scientific data

OPEN

DATA DESCRIPTOR

BULL Database – Spanish Basin attributes for Unravelling Learning in Large-sample hydrology

Javier Senent-Aparicio[1]✉, Gerardo Castellanos-Osorio[1], Francisco Segura-Méndez[1], Adrián López-Ballesteros[1], Patricia Jimeno-Sáez[1] & Julio Pérez-Sánchez[2]

We present a novel basin dataset for large-sample hydrological studies in Spain. BULL comprises data for 484 basins, combining hydrometeorological time series with several attributes related to geology, soil, topography, land cover, anthropogenic influence and hydroclimatology. Thus, we followed recommendations in the CARAVAN initiative for generating a truly open global hydrological dataset to collect these attributes. Several climatological data sources were used, and their data were validated by hydrological modelling. One of the main novelties of BULL compared to other national-scale datasets is the analysis of the hydrological alteration of the basins included in this dataset. This aspect is critical in countries such as Spain, which are characterised by rivers suffering from the highest levels of anthropisation. The BULL dataset is freely available at https://zenodo.org/records/10605646.

Background & Summary

Large-sample hydrology (LSH) yields reliable insights into hydrological processes and models by leveraging comprehensive basin datasets. Recent review studies[1] have underscored the fundamental role of these datasets in a wide range of hydrological investigations, including catchment classification[2], assessments of terrestrial water storage and extreme events[3], evaluations of hydrological models[4], benchmarking[5], parameter estimation[6], regionalisation through machine learning algorithms[7], analyses of human impacts on hydrology[8], streamflow forecasting[9], exploration of climate change impacts[10], and assessments of data and model uncertainties[11].

In recent years, several LSH datasets have been developed at a national scale. For instance, Addor *et al.*[12] continued the work published by Newman *et al.*[13] to create a dataset that included streamflow measurements, meteorological forcing data, and basin attributes for 671 watersheds in the contiguous United States.Other scientists extended this initiative in subsequentyears to develop similar databases in other countries, such as Chile[14], the Great Britain[15], Brazil[16,17], Australia[18], Central Europe[19], China[20], Iceland[21], and Switzerland[22]. The vast number of hydrological databases published in recent years motivated the recently published work of Kratzert *et al.*[23], which combined and standardised several LSH datasets into a global community dataset called CARAVAN. Recently, several datasets were published following the recommendations and codes provided by this initiative in Israel[24], Germany[25], Denmark[26], and Spain[27].

Despite being one of the driest countries in the European Union, Spain has the most irrigated croplands, accounting for 75% of total water resources consumption[28]. Paradoxically, the primary areas with irrigation are concentrated in the country's most arid regions. This considerable imbalance between water resources and demands has prompted substantial investments in hydraulic infrastructure, leading to varying degrees of water resource exploitation across basins. Spain boasts the world's largest reservoir capacity relative to its surface area; over 1,200 large dams (predominantly constructed in the mid-20th century) play a crucial role in the nation's socio-economic development[29]. The extensive dam network has placed Spanish rivers among the most regulated globally, as evident from the GlObal georeferenced Database of Dams (GOODD)[30] analysis. Notably, only 25% of the surface area of Peninsular Spain does not drain into one of the 823 largest dams recorded in the GOODD database, highlighting the pervasive influence of dams on Spanish rivers, which has resulted in the difficulty of finding flow gauging stations in a natural regime. Therefore, the need to characterise LSH datasets with the degree of hydrological anthropisation, as highlighted by Addor *et al.*[1], is especially relevant in a country like

[1]Department of Civil Engineering, Catholic University of San Antonio, Campus de los Jerónimos s/n, 30107, Murcia, Spain. [2]Department of Civil Engineering, Universidad de Las Palmas de Gran Canaria, Campus de Tafira, 35017, Las Palmas de Gran Canaria, Spain. ✉e-mail: jsenent@ucam.edu

Spain where a large number of dams makes Spanish rivers among the most regulated in the world, making it a great challenge to find those whose regime has not been altered.

However, recently published datasets following the recommendations of the CARAVAN initiative for Spain[27] have not provided a detailed analysis of this issue. For instance, the degree of regulation is calculated in the CARAVAN initiative based on the Global Reservoir and Dam (GRanD) database developed by Lehner *et al.*[31]. Considering Spain, CAMELS-ES includes the degree of alteration and, the information provided by the Spanish Ministry of Environment on whether there are dams upstream of the gauging station. However, these criteria are not enough to identify hydrological anthropisation, since groundwater exploitation upstream of the monitoring station in some gauging stations is so important that the hydrological regime of the river is clearly altered despite the absence of dams. Hence, our work compared the observed flows of all study basins with the flows simulated by the national-scale hydrological model (Integrated System for Rainfall-Runoff Modelling, SIMPA)[32] to identify basins with minimally altered hydrological regimes.

