Data in Brief 56 (2024) 110876

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

SiCLIMA: High-resolution hydroclimate and temperature dataset for Aragón (northeast Spain)

Roberto Serrano-Notivoli*, Miguel Ángel Saz, Luis Alberto Longares, Martín de Luis

Departamento de Geografía y Ordenación del Territorio. Instituto Universitario de Investigación en Ciencias Ambientales de Aragón (IUCA), Universidad de Zaragoza, Pedro Cerbuna, 12, 50009 Zaragoza, Spain

ARTICLE INFO

Article history:
Received 26 July 2024
Revised 13 August 2024
Accepted 19 August 2024
Available online 24 August 2024

Dataset link: SiCLIMA (Sistema de Información Climática de Aragón) (Original data)

Keywords:
Precipitation
Temperature
Gridded dataset
Quality control
Reconstruction
Climate extremes

ABSTRACT

A new high-resolution climatic gridded dataset was built for Aragón (northeast Spain) using a large collection of daily precipitation and temperature observations from more than 3000 weather stations. The grid covers, at the unprecedented spatial resolution of 0.25 km^2, daily maximum and minimum temperatures and precipitation in the 1950–2020 period. The complex orography (from 70 to 3,400 m.a.s.l.) of the medium-sized region (~48,000 km^2) required a climate modelling method based on a spatially-dense weather monitoring network and local predictors. The 3-step workflow for grid creation consisted of: 1) a comprehensive quality control of raw observations, based on a spatial comparison with nearest data; 2) a climate reconstruction based on the creation of reference values, through multivariate linear regressions, for every day and location, based on the observed climate and terrain-based environmental variables; and 3) the prediction of precipitation and temperature values in a regular 500 × 500 m grid, based on the reconstructed data series. The resulting dataset improves the spatial representativity of climate and allows for a detailed analyses at landscape

* Corresponding author.
E-mail address: roberto.serrano@unizar.es (R. Serrano-Notivoli).
Social media: @sn_rober (R. Serrano-Notivoli)

https://doi.org/10.1016/j.dib.2024.110876

scale not only in climate studies but also in related disciplines such as hydrology or biogeography, amongst others.

Specifications Table

Subject	Earth and Planetary Sciences.
Specific subject area	Climatology, Meteorology, Hydrology.
Type of data	Raster, Filtered, Processed.
Data collection	The raw daily precipitation and temperature data series were obtained from the Spanish Meteorological Agency, the Ministry of Agriculture, Fisheries and Food, the Meteorological Service of Catalonia, the Government of Navarra, and the Automatic Hydrological Information Systems (SAIHs) from the basins of Ebro, Júcar, Tajo, and Duero rivers. In total, 3602 stations were included (2315 of precipitation and 1290 of temperature), covering all the days from 1950 to 01–01 to 2020-12-31. Data series were quality controlled and completely reconstructed by estimating data in missing values.
Data source location	Institution: University of Zaragoza Region: Aragón (northeast Spain) Spatial extent: 2.7ºW, 39.4ºN; 1.3ºE, 43.1ºN
Data accessibility	Repository name: Zenodo Data identification number: DOI: 10.5281/zenodo.12822293 Direct URL to data: https://zenodo.org/doi/10.5281/zenodo.12822293
Related research article	*none*

1. Value of the Data

- Daily precipitation and temperature data allow for assessing extreme events, which are becoming more frequent and intense in Aragón, especially high temperatures and precipitation extremes.
- The gridded climatic data can be used in any research requiring climatic information in the study area. Some of the most demanding topics are, for example, species distribution modeling, climate-related health diseases modeling, past climate variability, or natural hazards, amongst others.
- The dataset includes information from different sources and data at an unprecedented spatial resolution (0.25 km^2), which substantially improves the representativity of the data in comparison with previous datasets covering the area.
- The information about the uncertainty of climate estimates allows to assess their reliability and their spatial consistence.

2. Background

Climate is a territorial continuum, yet the climatic information is gathered in a small portion of land, at the meteorological stations. Recorded data can become superficial information after adequate quality control and geostatistical techniques application, but its reliability depend on the decisions made in this process [1].

Aragón (NE Spain) is in a Mediterranean climatic context, however, its great inner variability produces a wide range of environments. From the semi-arid steppes of central lowlands ($<$ 100 m.a.s.l. –meters above sea level–) to the cold humid summits of Pyrenees ($>$3000 m.a.s.l.) at north, and the Mediterranean mountains ($<$ 2000 m.a.s.l.) at the southeastern limit, there is a rich gradient of temperatures and precipitations that remain hide to large-scale climate anal-

yses. The complex climatic mosaic is conditioned by the nature of the atmospheric circulation, geographical factors such as relief or continentality, and also by the uses and occupation of the land.

