified. Small location discrepancies among such stations may be due to differences in precision in reporting the latitude and longitude coordinates or may reflect different nearby measurement sites[39]. Therefore, some criteria must be applied to eliminate duplicated stations in both the subsequent procedures and the final dataset. In this work, we used a similar unification approach from previously constructed datasets[7], but with a few variations.

- Criteria 1: If the distance between two stations (or more) is less than 10 m, we calculated the correlation and mean absolute error with the five nearby stations. The selected station is the one with the highest correlation and lower mean absolute error.
- Criteria 2: If the distance between two stations (or more) is larger than 10 m but smaller than 25 km, we calculated the correlation, mean absolute error, and the percentage of similar data (excluding values below 0.5 mm). If the correlation exceeds 0.999, the mean absolute error is less than 0.1, and the percentage of identical data exceeds 50%; the station with a longer period is kept. At least two of the conditions should be fulfilled in order to remove a station.
- Criteria 3: The station elevation values are replaced by the nearest grid cell of a 250 m resolution DEM. This process was made due to concerns regarding the accuracy of the raw elevation data. To ensure the station location, we examined the surrounding area. If a station has more than 50% negative DEM values within a 500 m, it is removed. This is done to ensure that the stations are located above sea level (continental area).

The data generated at this stage represents the raw_obs (Fig. 1).

Quality Control. After the station unification, stations are quality-controlled (QC) using two strategies: standard and enhanced QC. The first approach detects (mostly) single suspicious values, while the second addresses recurring data quality issues that may remain undetected by standard QC processes[18,20].

Standard. Previous precipitation QC research provided the basis for the automatic standard QC checks[6,30,40]. Although these QC checks were used globally (or in large regions) with no climate-specific criteria, we decided to establish one based on the percentage of wet days (>0.1 mm). This means that some QC steps were applied differently depending on the type of climate (arid or wet; Supplementary Fig. 2). The standard QC steps are as follows:

- Repetition nonzero check (SQC-01): Daily records were flagged if constant values exceeding 10 mm/d persist for more than four days. If constant values were discovered within a month, all values were flagged.
- Repetition zero check (SQC-02): To detect suspicious zero values in the time series, we evaluated the annual frequency of zero precipitation days and compared it against the climatological frequency. A given year was flagged as suspicious if its zero-frequency exceeded the normal frequency by more than six times the interquartile range, or fell below it by the same threshold. This check was only applied if two conditions were met: (i) at least 85% of the data for that year were available, and (ii) the overall percentage of wet days was greater than 5%.
- Subsequent month duplicated records check (SQC-03): Duplicated daily records in the subsequent months (up to eleven) were identified by calculating a correlation and the number of equal values (excluding values below 0.5 mm). The criteria for the temporal correlation coefficient and the number of days with equal values are set at 0.3 and 10, respectively. If the conditions are met, both the 10 days (target) and duplicates are flagged.
- Subsequent year-month duplicated records check (SQC-04): Duplicated daily records in the same month for the subsequent years (up to eleven) were identified by calculating a correlation and the number of equal values (excluding values below 0.5 mm). The criteria for the temporal correlation coefficient and the number of days with equal values are set at 0.3 and 10, respectively. If the conditions are met, both the 10 days (target) and duplicates are flagged.
- Z-score-based outlier check (SQC-05): Daily records were flagged if their difference from the daily-normal mean was larger than nine sample standard deviations in stations with a percentage of wet days above 5%. The daily-normal mean is calculated from data within a 15-day window centered on all available years (at least 10 years). For stations with a percentage of wet days below 5%, the difference was set three times higher. This step was repeated three times.
- Spatiotemporally isolated value check (SQC-06): A daily record was flagged if it was extreme both in space and time. To meet these conditions, the percentile difference i) with the five nearby stations within a radius of 400 km (space) and ii) with the previous and next day (time) must be greater than the 99.99th percentile.
- Unique or full dry records check (SQC-07): Stations with fewer than 15 unique values or more than 99.5% dry records (<0.5 mm/d) are flagged. This step was only conducted in stations with a percentage of wet days above 5%.

After setting any standard QC-flagged observation as missing, we defined two criteria to select the best stations in terms of the amount of available data: time series i) with at least 10 years of data for each day of the year, and ii) with at least 5 years of continuous data (a year is full if it has 70% of data). The selected stations were used for the enhanced QC.

Enhanced. The enhanced QC process was developed in Hunziker *et al.*[18,20]. They created a comprehensive set of tests that inspect often overlooked issues (truncations, small gaps, asymmetric rounding patterns, and measurement precision inconsistencies, among others) through data visualization techniques that let users manually correct errors or remove specific periods of the dataset. Nevertheless, by increasing the number of weather stations and the size of the study area, the data visualization approach is impractical. In this regard, we automatized the tests by proposing a classification level to describe how good a station was. Therefore, instead of flagging time series periods, we flagged the stations based on the enhanced QC test levels. The enhanced QC tests are as follows:

- Truncation (EQC-01) refers to cases where extreme precipitation events are systematically missing or reduced above a certain threshold, often due to sensor or recording issues. Since no standard algorithm exists to detect truncation, we define it here as the condition where the maximum precipitation values in a series remain constant over a prolonged period (in years). To detect this, we computed a moving maximum value across the daily precipitation series. If this maximum remains unchanged over a predefined number of years, it is flagged as a potential truncation error. Thus, based on the length of years:
  - Level 0: no truncation (a constant maximum value lasts less than 3 years).
  - Level 1: a constant maximum value lasts longer than 3 years but less than 5 years.
  - Level 2: a constant maximum value lasts more than 5 years.
- Small gaps (EQC-02) can be seen as unreported precipitation events that result in a gap or a frequency reduction in values below a specific threshold. To define the small gaps, we calculated the total count of values in five precipitation ranges from 0-1, 1-2, 2-3, 3-4, and 4-5 mm (not including the values in the limits) for each year. Therefore, considering the percentage of years with zero counts:
  - Level 0: no small gaps (0%; years show at least one value in any range).
  - Level 1: small gaps persist in at least 20% of consecutive years.
  - Level 2: small gaps extend for more than 20% of consecutive years.
- Weekly cycles (EQC-03) are characterized by the occurrence of wet days that significantly differ between the days of the week. To compute the weekly cycles, first, for each day of the week, the probability of precipitation is calculated by dividing the total number of wet days by the total counts of values. Later, the number of wet days is tested by a two-sided binomial test (95% confidence level). Based on how many days were significant, we define:
  - Level 0: no atypical weekly cycle (similar probability between the days of the week).
  - Level 1: at least two days present an atypical probability (significant test).
  - Level 2: more than two days present an atypical probability (significant test) or one day presents an extremely different probability (more than 10%).
- Precision and rounding (EQC-04) patterns depict inconsistencies in the frequency of decimal values in the time series. As there is no absolute correct frequency of decimals, we decided to measure how similar the decimal patterns are in the time series. A decimal pattern is interpreted as the list of unique decimal values observed, sorted in descending order. In this way, the decimal pattern for each year is computed first, followed by the selection of the most dominant pattern (mode). Based on how much (in percentage) this dominating pattern represents the time series, we define:
  - Level 0: coherent precision and rounding pattern (similar decimal pattern in more than 70% of the time series).
  - Level 1: a similar decimal pattern in less than 70% but more than 50% of the time series.
  - Level 2: different decimal patterns (no dominant pattern).

Preliminary experiments showed that the automatic enhanced QC could characterize the described issues in both high- and low-quality time series (Supplementary Figures 3 and 4). However, some additions should be made before the application in South America as a whole. This is due to the variety of climates (from wet to extremely arid) and inherent issues with the precision (number of digits in a number)[41] and scale (number of digits to the right of the decimal point in a number) of daily precipitation values (Supplementary Fig. 2). For that purpose, we set to level 0 the application of EQC-01, EQC-02, and EQC-04 (EQC-03) in stations with a percentage of wet days below 15% (5%). Additionally, EQC-02 can be impacted by the length of the decimal patterns (EQC-04), so it is more probable to find small gaps in time series with fewer decimals or with full integers; an issue that is usual in South America. Therefore, we also set level 0 for the application of EQC-02 if the dominant decimal pattern of fewer decimals (less than 5 decimals) is more than 25% in the time series.

Each enhanced QC test indicates that level 0 represents the stations with the fewest quality issues. Ideally, we would select only those stations at level 0 after applying the enhanced QC process. However, allowing some quality issues in subsequent analyses to retain a greater number of stations is a reasonable trade-off. This approach can broaden spatial coverage, which is particularly beneficial in South America, where station density is limited, while still maintaining acceptable data quality. Based on this, we only flagged any station at level 2 in EQC-01, EQC-02, or EQC-03. Therefore, the selected stations in the enhanced QC that intersect the ones in standard QC represent the dataset (qc_obs) used for the following process (Fig. 1). Finally, it should be mentioned that we did not discard the other stations as they remain valid, particularly for the gap-filling process.