Precipitation is a pivotal factor in hydrological modelling since it significantly influences the accuracy[33].This relationship exhibits nonlinearity despite its undeniable connection to various processes within the hydrological cycle, including the amount, intensity, and distribution. Nonetheless, precise precipitation assessment remains paramount for hydrological modelling, as it furnishes meteorological data crucial for hydrological studies[34]. Thus, ensuring reliable and accurate precipitation data at adequate spatial and temporal resolutions is imperative for scrutinising climate trends and effective water resource management[35]. Estimating precipitation across space poses challenges due to its spatio-temporal diversity and the complexity of involved physical processes[36]. However, recent advancements have seen notable strides in developing global reanalysis systems that combine observations of diverse variables with numerical weather predictions through data assimilation techniques[37]. ERA5-Land[38] is one such reanalysis product the CARAVAN initiative recommends by the CARAVAN initiative for extracting meteorological forcing data. However, recent studies[39] evaluating this product for Peninsular Spain have highlighted its poor detection capacity on the Mediterranean coast, especially during the summer. Other recent initiatives[27] include the EMO-1 meteorological dataset[40] developed for Europe at a higher resolution than ERA5-Land within the CAMELS-ES dataset. However, its performance has not yet been evaluated for Peninsular Spain. The BULL database includes the weather data for ERA5-Land and EMO-1 for all catchments, as well as the official grid of the Spanish State Meteorological Agency, whose performance has been shown as highly suitable for hydrological modelling[34].

The main objective of this study was to present the BULL database, which was developed for application in large-scale hydrological studies following the CARAVAN initiative's procedures. In addition, the 484 catchments were analysed to determine which had unaltered hydrological regimes. Moreover, data from three meteorological reanalysis products were analysed, considering precipitation and temperature estimation as well as their influence on the hydrological simulation using the Témez[41] hydrological model. The BULL database expands on previous efforts by other authors, including hydrometeorological daily time series from 1951 to 2021. This database provides opportunities for identifying long-term trends for climate research over decades as well as conducting short-term local water cycle analyses in specific basins. The BULL database can also serve as a benchmark dataset for improved modelling and analysis tools in Peninsular Spain and holds potential for further extensions, such as refining the temporal resolution from daily to hourly, adding water quality and chemical data, and incorporating data from over 400 reservoirs available in the Official Gauging Station Network of Spain (ROEA).

Methods

The conceptual framework designed to build the BULL dataset is shown in a flowchart in Fig. 1. The first part of this study selected the basins to include in the BULL database. All available basins in the ROEA) were considered. Additional information from flow gauge stations provided by the Catalan Water Agency (ACA) and the Andalusian Environmental Information Network (REDIAM) was also obtained. These data were subjected to a series of selection criteria to obtain 484 basins. Secondly, the code provided by the CARAVAN initiative was used to extract meteorological forcing data from ERA5, and basin attributes to extend CARAVAN with data from 484 basins in Spain. Subsequently, an analysis was conducted to determine how many of these 484 basins were unaltered by comparing observational data with national-scale hydrological model SIMPA data. SIMPA is a distributed hydrological model that Spanish authorities use to evaluate water resources in natural regimes. It simulates the natural water balance and provides information about the main hydrological variables at a monthly time step and with a spatial resolution of 500 m. Finally, meteorological data from AEMET and EMO-1 products were extracted for all basins, analysing their performance, and simulating hydrological processes through the Témez hydrological modelto highlight the influence of the meteorological data in hydrological simulation.

Basin selection and delineation. Selecting basins for the BULL database began with data from the flow measurements available at the 1,634 stations registered in the ROEA, ACA, and REDIAM. Next, initial filtering excluded all flow measurement stations that only had data before 1951 since the available meteorological data from the Spanish Meteorological Agency began that year. As a result, the number of stations was reduced to 1,431. Then, the CARAVAN initiative's recommended filter regarding the size of the basins was applied, discarding those smaller than 100 km^2 or larger than 2000 km^2, which reduced the number of available stations to 848. Subsequently, the percentage of gaps in the data from these stations was analysed, eliminating those with a percentage of gaps greater than 10%, reducing the number of available stations to 764. Finally, the minimum number of years with complete data available at these stations was studied, discarding all stations with less than 20 complete years of data, resulting in the total number of basins analysed in the BULL database, which was 484 (Fig. 2a). Two hundred twenty-nine of these basins have forest as the primary land use. Regarding size, 59% (n = 286) have