This dataset was created with a widely used climatic reconstruction method [2,3] through the reddPrec R package [4] that covers three features: quality control, data series reconstruction, and gridding. To do this, local multivariate linear regressions were applied to the observations, using environmental information as covariates.

3. Data Description

The SiCLIMA (Sistema de Información del Clima en Aragón) dataset comprises daily gridded data of 6 variables: precipitation, maximum and minimum temperature, and their corresponding uncertainty values. They are stored in geospatial format at a 500 × 500 m spatial resolution, in the projected coordinate system ETRS89 / UTM zone 30 N (EPSG:25830). Data on each of the six variables are structured in 6 *NetCDF* files, one per variable (precipitation, maximum and minimum temperature, and their uncertainty values). Each file contains 25,933 layers corresponding to all days of the period. This structure allows for manageable partial loading of data series instead of loading the whole dataset, leading to large amounts of RAM memory consumption. *NetCDF* files can be accessed from any software or system with Geographic Information Systems capabilities (e.g.: ArcGIS, QGIS, R, CDO, Geoserver, etc.).

The layer structure of the dataset allows for map algebra calculations, which facilitates the analysis of spatial patterns related to climate. Layers (one per day) can be temporally aggregated in different ways, as if they were data series, resulting in a limitless body of climatic aggregations and indices to represent different features of climate. As an exemplar, Fig. 1a shows the annual average maximum precipitation in 24 h, calculated as the mean value of all the annual values of maximum daily precipitation. Calculations were made for all pixels, treated as single data series, and then colored as the resulting map. Fig. 1b shows the annual average number of

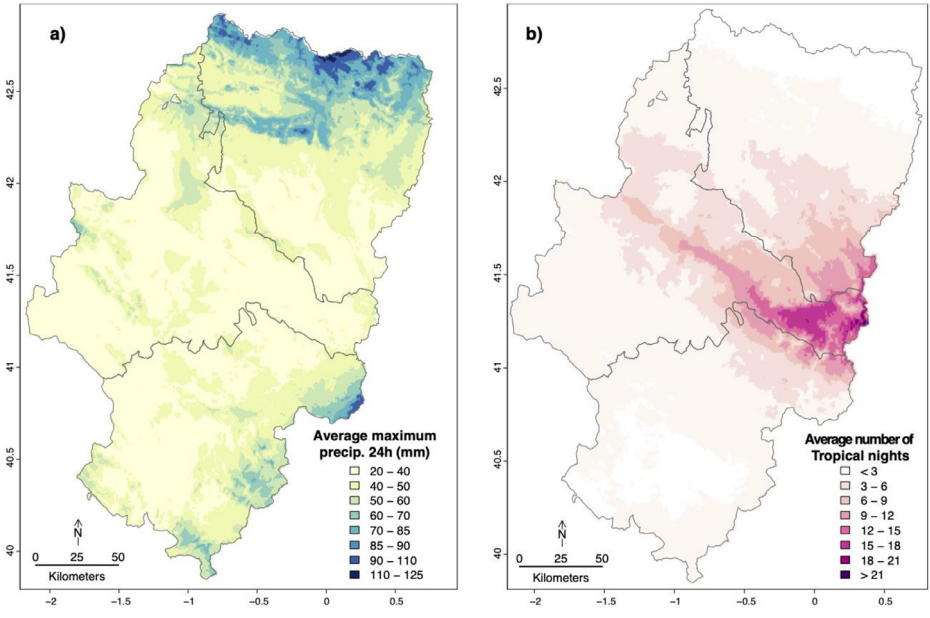

Fig. 1. a) Annual average maximum precipitation in 24 h (1950–2020). b) Annual average number of tropical nights (TMIN>20 ºC) (1950–2020).

tropical nights, calculated as the mean value of all the annual number of days with minimum temperatures higher than 20 °C.

4. Experimental Design, Materials and Methods

4.1. Data

The daily precipitation and temperature data recorded at the 3602 meteorological stations were collected, under request, within the framework of the "Atlas Climático de Aragón" project, from Spanish Meteorological Agency (AEMET) (2807 stations), the Ministry of Agriculture, Fisheries and Food (221), the Meteorological Service of Catalonia (SMC) (66), the Government of Navarra (26), and the SAIHs from the basins of Ebro (375), Júcar (91), Tajo (13), and Duero (3) basins.