Gap-filling. The key process for gap-filling is based on the concept of reference values (RVs) developed by Serrano-Notivoli *et al.*[42,43]. RVs are estimated independently for each day and location using the available data from nearby stations and topographic covariables (latitude, longitude, and elevation). This means that RVs are local models in space and time. Thus, this framework enables the building of highly flexible models that can represent local precipitation conditions.

RVs employ a hybrid modeling approach to predict daily precipitation. The logic is to first use a classification model to predict whether a day is wet or dry, and then apply a regression model only to the wet days to estimate the amount of precipitation. As a result, for each location and day, the RV is based on two predicted values: (i) a binomial prediction (BP) of the probability of occurrence of a wet day and (ii) a magnitude prediction of precipitation (MP), in the case where a wet day is predicted. The combination of these two values (RV = MP if BP > 0.5, else RV = 0) produces the estimated RV and its associated uncertainty (standard error) for each day and location. Generalized linear models (glm) are used as the modeling foundation in RVs.

Previous research demonstrated the usefulness of RVs in dense-stations and moderately-challenging-terrain regions[44–46]. However, some additions and changes must be made before implementing the framework in South America:

- Use of machine learning approaches as the modeling foundation in RVs. Besides glm, we tested support vector machines (svm), random forests (rf), extreme gradient boosting (xgboost), and neural network (nn) models. Preliminary experiments (Supplementary Fig. 5) in small areas evidenced that machine learning models were better than glm. In addition, it was found that glm, rf, and xgboost were more applicable over larger areas (arid to wet) without many fine-tuning adjustments. On the other hand, nn and svm required adjustments to parameters for different areas, making them less generalizable. Although rf and xgboost yielded similar mean efficiency, we opted for xgboost due to the greater number of stations with higher efficiency. Hence, we used xgboost as the modeling foundation for the gap-filling process. It must be pointed out that we used the default xgboost model[47,48] (*nthread* = 1, *nrounds* = 5) as the hyperparameter tuning only offered slight improvements.
- Use virtual stations to enhance the density of the original station network[49,50]. Virtual stations come from the ERA5-Land and were particularly useful for the early period of the dataset. These time series were not directly used, but a bias-correction version (quantile mapping) with the closest station to the grid point. Similarly, the topographic covariables were also obtained. So, we obtained a virtual station for each station, and they were fed into the gap-filling process as real stations.
- Use of more than three topographic covariables in the classification and regression modeling. Besides latitude, longitude, and elevation, we also employed 16 variables listed in section 3.1. To reduce both autocorrelation and dimensionality, we remove the autocorrelated (>0.8) variables first and then retain the first principal components. Based on Horn's parallel analysis[51], we noted that four components explained more than 65% of the variance (Supplementary Fig. 6). In this fashion, we used latitude, longitude, elevation, and the first four principal components as independent variables in the gap-filling process.
- Use of an iterative framework of RVs. This was done to take advantage of those stations that did not have a common period at the beginning of the gap-filling process[49,50,52], particularly in the early period. For the gap-filling procedure, we employed up to three cycles in which we searched for nearby stations within i) 175 km, ii) 275 km, and iii) 650 km. It should be noted that, despite the enormous distance ratio, we limited the number of stations to a minimum of 8 and a maximum of 16 stations for a target station.

From 1960 to 2015, we produced RVs for each station and day under the abovementioned settings. The RVs framework generated two primary outputs: the predicted RV (mod_pred) and standard error (err). However, we also obtained a bias-correction version of the predicted RV named bc_pred. This bias correction (quantile mapping) was conducted in the final stage (of the iterative framework), using the original time series to enhance the model estimation. Therefore, the gap-filling step produces two serially complete daily precipitation series for each station: obs_mod, based on the raw model predictions, and obs_bc, based on the bias-corrected (Fig. 1). These two versions serve as the input for the subsequent homogenization step, which aims to correct potential non-climatic inhomogeneities and generate the final datasets (hmg_obs_mod and hmg_obs_bc).

Evaluation metrics. After calculating mod_pred and bc_pred, we compared them to days with available observations in qc_obs to assess the gap-filling framework's efficiency. This evaluation is leave-one-out cross-validation because the RVs were predicted without using the station's observations. For this purpose, we employed a variety of metrics that evaluate both the continuous and categorical nature of precipitation. The continuous metrics were: the refined index of agreement (dr), mean absolute error (mae), root mean squared error (rmse), normalized mae (nmae), and normalized rmse (nrmse). The categorical metrics were: accuracy, precision, recall, F-measure (f1), balanced accuracy (bcc), and G-mean (g_mean). Despite the amount of metrics, it should be stated that we focus the evaluation on two key metrics: dr and bcc (see Technical Validation section). These are defined as:

- Refined index of agreement[53]:

$$dr = 1 - \frac{\sum_{i=1}^{n}\left|p_i - \hat{y}_i\right|}{2\sum_{i=1}^{n}\left|\hat{y}_i - \overline{y}\right|}$$

where n is the number of observations, p_i is the predicted precipitation on day i, $\hat{y}_i$ is the observed precipitation on day i, and $\overline{y}$ is the mean of $\hat{y}_i$.

- Balanced accuracy[54]:

$$bcc = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

where TP denotes the number of days correctly classified as wet (≥ 0.1 mm), TN as days correctly classified as dry (< 0.1 mm), FP as days incorrectly classified as wet, and FN as days incorrectly classified as dry.

Homogeneity. Non-climatic factors (changes in station location, instrumentation, and observation techniques) can impact measurements. Time series must be homogenized to reduce inhomogeneities and provide more reliable observations[55]. The literature provides a range of homogenization algorithms, particularly for high-density networks with strongly correlated data[55,56]. As a result, homogenization performance worsens significantly when applied in sparse networks, with erroneous corrections likely due to the low signal-to-noise ratio[57]. In this view, the homogenization approach must be tailored to the characteristics of the study area and climate variable.

In this study, we used a similar homogenization strategy based on previously applied approaches on global and continental scales[58,59]. Therefore, we used an automatic algorithm for both detection and adjustment without the use of metadata information. In addition, relative and absolute approaches are combined for situations in which relative homogenization can not be performed. The absolute test, which has a lower power of detection than the relative tests[55], is thus intended as a backup test for when a relative test is hardly possible[59].

Detection. To ensure high confidence in breakpoint detection, we used a combination of different statistical tests and intercomparison of their results. Five univariate breakpoint tests were applied[60]: Student's, Mann-Whitney, Buishand-R, Pettit, and Standard Normal Homogeneity Test. We opted for univariate rather than multivariate tests because they are more straightforward to apply and are less likely to include spurious break detections[61]. Depending on the availability of nearby stations for a target time series:

- Relative: The algorithm searches for up to eight well-correlated (correlation > 0.6) nearby stations within a 1000 km radius. Later, the five tests are applied to difference series (target minus nearby) created with three different temporal aggregations (annual, April-to-September, and October-to-March) and two indices (PRCPTOT: total precipitation; R1mm: number of wet days). Overall, we applied up to 54 different combinations of the approach. Finally, the breakpoint is assigned using the pairwise comparison approach[62]: a breakpoint is set to a certain year if it is found in at least the 7% of the number of difference time series that are significant (p-value < 0.05), using a tolerance of ± 1 years.
- Absolute: The absolute approach is used if the algorithm detects fewer than four nearby stations or none at all. As a result, the five tests are only used in the actual PRCPTOT and R1mm series at the three temporal aggregations. Here, it is used up to nine distinct combinations. Finally, the breakpoint is assigned similarly to the relative approach.

Adjustment. We adapted the quantile-matching technique outlined in Squintu *et al.*[58] to correct the inhomogeneous time series. This method applies adjustments of varying magnitudes based on the value being corrected. Particularly, this approach offers a more robust correction for extreme records, as it tailors adjustments to reflect the intensity of the observed values rather than applying a uniform correction across all dates[63]. It should be mentioned that this algorithm was created for temperature data, therefore, we made some changes to be used for precipitation. Dry values (<0.1 mm) are not corrected, and wet values were transformed twice (square root and log) before the algorithm execution to force a normal distribution. Following the correction, values were reversed to provide the actual precipitation values. Based on this consideration, the correction was applied in two ways:

- Relative: The adjustment factor was computed using the target and nearby time series of the detection stage. It is assumed that the data after the break is correct; thus, the correction is backward. The correction is performed if there is a detected break year, otherwise, the original data is kept.
- Absolute: The adjustment factor was computed using the target time series. This can be seen as an application of quantile mapping, as there are no nearby stations. Data before the break year is corrected based on the quantiles of the sample after the break year. Similarly, the correction is only if there is a detected break year.