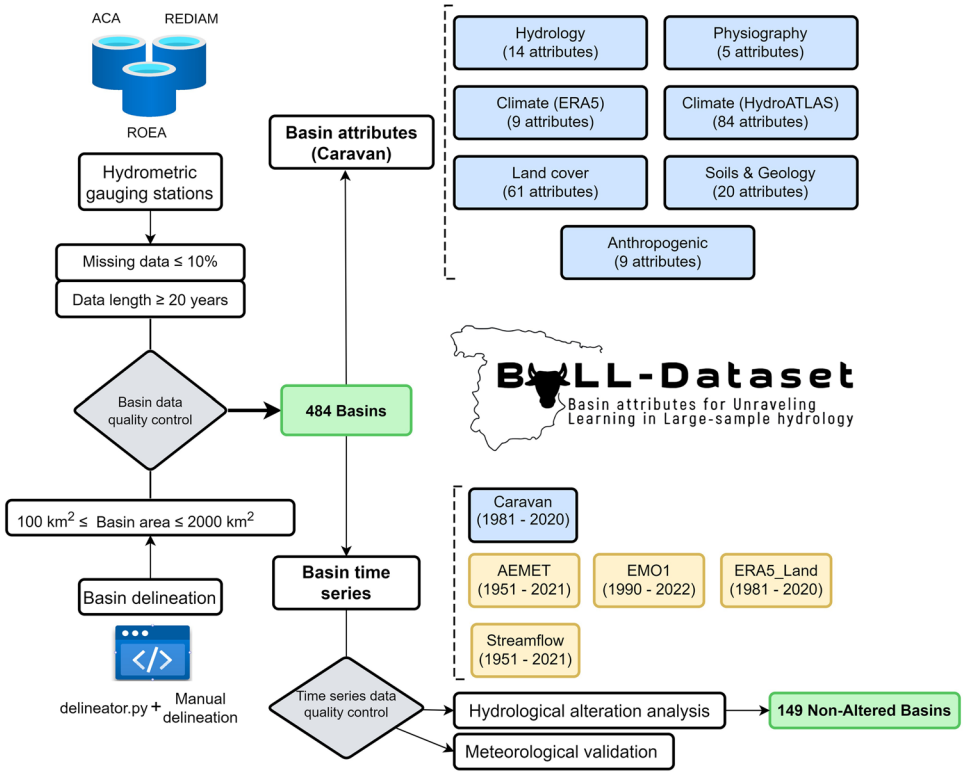

Fig. 1 Schematic figure of the approach to generating the BULL dataset for Peninsular Spain.

an area between 100 and 500 km^2, while only 6% of the basins (n = 28) have an area greater than 1500 km^2. It is important to highlight that 287 basins, representing 60% of the total, have complete series without gaps for those 20 years. Additionally, approximately 19% of the basins have complete time series covering at least 60 years of observed daily runoff, while approximately 26% have 90% data coverage for that period. Figure 2b shows that the stations in the BULL dataset tend to have long-term gauging records, with the shortest record being 20 years, and more than half of the records (68% of the stations) being at least 40 years long. The distribution of stations according to the time-series length is uniform across the entire area, except for those with at least a span of 60 years, which are more abundant in the central-northern region.

Spain currently lacks a dedicated spatial database for extracting the geometrical attributes of gauged basins. Consequently, these basins' dimensions were derived from previously acquired gauge data, designating them as the drainage focal points for Spanish gauged basins. The accuracy and spatial resolution of the digital elevation models (DEMs) were fundamental technical aspects during this phase[17]. Therefore, the choice was made to utilise the Delineator.py code[42] to achieve this task, which employs hybrid methodologies, integrating vector- and raster-based approaches alongside data from MERIT-Hydro to delineate basins. These DEMs accurately depict terrain elevations at a 3-second resolution (approximately 90 metres at the equator), encompassing land areas between 90°N and 60°S for basin delineation purposes. Subsequently, geoprocessing techniques were employed to calculate the surface area of each delineated basin. The quality control process began by cross-referencing the information provided by the ROEA, ACA, and REDIAM, as well as the initial watershed delineation results obtained using delineator.py. Basins with significant differences in surface area were addressed by verifying the location of gauging points and adjusting the basin boundaries accordingly. Once the initial errors were rectified, the QGIS 3.23.3 command "fix geometries" was applied to ensure the integrity of the final basin geometries. Subsequently, the CARAVAN initiative scripts were applied to obtain each basin's attributes.