The environmental variables used to model the precipitation and temperature estimates were (Fig. 2): (1) elevation, sourced from EU-DEM at 25 m spatial resolution [5] and resampled to 500 m; (2) latitude and (3) longitude of each pixel covering the spatial domain; (4) A cost-distance model computed by summing the distances of each cell to the nearest coastline and multiplied by a friction surface (the elevation in this case) which express the cost per unit distance. The values in the map are the logarithm of the cost-distance values for representation purposes, and (5) Ruggedness of terrain, expressed as the mean of the absolute differences between the elevation of a cell and its 8 surrounding cells.

The gridded dataset was built through 3 different stages: quality control, series reconstruction and gridding. The core of the method is the creation of reference values (RVs), which are estimates of precipitation and temperature based on the nearest stations and their

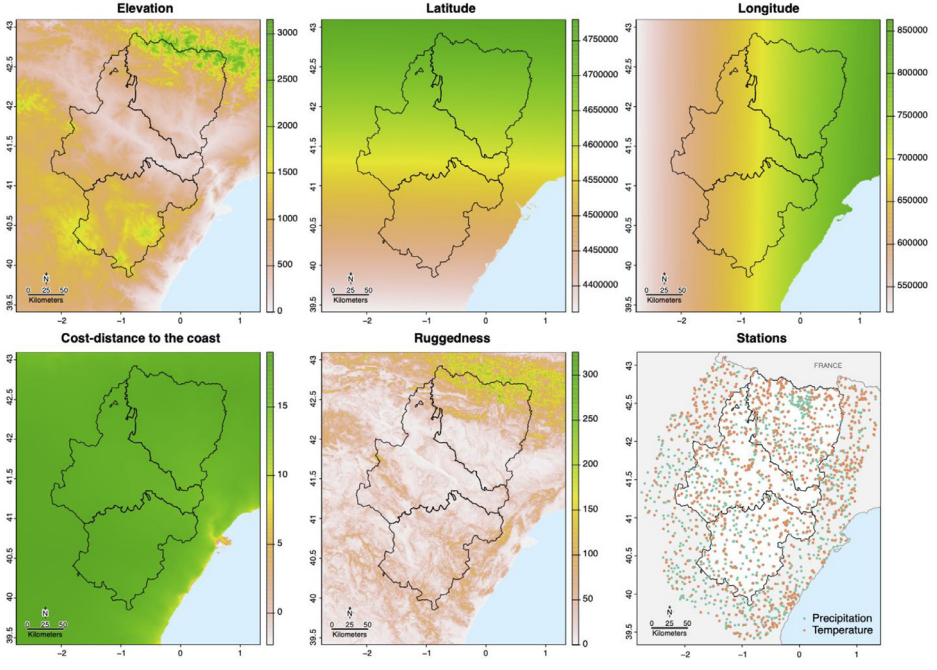

Fig. 2. Environmental variables used as covariates in precipitation and temperature modelling (elevation (meters); latitude and longitude (degrees), cost-distance to the coast (unitless), ruggedeness (meters)) and location of the meteorological observatories with the original data.

corresponding values of the environmental variables. The RVs are estimated through the combined use of generalized linear mixed models (GLMMs) and generalized linear models (GLMs) for temperature and GLMs for precipitation. This method has been described in detail and applied before to create climatic gridded datasets in different parts of the world [2,3,6,7].

4.2. Quality control

The quality control (QC) process for temperature checks several parameters to flag and remove abnormal data: internal incoherences (TMAX<TMIN), months with less than 10% of observations, extreme values, and series of repeated values. After these first inspections, RVs are computed to obtain a paired estimate to each observation. Then, correlation and difference analyses [2,3] are applied to both observed and estimated data series to remove those observations with very low correlations and very high differences with their corresponding estimates. At this point, the RVs represent the spatial coherence of the observations since they are built from neighbouring observations and, because of that, removals based on spatial and temporal anomalies result in a consistent dataset of observations, ready to be used for reconstruction.

Precipitation, on the other side, follow a different approach. The QC consists on the application of five criteria that compare every original observation with its surrounding observations in the same day, or with the RV created from them. In this work, the 15 nearest observations (NNS) were considered. The five criteria check: 1) suspect rain values, being removed the observation when it is higher than zero and all the NNS are zero; 2) suspect zeros, which are the inverse situation than (1); 3) suspect outliers are flagged when the magnitude of the observed value is 10 times higher or lower than the RV; 4) suspect dry days are observations of value zero, a probability of wet day higher than 0.99 and an RV higher than 5 mm; and 5) suspect wet days, which flags observations higher than 5 mm with a precipitation occurrence probability lower than 0.01 and an RV lower than 0.1. This process is iterative until no suspect values are detected.