The homogeneity framework was applied after gap-filling to: (i) detect inhomogeneities caused by the gap-filling approach, and (ii) because the method was more reliable when there were no gaps in the time series[5,49,50,64]. We apply the homogenization to each ecoregion only once. In addition, adjusted values were set to not exceed one unit difference of the root cubic difference with the raw data. This was done to decrease the influence of the adjustment in the extreme tails[65,66]. It still keeps the extreme adjustment while preventing the creation of extremely excessive values, especially in extremely arid and wet areas. Finally, two new databases have been created: hmg_obs_mod and hmg_obs_bc (Fig. 1), which are homogenized versions of the gap-filled databases obs_mod and obs_bc, respectively.

Data Records

The set of data generated in SC-PREC4SA consists of three key components. For rapid access, the data are divided into different repositories and are stored in a figshare collection[67] (https://doi.org/10.6084/m9.figshare.c.7588178.v4).

- SC-PREC4SA metadata: A file (.csv) that provides information about each station. The file contains the following information (*headers*): station code (*ID*), name (*NAME*), longitude in decimal degrees (*LON*), latitude in decimal degrees (*LAT*), elevation from sources (*ALTs*), elevation from DEM in meters sea above level (*ALT*), country (*COUNTRY*), source (*SOURCE*), ecoregion (*ECOREGIONS*).
- SC-PREC4SA data: Files (.csv) for each station that include the nine daily precipitation (mm/day) outputs (Fig. 1) from 1960 to 2015. Each file contains the following information (*headers*): time step (*time_step*), raw time series (*raw_obs*), quality-controlled time series (*qc_obs*), gap-filling model prediction (*mod_pred*), gap-filling model prediction with a bias correction (*bc_pred*), error of gap-filling model prediction (*err*), quality-controlled time series plus mod_pred (*obs_mod*), quality-controlled time series plus bc_pred (*obs_bc*), homogenized time series of obs_mod (*hmg_obs_mod*); and, homogenized time series of obs_bc (*hmg_obs_bc*). Due to the number of stations, files are subdivided by ecoregions into compressed folders (.zip).
- SC-PREC4SA gap-filling metrics: A file (.csv) provides information about each station's gap-filling evaluation metrics. The file contains the following information (*headers*): station code (*ID*), station ecoregion (*ECOREGIONS*), type of gap-filling model (mod_pred or bc_pred; *MOD*), pairwise number of dates used to calculate the metrics (*n_data*), refined index of agreement (*dr*), mean absolute error (*mae*), root mean squared error (*rmse*), normalized mean absolute error (*nmae*), normalized root mean squared error (*nrmse*), accuracy (*accuracy*), precision (*precision*), recall (*recall*), F-measure (*f1*), balanced accuracy (*bcc*), G-mean (*g_mean*). In addition, the percentage of wet days was added (*wet_day*).

It should be pointed out that we can not share raw data from Bolivia due to data-sharing restrictions. Therefore, we set as missing data (*NA*) the *raw_obs* and *qc_obs* columns in each file that belongs to these countries in the SC-PREC4SA data repository.

Finally, we uploaded the data generated with the other RV foundation models tested in the gap-filling procedure (glm and rf). Those will also be available in the main repository. The purpose of providing these different versions is for further research (see Usage Notes section).

Technical Validation

We report the suitability of SC-PREC4SA by exploring three key procedures: quality control, gap-filling, and homogenization. First, we summarise the results of the applied quality control. Second, we use statistical indicators to evaluate the gap-filling model's efficiency. Lastly, we assess the impact of homogenization by measuring the magnitude of breaks/adjustments as well as the temporal variability of PRCPTOT and R1mm.

Quality control. Following the unification process, we obtained a total of 14624 stations that underwent both standard and enhanced quality control (QC) procedures.

The results of the standard QC process, summarized in Table 1 by ecoregion, are expressed in terms of the number of flagged daily records and their corresponding percentages. Overall, flagged data accounted for less than 0.15% of the total dataset in South America, indicating a minimal portion of flagged values. Most issues stemmed from suspected zero values, duplicate records, and outliers, as identified in steps SQC-02, SQC-03, SQC-04, and SQC-05. At the ecoregion level, MPN exhibited the highest percentage of flagged data, despite having fewer stations (Supplementary Table 2). However, as expected, ecoregions with a higher station density showed greater flagged daily data.

From a temporal perspective (Fig. 3), the proportion of flagged data per year was consistent with the overall average, with minor fluctuations across QC steps. However, pre-1965 data showed elevated percentages, peaking at 0.3-0.4% in 1960 and 1963. This can be attributed to the frequent use of repeated zero values (SQC-02), indicating that early South American precipitation data may include a substantial number of false zeros.

To address systematic data quality issues undetected by standard QC, we implemented enhanced QC procedures. Table 2 presents results from Level 2 of EQC-01 to EQC-04, considered the "worst-case" scenario in terms of quality issues. On average, about 30 % of stations exhibited previously undetected issues, primarily due to small gaps (EQC-02) and precision/rounding patterns (EQC-04). These issues affected more than 15 % of the time series across South America (Supplementary Fig. 7).

At the ecoregion level, every region except PAD exhibited at least ≈ 20 % of stations with undetected issues. Regions such as SAS and GCH (and to a lesser extent CAS, EHL, and PPS) had over 40 % (30 %) of stations affected. The PAD region showed no issues due to its specific wet-day percentage conditions. Overall, the enhanced QC results underscore the presence of significant quality issues in South American precipitation data, potentially impacting processes like gap-filling, homogenization, and previous analyses.

The automatic enhanced QC findings for CAS align closely with previous research. For example, Hunziker *et al.*[20] reported that approximately 40 % of observations were unsuitable for calculating monthly temperature means and precipitation sums due to quality issues. Our study identified similar problems in 34.79 % of precipitation time series. The discrepancy may be due to differences in methods; Hunziker *et al.*[20] manually applied additional tests beyond the four primary tests developed here. These additional tests, which often require visual inspection, were excluded due to challenges in automating such analyses.

The thresholds for the enhanced QC (and standard QC) were designed to account for South America's diverse climates. For example, in extremely arid areas, the lack of wet days poses challenges for reliably identifying patterns in the time series. While Hunziker *et al.*[20] did not encounter such issues in CAS (where wet-day

Standard Quality Contro	Ecoregions									South America
	NAS	PAD	CAS	SAS	AOL	EHL	GCH	PPS	MPN	
SQC-01	154	0	62	22	57	90	0	9	0	394
	0	0	0	0	0	0	0	0	0	0
SQC-02	9134	730	1461	1460	4382	18266	0	365	366	36164
	0.02	0.03	0.01	0.03	0.05	0.04	0	0.02	0.05	0.03
SQC-03	15538	0	1102	852	2450	3066	62	0	60	23130
	0.04	0	0.01	0.02	0.03	0.01	0.01	0	0.01	0.02
SQC-04	22220	56	2658	1038	3890	6529	118	62	60	36631
	0.06	0	0.02	0.02	0.05	0.01	0.02	0	0.01	0.03
SQC-05	7915	681	6717	3319	1143	19196	323	205	725	40224
	0.02	0.03	0.06	0.07	0.01	0.04	0.05	0.01	0.1	0.04
SQC-06	448	24	136	33	125	700	10	26	15	1517
	0	0	0	0	0	0	0	0	0	0
SQC-07	1819	196	1445	235	440	849	64	74	520	5642
	0	0.01	0.01	0	0.01	0	0.01	0	0.07	0
SQC	57228	1687	13581	6959	12487	48696	577	741	1746	143702
	0.16	0.07	0.13	0.14	0.15	0.1	0.08	0.03	0.24	0.13

Table 1. Summary of standard quality control (SQC) process. The number of flagged data and percentage (compared to the total daily data) are displayed for each quality control step by ecoregion.The last column (row) shows the same, but encompassing the entire South America area (SQC steps). The bottom right corner displays the unified results.

percentages range from 10-70%), these challenges are more pronounced in other parts of South America (Supplementary Fig. 7).

To balance data quality and spatial coverage, we flagged only time series classified as Level 2 in EQC-01, EQC-02, or EQC-03. This approach prioritized retaining a larger number of stations while ensuring adequate spatial representation across South America (Supplementary Fig. 7). As a result, 7794 stations (qc_obs) were retained, providing comprehensive coverage in the region. While this represents a reduction in the number of stations, the improvement in data quality significantly outweighs the drawback of fewer observations. Notably, this number of stations is still greater than those in existing global datasets[7,28,29].

Gap-filling. The results of the gap-filling framework were evaluated using statistical metrics computed for each output: mod_pred (model prediction) and bc_pred (bias-corrected prediction). The evaluation focused on the metrics dr (continuous) and bcc (categorical), as these metrics provide a standardized and intuitive measure of model performance (in both > 0.5 means good results). The dr metric was chosen for its broad applicability and resistance to counterbalanced errors, as noted in a previous study[68]. The bcc metric was selected to ensure fair representation of imbalanced and balanced classification classes[54], particularly critical for South America, where extremely arid and wet regions coexist with semi-arid, semi-wet, and mixed climates. By using dr and bcc, the framework effectively addresses both continuous and categorical aspects of precipitation modeling.