Methodological approach for identifying non-altered basins. As mentioned, Spain's rivers exhibit a higher degree of anthropisation than other countries, mainly due to the large number of existing dams and the need to supply water to the entire Mediterranean Spain[43]. Therefore, the BULL initiative sought to analyse which basins included in the database had a natural or quasi-natural regime. Hence, the monthly observed flows of the 484 basins included in the BULL database were compared with the flows simulated by the national-scale hydrological model SIMPA, which Spain uses to calculate water resources in the various river basin plans developed under the framework of the European Water Directive[44]. To identify unaltered basins in the BULL database, we established a criterion of those whose observed flows compared to those of SIMPA produce a Nash-Sutcliffe coefficient[45] (NSE) equal to or greater than 0.50. This NSE value is commonly used in hydrological modelling following Moriasi *et al.*[46]. A significant limitation of this approach was the assumption that the SIMPA model adequately reflected the natural regime flow series. The results of the model presented different sources of uncertainty due to

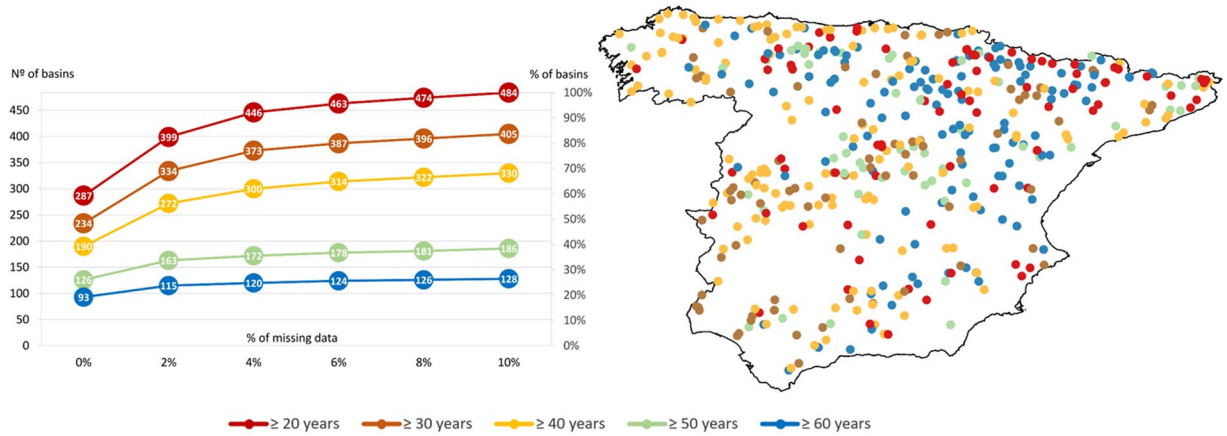

Fig. 2 Analysis of missing data and time series length in streamflow stations: (**a**) the number and percentage of stations with different percentages of missing data across various time periods and (**b**) the length of the streamflow time series for each station.

input data error, model parameters, and model structure. However, as seen in the analysis of water resources for Peninsular Spain conducted by the Centre of Hydrographic Studies of the Centre for Studies and Experimentation of Public Works (CEDEX)[47], the results were acceptable in most of the territory. Nevertheless, the efficiency decreased in the more arid areas of the Southeast with few unaltered gauging stations.

Evaluation of reanalysis datasets for hydrological modelling. The meteorological data used to produce BULL were statistically validated by comparing and evaluating the robustness and accuracy of the datasets. AEMET was used as a benchmark[34,39,48] to evaluate the performance of ERA5-Land and EMO1. The Spearman correlation coefficient (ρ), relative bias (RBIAS), root mean square error (RMSE), and Kling-Gupta efficiency (KGE) were calculated separately for each basin according to four meteorological variables: the total monthly precipitation and the monthly maximum, minimum, and mean temperatures. The equations of these statistics used for continuous analysis were described by Gomis-Cebolla *et al.*[39]. The coefficient ρ offers a robust assessment of the degree to which two variables exhibit a monotonic relationship, regardless of the data's distributional properties. The RBIAS, expressed as a percentage, denotes the systematic bias in the estimates of precipitation and temperature, while the RMSE quantitatively represents the error characteristics between the different estimates of the variables and those taken as reference. KGE consolidates multiple statistical metrics into one measure, evaluating the model's accuracy, variability reproduction, and temporal alignment with the reference data.