Temperature quality control followed the same temporal evolution of flagged observations for both maximum and minimum values (Fig. 3). Despite the varied temporal evolution of the detected abnormal values, an annual average of 2.81 % and 2.86 % of data were respectively removed from the original observations. These values were lower until the beginning of 1990s decade, when a slight increase occurred coinciding with the generalized inclusion of automatic weather stations in the observation networks. However, the precipitation quality control followed a different evolution, with a decreasing number of flagged observations. This situation was probably due to an increase in: i) the precision of measurements and ii) the spatial density of observations, allowing for a more robust comparison between observations and RVs.

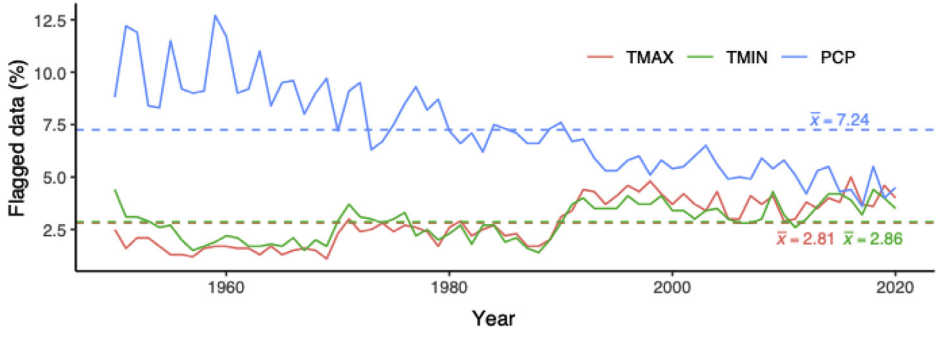

Fig. 3. Temporal evolution of flagged daily observations of precipitation (solid blue line), maximum (solid red line) and minimum (solid green line) temperatures. Average values are indicated and highlighted with dashed lines.

4.3. Climate reconstruction

The filtered series after the QC process are used to calculate new RVs for all days and locations of the observatories. The RVs corresponding to the days with no observations serve to fill the missing values, creating serially–complete data series. Those coinciding with the observations are used to evaluate the performance of the modelling by a pairwise comparison in which a leave-one-out process cross validation (LOO–CV) helps to easily compare the observations with their corresponding estimates.

The high correlation between observations and predictions is generalized, considering the average by stations (Fig. 4), for maximum and minimum mean daily temperatures (Pearson $>$ 0.97 and 0.95, respectively) and mean daily precipitation (0.96), as well as the extremes, considering the 95th percentile of maximum temperatures and 5th percentile for minimum (0.94 for both variables) and the 95th percentile daily precipitation (0.95).

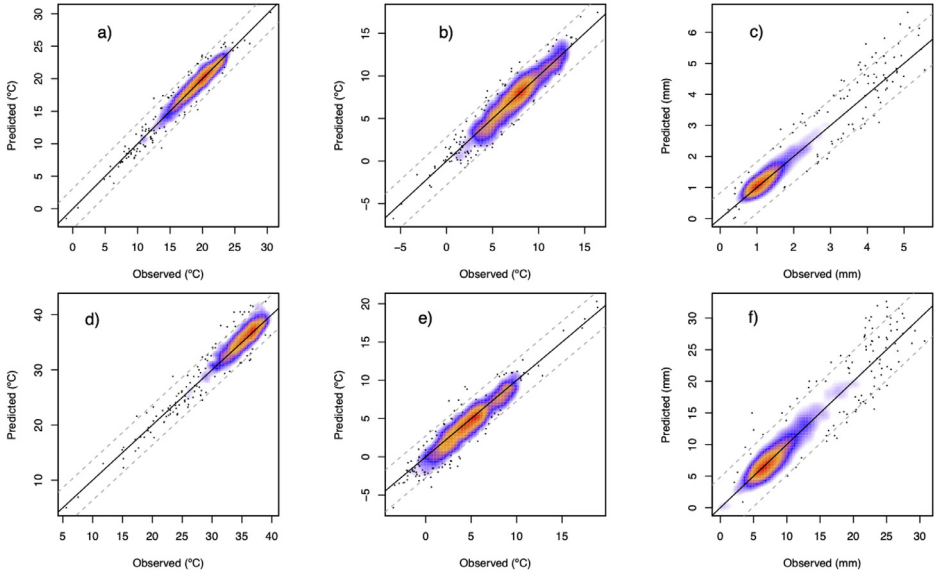

Fig. 4. Comparison between observations (X axis) and estimates (Y axis), by stations, of the mean maximum temperatures (a) and their 95th percentiles (d), the mean minimum temperatures (b) and their 5th percentiles (e), and the mean daily precipitation (c) and their 95th percentiles (f). Dashed lines represent ±1 standard deviation of the data.