The summarized results for dr and bcc by ecoregion (Table 3) revealed that the gap-filling framework performed relatively well across South America, with both metrics exceeding 0.5 on average. This indicates the model's capability to capture general precipitation patterns and reliably classify wet and dry days, despite the climatic diversity. Among the ecoregions, the SAS region exhibited the highest performance, with dr and bcc values consistently above 0.75, while NAS, AOL, and PAD demonstrated the lowest performance, with metric values approaching 0.5 for dr and 0.7 for bcc.

On average, the bias-corrected predictions (bc_pred) improved regression performance, as indicated by higher dr values compared to mod_pred. However, this improvement was not uniform, with the MPN ecoregion showing no evident enhancement. The classification results remained consistent between mod_pred and bc_pred, as the bias correction only altered wet day estimates. Despite improving overall agreement, the bias correction introduced larger errors, as reflected in higher mae and rmse values (Supplementary Table 3). This suggests that while the bias correction reduces systematic bias, it may overcorrect certain high-precipitation days, leading to larger variability in errors. The observed disparity between mod_pred and bc_pred underscores the rationale for providing two gap-filling outputs, allowing users to balance overall accuracy (dr) against precision in individual predictions (mae and rmse).

Visual analysis of dr and bcc at the station level (Fig. 4a) further validated these findings. The framework performed better in semi-arid, semi-wet, and mixed environments compared to extremely arid or wet regions (Fig. 4b). This trend was more pronounced in the classification (bcc) than in regression (dr), particularly in stations with wet day percentages ranging from 10-50% in bcc or 5-30% in dr. Notably, the bias correction aligned with these patterns, as dr values improved within the 5-30% wet-day range. However, some stations in extreme climates (arid) fell below acceptable regression performance thresholds (dr < 0.5).

When compared to previous gap-filling studies on a South American scale, similar patterns were observed. Albeit not from a climate diversity perspective (arid to wet), Tang *et al.*[7] reported better performance in regions

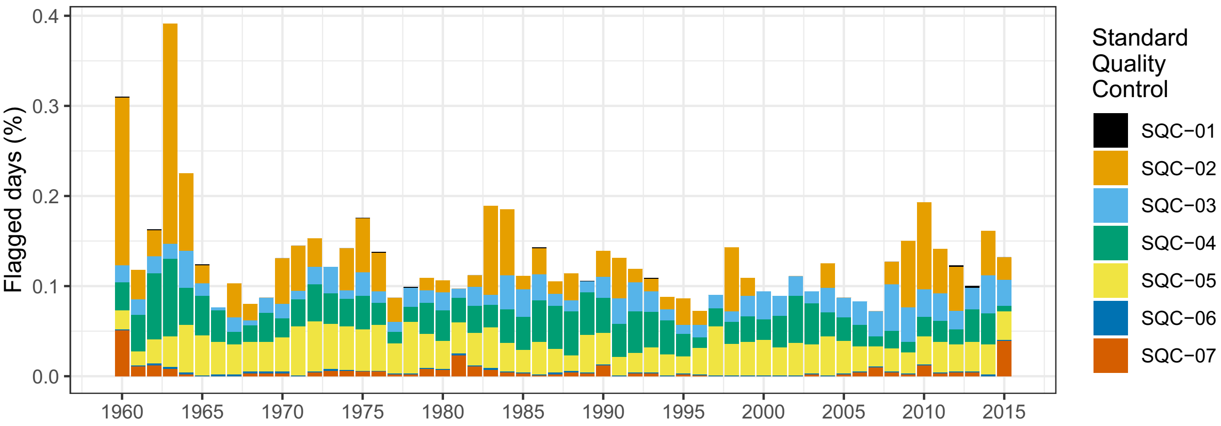

Fig. 3 Flagged daily data (%) by each standard quality control step from 1960 to 2015 in South America.

Enhanced Quality Control	Ecoregions									South America
	NAS	PAD	CAS	SAS	AOL	EHL	GCH	PPS	MPN	
EQC-01	58	0	91	20	10	39	8	2	8	236
	2.02	0	11.68	5.13	1.27	0.83	13.56	0.89	12.9	2.34
EQC-02	249	0	144	40	170	1340	17	63	3	2026
	8.66	0	18.49	10.26	21.66	28.42	28.81	28	4.84	20.09
EQC-03	46	0	6	7	20	26	1	3	2	111
	1.6	0	0.77	1.79	2.55	0.55	1.69	1.33	3.23	1.1
EQC-04	261	0	124	158	79	848	9	38	13	1530
	9.08	0	15.92	40.51	10.06	17.99	15.25	16.89	20.97	15.17
EQC	560	0	271	186	225	1678	26	79	18	3043
	19.47	0	34.79	47.69	28.66	35.59	44.07	35.11	29.03	30.18

Table 2. Summary of enhanced quality control (EQC) process. The number of stations (Level = 2) and percentage (compared to the total stations) are displayed for each quality control step by ecoregion. The last column (row) shows the same, but encompassing the entire South America area (EQC steps). The bottom right corner displays the unified results.

Statistical Metrics	Model output	Ecoregions									South America
		NAS	PAD	CAS	SAS	AOL	EHL	GCH	PPS	MPN	
dr	mod_pred	0.53	0.61	0.61	0.74	0.53	0.63	0.64	0.68	0.63	0.62
	bc_pred	**0.58**	**0.63**	**0.64**	**0.75**	**0.57**	**0.66**	**0.65**	**0.7**	0.63	**0.65**
bcc	mod_pred	0.7	0.68	**0.73**	0.81	0.71	0.74	**0.73**	0.76	0.68	**0.73**
	bc_pred	0.7	0.68	0.72	0.81	0.71	0.74	0.72	0.76	0.68	0.72

Table 3. Summary of gap-filling evaluation metrics: refined index of agreement (dr) and balanced accuracy (bcc). The mean value is displayed for each metric by model output (model prediction without [mod_pred] and with bias-correction [bc_pred]) and ecoregion. The last column displays the mean value at the South American scale. In bold when the statistical metric is best depending on the model output.

with dense station networks and lower performance in sparse networks. This study also identified higher metric values in dense areas such as SAS and southern EHL. However, significant differences emerged regarding the Andes Cordillera (CAS), where Tang *et al.*[7] reported lower performance. This discrepancy may arise from differences in the frameworks: Tang *et al.*[7] relied heavily on the temporal correlation of ERA5 for gap-filling, whereas this study employed flexible local models that relax the need for spatiotemporal correlation, addressing known issues with ERA5 precipitation in complex terrains like the Andes[69].

Overall, the gap-filling framework demonstrated strong performance across diverse climates and terrains in South America, effectively addressing challenges related to station sparsity and climatic variability. Nevertheless, limitations remain, particularly in capturing extreme precipitation events, highlighting the need for further refinement and improvement.

Homogeneity. Following the gap-filling process, two datasets were constructed: obs_mod and obs_bc, representing observations filled with model predictions and bias-corrected predictions, respectively. These datasets underwent a homogenization procedure, resulting in two additional datasets: hmg_obs_mod and hmg_obs_bc.

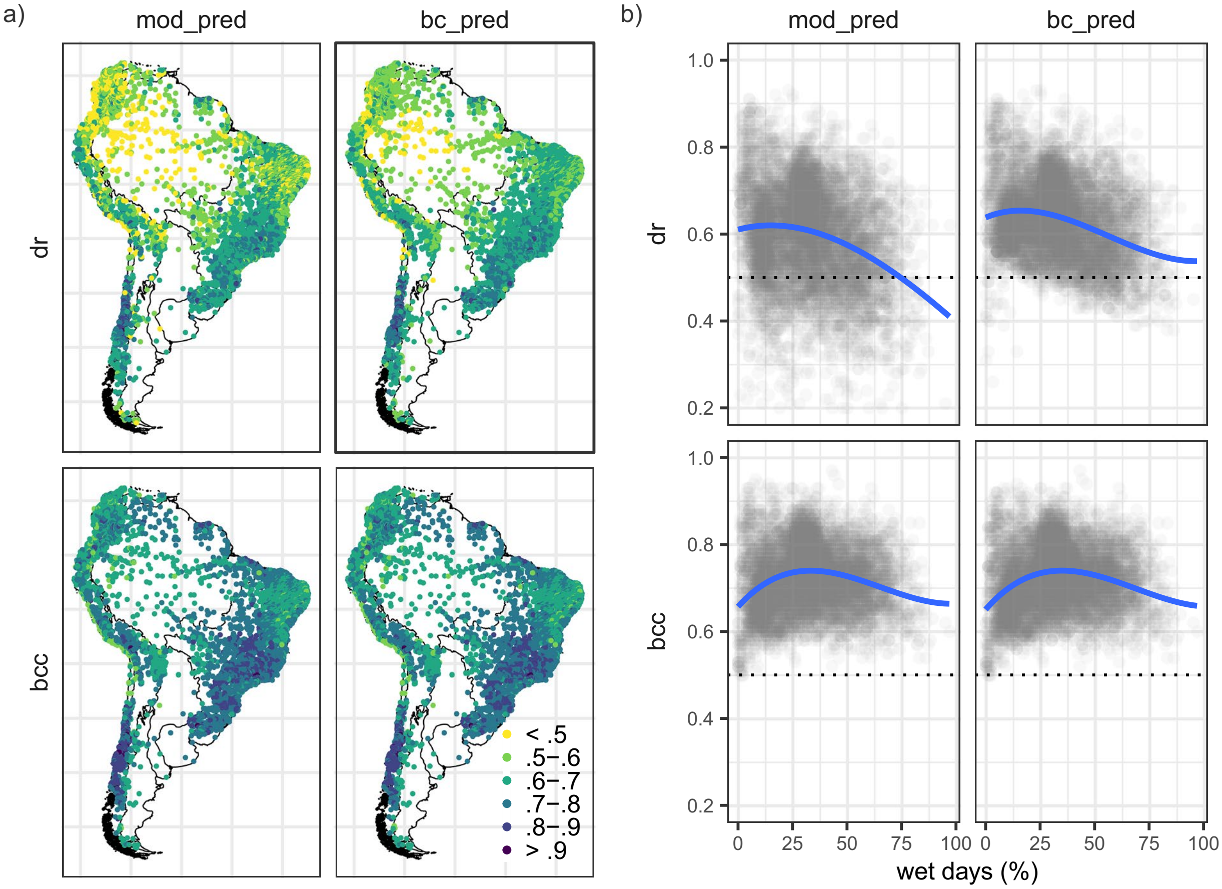

Fig. 4 (**a**) Spatial distribution gap-filling evaluation metrics: refined index of agreement (dr) and balanced accuracy (bcc). (**b**) Relationship between evaluation metrics versus the percentage of wet days (wet days). The blue line in (**b**) represents the trend line between both variables computed using splines. The dotted line in (**b**) displays the 0.5 value in both metrics. Statistical metrics are divided by model output (model prediction without [mod_pred] and with bias correction [bc_pred]).

The Table 4 displays the main results of the applied homogenization procedure: detection and adjustment. At the South American scale, approximately 75 % of the time series employed a relative detection approach, leveraging correlations with nearby stations, while the remaining 25 % relied on the absolute approach due to sparse station networks or challenging terrain. Ecoregions such as SAS, EHL, and PPS showed higher applicability of the relative approach, while areas like CAS and GCH required the absolute method. This reliance on the absolute approach in some regions underscores its utility in sparsely populated or complex terrains where relative detection methods struggle.

Breakpoints were detected in nearly all time series, with more than 95 % of the series presenting statistical significance. PAD exhibited slightly lower rates of breakpoints (≈90%). The temporal distribution of breakpoints varied across ecoregions (Supplementary Fig. 8), with substantial distribution between 1970 and 2010. However, some regions, such as NAS and AOL, displayed frequent breakpoints in the 1970s, whereas GCH and PPS showed peaks in the 2000s. The prominence of breakpoints in indices like PRCPTOT and R1mm suggests that certain inhomogeneities were introduced during the gap-filling stage that may not have been fully accounted for in the gap-filling evaluation. Approximately 48 % of the daily data in both obs_mod and obs_bc datasets is synthetic, contributing to these patterns.

The adjustments applied through the quantile-matching method varied based on precipitation deciles, with mean values ranging between 2 and 4.5 mm across South America. Most adjustments fell within a -0.5 to 0.5 range (difference between roots of cubic), though extreme values near ± 1 were observed (Supplementary Fig. 9), particularly in stations that belong to extremely arid or wet ecoregions. Despite the similarity in the mean adjusted magnitude, we observed slightly higher peaks on the adjusted distribution in PAD, EHL, AOL, PPS, and MPN in obs_mod. The evidence that there were fewer adjusted values in obs_bc could be attributed to the fact that it was bias-corrected (quantile mapping) before the adjustment. This outcome is also supported by a slightly higher number of breakpoints in obs_mod rather than obs_bc (Supplementary Fig. 8).

Homogenization impacts on PRCPTOT and R1mm indices were explored at both ecoregional and continental scales (Figs. 5 and 6). As expected, we noted a higher impact in PRCPTOT rather than R1mm due to the homogeneity focused on (magnitude) wet days rather than on dry days. Although significant breakpoints were detected in the R1mm indices in the precipitation time series, these were not adjusted. So, in general, the

		Ecoregions									South
Homogenization process	**Database**	**NAS**	**PAD**	**CAS**	**SAS**	**AOL**	**EHL**	**GCH**	**PPS**	**MPN**	**America**
Relative detection (%)	obs_mod	55.36	51.81	44.1	82.21	70.95	95.47	41.03	98.09	56	74.69
	obs_bc	63.94	54.92	50.52	86.81	71.28	96.1	41.03	98.09	56	78.51
Significative detection (%)	obs_mod	99.05	90.16	97.57	100	99.16	99.46	97.44	100	96	98.94
	obs_bc	99.01	88.6	95.83	98.47	99.66	99.22	92.31	100	96	98.6
Mean adjustment (mm)	obs_mod	3.32	3.27	1.92	3.09	3.56	3.09	3.94	3.61	1.43	3.19
	obs_bc	3.37	3.42	2.15	3.08	4.35	4.2	4.03	4.19	2.06	3.73

Table 4. Summary of the homogenization process for database and ecoregions: percentage of stations where the relative detection test was performed (relative detection), percentage of stations where a significant break detection was found (significant detection), and the mean value of the applied adjustment (mean adjustment). The last column (row) shows the same, but encompasses the entire South American area.

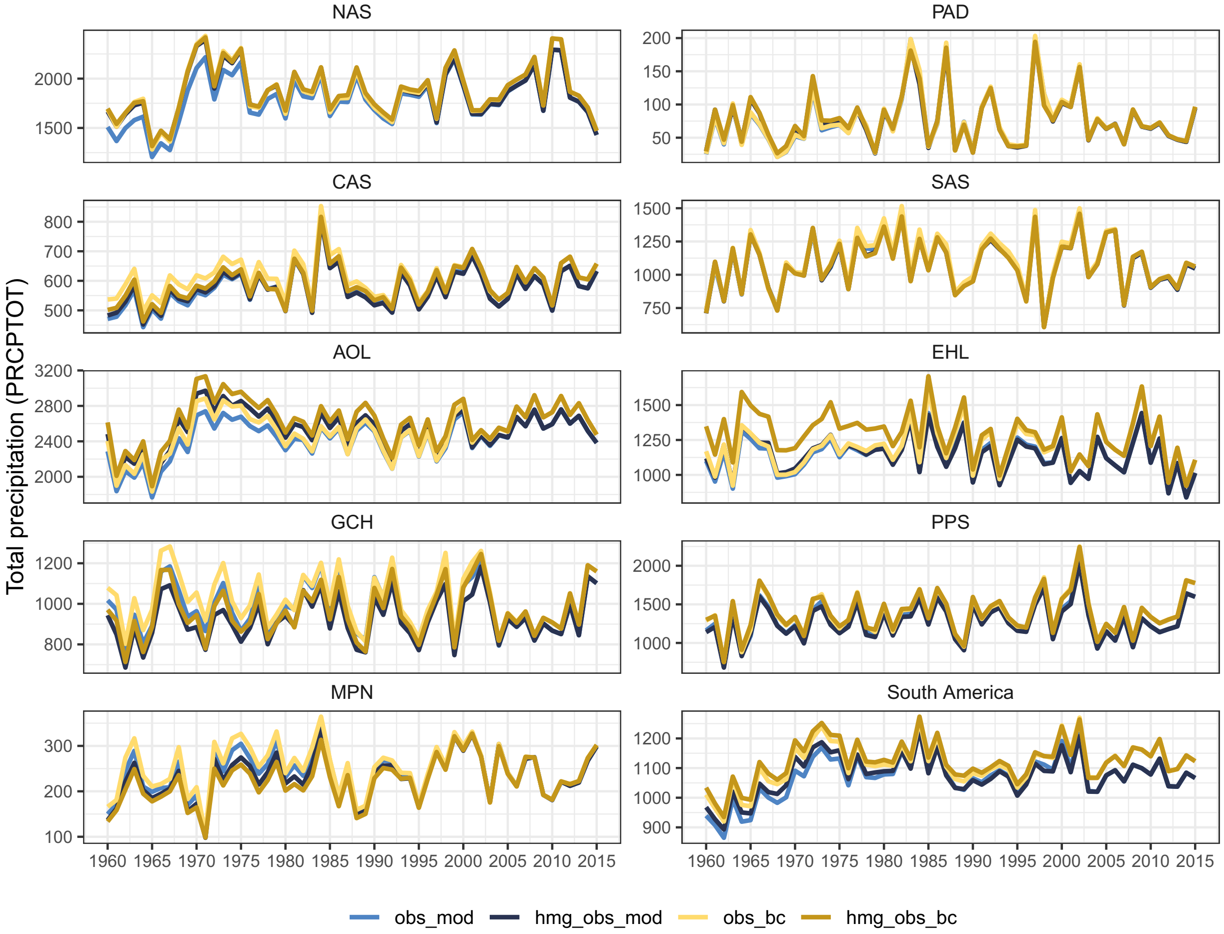

Fig. 5 Mean time series of the total precipitation (mm/year) by ecoregions and South America after (observed data plus model prediction without [hmg_obs_mod] and with bias-correction [hmg_obs_bc]) and before (observed data plus model prediction without [obs_mod] and with bias-correction [obs_bc]) the homogenization process.

adjusted wet day magnitude was done to follow the wet day distribution. Ecoregions like PAD and SAS exhibited similar patterns in both gap-filled and homogenized datasets, while AOL and EHL demonstrated notable differences. The variability of PRCPTOT demonstrates not only the impact of homogenization and gap-filling but also the distribution of missing data (Supplementary Fig. 1).