In a final stage, the reconstructed series are used to build the gridded dataset. For each grid box centroid with the information extracted from the environmental variables acting as covariates, RVs are calculated for each day. In this case, we used a 500×500 m spatial resolution grid that created 19,4662 grid points in which estimates of precipitation, maximum and minimum temperature were calculated for all the 25,933 days from 1950 to 2020. These resulted in more than 5000 millions of data estimates, and the same number for associated uncertainty values.

Limitations

Inherent to all gridded data, varied differences with observations can arise. It must be considered that daily values in the raster layers are not observations but estimates. Therefore, a direct comparison of a pixel with its overlaying observatory is not fair since the pixel is built with neighboring observations.

Mountain areas are underrepresented in terms of density of observations, which is reflected in a greater uncertainty of the estimates. These areas must be taken with caution since they may

not accurately reflect what is happening in the territory. Furthermore, the orographic complexity means that the mixture of observations to build the model can be insufficient to produce a reliable estimate.

Lastly, both the reconstruction and the grid maintain the temporal structure of the series of original observations, meaning that if there were temporal inconsistencies in the raw data series, such as inhomogeneities, these will remain in the resulting data. By mixing information from several stations, these inhomogeneities can be transferred to an uncertain degree to the pixels of the grid.

Ethics Statement

This dataset does not include any human subjects, animal experiment, or social media platforms.

Data Availability

SiCLIMA (Sistema de Información Climática de Aragón) (Original data) (Zenodo).

CRediT Author Statement

Roberto Serrano-Notivoli: Conceptualization, Methodology, Formal analysis, Data curation, Validation, Software, Visualization, Writing – original draft; **Miguel Ángel Saz:** Validation, Resources, Project administration, Writing – review & editing; **Luis Alberto Longares:** Validation, Methodology, Resources, Writing – review & editing; **Martín de Luis:** Conceptualization, Methodology, Formal analysis, Data curation, Writing – review & editing.

Acknowledgments

We acknowledge the Government of Aragón for funding this research through the "Atlas Climático de Aragón" project, and AEMET for providing the raw climatic data to build the gridded dataset. R.S-N is funded by grant RYC2021-034330-I funded by MCIN/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/PRTR". All the authors are supported by the Government of Aragón, through the "Programme of research groups" (group S74_23R, "Clima, Agua, Cambio Global y Sistemas Naturales").

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Serrano-Notivoli, E. Tejedor, From rain to data: a review of the creation of monthly and daily station-based gridded precipitation datasets, WIREs Water 8 (6) (2021) e1555, doi:10.1002/wat2.1555.
[2] R. Serrano-Notivoli, S. Begueriá, M.A. Saz, L.A. Longares, M. de Luis, SPREAD: a high-resolution daily gridded precipitation dataset for Spain – an extreme events frequency and intensity overview, Earth Syst. Sci. Data 9 (2017) 721–738, doi:10.5194/essd-9-721-2017.
[3] R. Serrano-Notivoli, S. Begueriá, M. de Luis, STEAD: a high-resolution daily gridded temperature dataset for Spain, Earth Syst. Sci. Data 11 (2019) 1171–1188, doi:10.5194/essd-11-1171-2019.
[4] R. Serrano-Notivoli, M. de Luis, S. Begueriá, An R package for daily precipitation climate series reconstruction, Environ. Modell. Softw. 89 (2017) 190–195, doi:10.1016/j.envsoft.2016.11.005.

[5] [dataset] European Environment Agency, Digital Elevation Model in Europe (EU-DEM). Accessed 2024-05-23, http://data.europa.eu/88u/dataset/eu-dem
[6] A. Centella-Artola, A. Bezanilla-Morlot, R. Serrano-Notivoli, R. Vázquez-Montenegro, M. Sierra-Lorenzo, D. Chang-Dominguez, A new long term gridded daily precipitation dataset at high-resolution for Cuba (CubaPrec1), Data Br. 48 (2023) 109294, doi:10.1016/j.dib.2023.109294.
[7] N. Škrk, R. Serrano-Notivoli, K. Cufar, M. Merela, Z. Crepinšek, L. Kajfež Bogataj, M.de Luis, SLOCLIM: a high-resolution daily gridded precipitation and temperature dataset for Slovenia, Earth Syst. Sci. Data 13 (2021) 3577–3592, doi:10.5194/essd-13-3577-2021.