Regardless of the variability in PRCPTOT (and R1mm) in the different datasets, these revealed important climatological features, such as high and low precipitation years associated with extreme El Niño or La Niña events[70]. Evidence of a climate shift in the 1970s was particularly pronounced in NAS and AOL[71,72], where a marked increase in precipitation was observed. Some regions, such as PAD and EHL, displayed clear fluctuations, implying increasing or decreasing trends, although detailed trend analysis was not performed. These

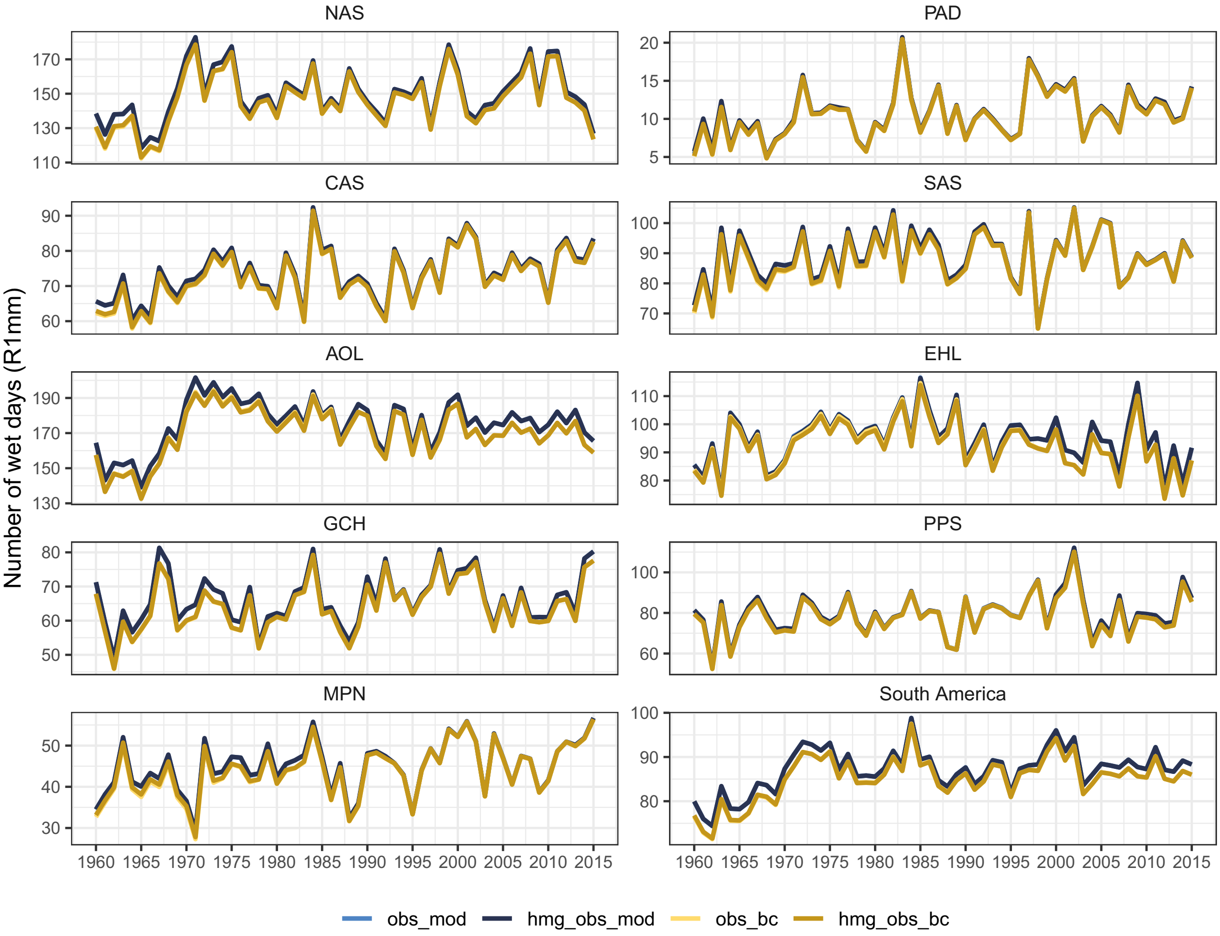

Fig. 6 Mean time series of the number of wet days (days/year) by ecoregions and South America after (observed data plus model prediction without [hmg_obs_mod] and with bias-correction [hmg_obs_bc]) and before (observed data plus model prediction without [obs_mod] and with bias-correction [obs_bc]) the homogenization process.

results suggest that some important features found in previous work are present in the reconstructed data. Nevertheless, more in-depth analysis is required for a better understanding.

In summary, the homogenization process primarily impacted wet days, aligning the datasets with known climatological features. Despite its limitations, including challenges in dry/wet day corrections and variations in PRCPTOT and R1mm, the results provide a robust foundation for understanding precipitation variability in South America. The availability of multiple datasets reflects the inherent uncertainty in a complex region such as South America. The plurality of datasets is an advantage as they provide different views of the same variable

Usage Notes

The SC-PREC4SA database is a very useful dataset for a variety of applications. The single database would simplify access to various datasets, hence improving research. Combining data from several sources into a single repository simplifies analysis, maintains consistency, and facilitates regional study and decision-making. This dataset not only provides different outputs targeted to researchers and practitioners, but it also shares the procedures used to create them. This results in a more consistent, traceable, and reproducible resource, increasing its usefulness for scientific and practical purposes.

Despite rigorous quality control, discrepancies in observation time continue to be an issue. Stations frequently have inconsistent or shifting reporting timings[18], and most databases, including this one, lack the hourly data and metadata required to solve this. While existing methods provide possible solutions[30], they might alter precipitation intensities. Users should be cautious when interpreting results.

The gap-filling process in this dataset primarily relies on ERA5-Land data. Other gridded precipitation datasets, such as satellite-based products, could also be used. Satellite products were not used in this study because they have been mostly available since the 2000s, but their potential to improve precipitation estimation represents an important opportunity. This challenge will be investigated in future studies to improve the accuracy of gap-filled data, particularly in the past and sparse areas.

The dataset was developed using the xgboost model for gap-filling. However, other versions based on glm and rf are also available, allowing users to combine the best estimations from different models using a multi-strategy merge technique. Although this method has been used in previous datasets[6,7], we chose not to employ a multi-strategy merging framework, in this case, to ensure uniformity in error and estimation over the entire region utilizing a single model foundation. This results in a more consistent approach to the dataset's application.

The homogenization process used on the dataset may affect the gap-filling results (regression part), particularly for extreme precipitation indices. Although no abrupt changes were seen in the homogenized PRCPTOT and R1mm mean time series. It is worth noting, however, that we did not test other extreme indices, which may be more sensitive to homogenization processes. Homogenization is especially difficult in locations with sparse station data, where a lack of nearby stations might restrict the detection of inhomogeneities and result in non-larger (or too large) adjustments in time series. In locations with low station density, the uncertainties imposed by homogenization may be more noticeable[57]. These problems highlight the importance of careful interpretation of homogenized data, particularly when working with sparse datasets.

Furthermore, uncertainty values were also calculated in the created dataset, which can be important for improving precipitation data analysis. These uncertainty estimations allow for more robust results since they account for spatial and temporal variations. This approach is consistent with prior efforts that employed uncertainty quantification to improve the reliability of precipitation analyses and extreme event evaluations[43,44]. Users can improve their interpretations of precipitation patterns by incorporating these uncertainty estimates, especially in areas with scarce or inconsistent data.

In line with our philosophy of transparency and flexibility, we provide two versions of the homogenized series: hmg_obs_mod and hmg_obs_bc. While quantile-quantile mapping methods have known limitations, particularly in extreme values, we applied the corrections conservatively to minimize distortions in higher quantiles. Given the complex nature of daily precipitation and its sensitivity to homogenization methods, we recommend that users consider both versions of the dataset. Users are encouraged to explore both versions and assess sensitivity, especially for analyses involving extremes or uncertainty. Alternative gap-filling model outputs (rf, glm) are also available to support this.

Finally, it is crucial to note that an update to SC-PREC4SA is not currently planned. Nonetheless, because the development of SC-PREC4SA is part of the ANDEX program, there are some initiatives[73]. ANDEX proposes solutions to make high-quality information available throughout all Andean countries that meet their objectives. Policies and strategies for collecting data and establishing observational networks are proposed.

Code availability

SC-PREC4SA was constructed using the R (v4.5.0) programming language. The entire code used is freely available at GitHub (https://github.com/adrHuerta/sc-prec4sa) under the GNU General Public License v3.0.

Received: 23 December 2024; Accepted: 2 June 2025;
Published online: 14 June 2025

References

1. Rodell, M. *et al*. The observed state of the water cycle in the early twenty-first century. *Journal of Climate* **28**, 8289–8318, https://doi.org/10.1175/JCLI-D-14-00555.1 (2015).
2. L'Ecuyer, T. S. *et al*. The observed state of the energy budget in the early twenty-first century. *Journal of Climate* **28**, 8319–8346, https://doi.org/10.1175/JCLI-D-14-00556.1 (2015).
3. Sun, Q. *et al*. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics* **56**, 79–107, https://doi.org/10.1002/2017RG000574 (2018).
4. Vicente-Serrano, S. M., Beguería, S., López-Moreno, J. I., García-Vera, M. A. & Stepanek, P. A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. *International Journal of Climatology* **30**, 1146–1163, https://doi.org/10.1002/joc.1850 (2010).
5. Woldesenbet, T. A., Elagib, N. A., Ribbe, L. & Heinrich, J. Gap filling and homogenization of climatological datasets in the headwater region of the Upper Blue Nile Basin, Ethiopia. *International Journal of Climatology* **37**, 2122–2140, https://doi.org/10.1002/joc.4839 (2017).
6. Tang, G. *et al*. SCDNA: A serially complete precipitation and temperature dataset for North America from 1979 to 2018. *Earth System Science Data* **12**, 2381–2409, https://doi.org/10.5194/essd-12-2381-2020 (2020).
7. Tang, G., Clark, M. P. & Papalexiou, S. M. SC-Earth: a station-based serially complete earth dataset from 1950 to 2019. *Journal of Climate* **34**, 6493–6511, https://doi.org/10.1175/JCLI-D-21-0067.1 (2021).
8. Garreaud, R. D., Vuille, M., Compagnucci, R. & Marengo, J. Present-day South American climate. *Palaeogeography, Palaeoclimatology, Palaeoecology* **281**, 180–195, https://doi.org/10.1016/j.palaeo.2007.10.032 (2009).
9. Espinoza, J. C. *et al*. Hydroclimate of the Andes part I: main climatic features. *Frontiers in Earth Science* **8**, 64, https://doi.org/10.3389/feart.2020.00064 (2020).
10. Junquas, C., Li, L., Vera, C., Le Treut, H. & Takahashi, K. Influence of South America orography on summertime precipitation in Southeastern South America. *Climate Dynamics* **46**, 3941–3963, https://doi.org/10.1007/s00382-015-2814-8 (2016).
11. Mejía, J. F. *et al*. Towards a mechanistic understanding of precipitation over the far eastern tropical Pacific and western Colombia, one of the rainiest spots on Earth. *Journal of Geophysical Research: Atmospheres* **126**, e2020JD033415, https://doi.org/10.1029/2020JD033415 (2021).
12. Schween, J. H., Hoffmeister, D. & Löhnert, U. Filling the observational gap in the Atacama Desert with a new network of climate stations. *Global and Planetary Change* **184**, 103034, https://doi.org/10.1016/j.gloplacha.2019.103034 (2020).
13. Ferreira, G. W. & Reboita, M. S. A new look into the South America precipitation regimes: observation and forecast. *Atmosphere* **13**, 873, https://doi.org/10.3390/atmos13060873 (2022).
14. Bazzanela, A. C., Dereczynski, C., Luiz-Silva, W. & Regoto, P. Performance of CMIP6 models over South America. *Climate Dynamics* **62**, 1501–1516, https://doi.org/10.1007/s00382-023-06979-1 (2024).
15. Poveda, G. *et al*. High impact weather events in the Andes. *Frontiers in Earth Science* **8**, 162, https://doi.org/10.3389/feart.2020.00162 (2020).
16. Ozturk, U. *et al*. How climate change and unplanned urban sprawl bring more landslides. *Nature* **608**, 262–265, https://doi.org/10.1038/d41586-022-02141-9 (2022).
17. de los Milagros Skansi, M. *et al*. Warming and wetting signals emerging from analysis of changes in climate extreme indices over South America. *Global and Planetary Change* **100**, 295–307, https://doi.org/10.1016/j.gloplacha.2012.11.004 (2013).

18. Hunziker, S. *et al*. Identifying, attributing, and overcoming common data quality issues of manned station observations. *International Journal of Climatology* **37**, 4131–4145, https://doi.org/10.1002/joc.5037 (2017).
19. Condom, T. *et al*. Climatological and hydrological observations for the South American Andes: in situ stations, satellite, and reanalysis data sets. *Frontiers in Earth Science* **8**, 92 (2020).
20. Hunziker, S. *et al*. Effects of undetected data quality issues on climatological analyses. *Climate of the Past* **14**, 1–20, https://doi.org/10.5194/cp-14-1-2018 (2018).
21. Aguayo, R. *et al*. PatagoniaMet: A multi-source hydrometeorological dataset for Western Patagonia. *Scientific Data* **11**, 6, https://doi.org/10.1038/s41597-023-02828-2 (2024).
22. Rozante, J. R., Moreira, D. S., de Goncalves, L. G. G. & Vila, D. A. Combining TRMM and surface observations of precipitation: technique and validation over South America. *Weather and forecasting* **25**, 885–894, https://doi.org/10.1175/2010WAF2222325.1 (2010).
23. Xavier, A. C., Scanlon, B. R., King, C. W. & Alves, A. I. New improved Brazilian daily weather gridded data (1961–2020). *International Journal of Climatology* **42**, 8390–8404, https://doi.org/10.1002/joc.7731 (2022).
24. Boisier, J. P. *et al*. CR2MET: A high-resolution precipitation and temperature dataset for hydroclimatic research in Chile. In *EGU general assembly conference abstracts*, 19739, https://doi.org/10.5281/zenodo.7529682 (2018).
25. Fernandez-Palomino, C. A. *et al*. A novel high-resolution gridded precipitation dataset for Peruvian and Ecuadorian watersheds: Development and hydrological evaluation. *Journal of Hydrometeorology* **23**, 309–336, https://doi.org/10.1175/JHM-D-20-0285.1 (2022).
26. Aybar, C. *et al*. Construction of a high-resolution gridded rainfall dataset for Peru from 1981 to the present day. *Hydrological Sciences Journal* **65**, 770–785, https://doi.org/10.1080/02626667.2019.1649411 (2020).
27. Huerta, A., Lavado-Casimiro, W. & Felipe-Obando, O. High-resolution gridded hourly precipitation dataset for Peru (PISCOp_h). *Data in Brief* **45**, 108570, https://doi.org/10.1016/j.dib.2022.108570 (2022).
28. Schamm, K. *et al*. Global gridded precipitation over land: A description of the new GPCC First Guess Daily product. *Earth System Science Data* **6**, 49–60, https://doi.org/10.5194/essd-6-49-2014 (2014).
29. Funk, C. *et al*. The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. *Scientific data* **2**, 1–21, https://doi.org/10.1038/sdata.2015.66 (2015).
30. Beck, H. E. *et al*. MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society* **100**, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1 (2019).
31. Tang, G., Clark, M. P. & Papalexiou, S. M. EM-Earth: the ensemble meteorological dataset for planet earth. *Bulletin of the American Meteorological Society* **103**, E996–E1018, https://doi.org/10.1175/BAMS-D-21-0106.1 (2022).
32. Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. & Houston, T. G. An overview of the global historical climatology network-daily database. *Journal of atmospheric and oceanic technology* **29**, 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1 (2012).
33. Lewis, E. *et al*. GSDR: a global sub-daily rainfall dataset. *Journal of Climate* **32**, 4715–4729, https://doi.org/10.1175/JCLI-D-18-0143.1 (2019).
34. Griffith, G. E., Omernik, J. M. & Azevedo, S. H. Ecological classification of the Western Hemisphere. *Unpublished Report***49**, http://ecological-regions.info/htm/sa_eco.htm (1998).
35. Van Den Besselaar, E. J. *et al*. International climate assessment & dataset: Climate services across borders. *Bulletin of the American Meteorological Society* **96**, 16–21, https://doi.org/10.1175/BAMS-D-13-00249.1 (2015).
36. Muñoz-Sabater, J. *et al*. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data* **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021 (2021).
37. Amatulli, G. *et al*. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific data* **5**, 1–15, https://doi.org/10.1038/sdata.2018.40 (2018).
38. Danielson, J. J. & Gesch, D. B. Global multi-resolution terrain elevation data 2010 (GMTED2010), https://doi.org/10.5066/F7J38R2N (2011).
39. Applequist, S., Durre, I. & Vose, R. The Global Historical Climatology Network Monthly Precipitation Dataset, Version 4. *Scientific Data* **11**, 633, https://doi.org/10.1038/s41597-024-03457-z (2024).
40. Hamada, A., Arakawa, O. & Yatagai, A. An automated quality control method for daily rain-gauge data. *Global Environ. Res* **15**, 183–192 (2011).
41. Rhines, A., Tingley, M. P., McKinnon, K. A. & Huybers, P. Decoding the precision of historical temperature observations. *Quarterly Journal of the Royal Meteorological Society* **141**, 2923–2933, https://doi.org/10.1002/qj.2612 (2015).
42. Serrano-Notivoli, R., de Luis, M. & Beguería, S. An R package for daily precipitation climate series reconstruction. *Environmental modelling & software* **89**, 190–195, https://doi.org/10.1016/j.envsoft.2016.11.005 (2017).
43. Serrano-Notivoli, R., de Luis, M., Saz, M. Á. & Beguería, S. Spatially based reconstruction of daily precipitation instrumental data series. *Climate Research* **73**, 167–186, https://doi.org/10.3354/cr01476 (2017).
44. Serrano-Notivoli, R., Beguería, S., Saz, M. Á., Longares, L. A. & de Luis, M. SPREAD: a high-resolution daily gridded precipitation dataset for Spain–an extreme events frequency and intensity overview. *Earth System Science Data* **9**, 721–738, https://doi.org/10.5194/essd-9-721-2017 (2017).
45. Škrk, N. *et al*. SLOCLIM: a high-resolution daily gridded precipitation and temperature dataset for Slovenia. *Earth System Science Data* **13**, 3577–3592, https://doi.org/10.5194/essd-13-3577-2021 (2021).
46. Centella-Artola, A. *et al*. A new long term gridded daily precipitation dataset at high-resolution for Cuba (CubaPrec1). *Data in Brief* **48**, 109294, https://doi.org/10.1016/j.dib.2023.109294 (2023).
47. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794, https://doi.org/10.1145/2939672.2939785 (2016).
48. Chen, T. *et al*. Extreme Gradient Boosting [R Package Xgboost Version 1.2. 0.1]. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*, 13–17, https://cran.r-project.org/web/packages/xgboost/index.html (2020).
49. Huerta, A. *et al*. PISCOeo_pm, a reference evapotranspiration gridded database based on FAO Penman-Monteith in Peru. *Scientific data* **9**, 328, https://doi.org/10.1038/s41597-022-01373-8 (2022).
50. Huerta, A. *et al*. High-resolution grids of daily air temperature for Peru-the new PISCOt v1. 2 dataset. *Scientific data* **10**, 847, https://doi.org/10.1038/s41597-023-02777-w (2023).
51. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185, https://doi.org/10.1007/BF02289447 (1965).
52. Gonzalez-Hidalgo, J. C., Peña-Angulo, D., Brunetti, M. & Cortesi, N. MOTEDAS: a new monthly temperature database for mainland Spain and the trend in temperature (1951–2010). *International Journal of Climatology* **35**, 4444–4463, https://doi.org/10.1002/joc.4298 (2015).
53. Willmott, C. J., Robeson, S. M. & Matsuura, K. A refined index of model performance. *International Journal of climatology* **32**, 2088–2094, https://doi.org/10.1002/joc.2419 (2012).
54. Thölke, P. *et al*. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage* **277**, 120253, https://doi.org/10.1016/j.neuroimage.2023.120253 (2023).
55. Venema, V. K. *et al*. Benchmarking homogenization algorithms for monthly data. *Climate of the Past* **8**, 89–115, https://doi.org/10.5194/cp-8-89-2012 (2012).

56. Guijarro, J. A. *et al*. Homogenization of monthly series of temperature and precipitation: Benchmarking results of the MULTITEST project. *International Journal of Climatology* **43**, 3994–4012, https://doi.org/10.1002/joc.8069 (2023).
57. Gubler, S. *et al*. The influence of station density on climate data homogenization. *International journal of climatology* **37**, 4670–4683, https://doi.org/10.1002/joc.5114 (2017).
58. Squintu, A. A., van der Schrier, G., Brugnara, Y. & Klein Tank, A. Homogenization of daily ECA&D temperature series. *International journal of climatology* **39**, 1243–1261, https://doi.org/10.1002/joc.5874 (2018).
59. Brugnara, Y., Good, E., Squintu, A. A., van der Schrier, G. & Brönnimann, S. The EUSTACE global land station daily air temperature dataset. *Geoscience Data Journal* **6**, 189–204, https://doi.org/10.1002/gdj3.81 (2019).
60. Hurtado, S. I., Zaninelli, P. G. & Agosta, E. A. A multi-breakpoint methodology to detect changes in climatic time series. An application to wet season precipitation in subtropical Argentina. *Atmospheric Research* **241**, 104955, https://doi.org/10.1016/j.atmosres.2020.104955 (2020).
61. Lund, R. B., Beaulieu, C., Killick, R., Lu, Q. & Shi, X. Good practices and common pitfalls in climate time series changepoint techniques: A review. *Journal of Climate* **36**, 8041–8057, https://doi.org/10.1175/JCLI-D-22-0954.1 (2023).
62. Caussinus, H. & Mestre, O. Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society Series C: Applied Statistics* **53**, 405–425, https://doi.org/10.1111/j.1467-9876.2004.05155.x (2004).
63. Brugnara, Y., McCarthy, M. P., Willett, K. M. & Rayner, N. A. Homogenization of daily temperature and humidity series in the UK. *International Journal of Climatology* **43**, 1693–1709, https://doi.org/10.1002/joc.7941 (2023).
64. Tomas-Burguera, M., Vicente-Serrano, S. M., Beguería, S., Reig, F. & Latorre, B. Reference crop evapotranspiration database in Spain (1961–2014). *Earth System Science Data* **11**, 1917–1930, https://doi.org/10.5194/essd-11-1917-2019 (2019).
65. Gutmann, E. *et al*. An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research* **50**, 7167–7186, https://doi.org/10.1002/2014WR015559 (2014).
66. Berg, P. *et al*. Robust handling of extremes in quantile mapping—Murder your darlings-. *Geoscientific Model Development* **17**, 8173–8179, https://doi.org/10.5194/gmd-17-8173-2024 (2024).
67. Huerta, A., Serrano-Notivoli, R. & Brönnimann, S. A serially complete daily precipitation dataset for South America (SC-PREC4SA), https://doi.org/10.6084/m9.figshare.c.7588178.v4 (2024).
68. Cinkus, G. *et al*. When best is the enemy of good–critical evaluation of performance criteria in hydrological models. *Hydrology and Earth System Sciences* **27**, 2397–2411, https://doi.org/10.5194/hess-27-2397-2023 (2023).
69. Imfeld, N. *et al*. Summertime precipitation deficits in the southern Peruvian highlands since 1964. *Int. J. Climatol* **39**, 4497–4513, https://doi.org/10.1002/joc.6087 (2019).
70. Cai, W. *et al*. Climate impacts of the El Niño–southern oscillation on South America. *Nature Reviews Earth & Environment* **1**, 215–231, https://doi.org/10.1038/s43017-020-0040-3 (2020).
71. Carvalho, L. M., Jones, C., Silva, A. E., Liebmann, B. & Silva Dias, P. L. The South American monsoon system and the 1970s climate transition. *international Journal of Climatology* **31**, 1248–1256, https://doi.org/10.1002/joc.2147 (2011).
72. Jacques-Coper, M. & Garreaud, R. D. Characterization of the 1970s climate shift in South America. *International Journal of Climatology* **35**, 2164–2179, https://doi.org/10.1002/joc.4120 (2015).
73. IANIGLA-CONICET & CR^2. Observatorio de Nieve en los Andes de Argentina y Chile. https://observatorioandino.com. Accessed 04-12-2024.

Author contributions

AH developed the dataset methodology and creation in consultation with RSN and SB. RSN collected the raw observed dataset. AH prepared the data, conducted the experiments, and wrote the first draft of the manuscript. All authors were involved in discussions with regard to data development and quality, and all reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05312-1.

Correspondence and requests for materials should be addressed to A.H.

Reprints and permissions information is available at www.nature.com/reprints.

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