The Témez hydrological model was employed to validate the meteorological sources from a hydrological perspective. The Témez lumped rainfall-runoff model has been extensively employed in Spain for water resource management[34,49–51] and in several other countries[52]. More details of the Témez model can be found in Pérez-Sánchez *et al.*[41]. The calibration process of the hydrological model was conducted to adjust its parameters to achieve an accurate streamflow simulation. AEMET data served as our reference dataset for calibration. The calibration involved a loss function based on the mean squared error (MSE), which compared observed and simulated streamflows. Once the parameters were calibrated, the streamflows were simulated using other available data sources. Optimisation was conducted using the least squares method with defined search bounds to identify the best model parameters. Specifically, the implementation in Python involved defining the loss function (MSE) and performing optimisation using the SciPy library's minimise function with the L-BFGS-B algorithm[53]. To evaluate each hydrological model, the NSE, PBIAS, and RMSE observations' standard deviation ratio (RSR) statistics were employed, following the criteria established by Moriasi *et al.*[54], whose work provided comprehensive details on these widely used statistics for hydrological model evaluation. The RSR standardised the RMSE using observation standard deviation, providing an integrated error index.

Data Records

The BULL dataset presented in this work[55], encompassing 484 watersheds, can be accessed, and downloaded at https://zenodo.org/records/10605646. The dataset is organised according to the following folder structure:

- The attributes folder contains three CSV files obtained by using the CARAVAN script. One file, labelled "attributes_caravan_.csv", comprises climate indices derived from ERA5-Land. The "attributes_hydroatlas_.csv" file encompasses attributes derived from HydroATLAS. In contrast, the "attributes_other_.csv" file incorporates other data relating to additional attributes such as the latitude and longitude coordinates, name, country of each gauge station and catchment area. The initial column in all attribute files is designated as "gauge_id", featuring a unique basin identifier in the format "BULL_{id}". Here, BULL corresponds to the source dataset, while {id} represents the basin ID defined in the original source dataset.
- A README.md file with the link to the scripts used is included in the code folder.

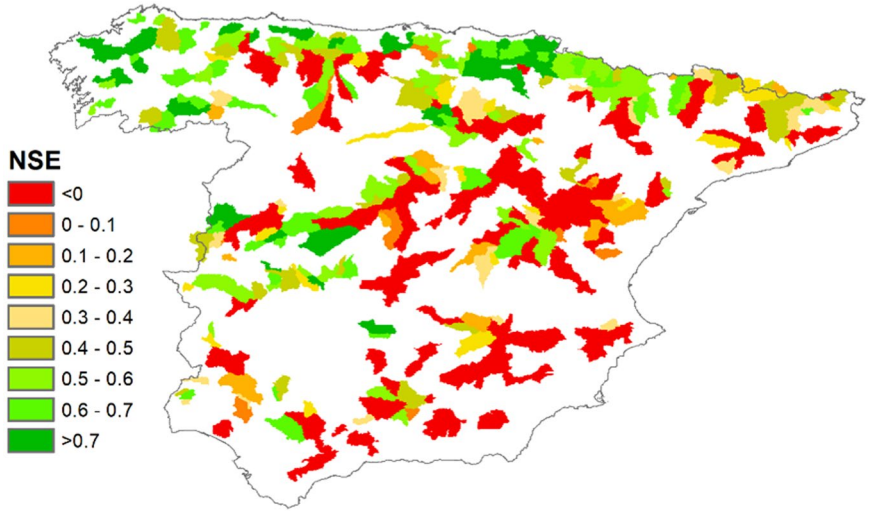

Fig. 3 NSE values comparing observed data with SIMPA national scale hydrological model values.

- The licenses folder encompasses licensing details for BULL and the incorporated dta. The README.md file in this directory provides general information about licenses and specific details about the source datasets used.
- The shapefiles folder contains a shapefile depicting the catchment boundaries of each basin included in the dataset. This shapefile was the basis for deriving the catchment attributes and ERA5-Land time series data. Each polygon in the shapefile is associated with a "gauge_id" field, containing the unique basin identifier.
- The timeseries folder comprises two subfolders, csv and netcdf, with the same structure and data—presented in CSV and netCDF formats. Within these subdirectories, five other subdirectories are labelled according to the source datasets, and individual files (CSV or netCDF) are allocated and encompass comprehensive time series data, including meteorological forcings, state variables, and streamflow. Moreover, the netCDF files incorporate metadata information such as physical units, time zones, and details regarding the data sources.
- The coordinate system used for shapefiles and netCDF files is the WGS 1984 Geographic Coordinate System (EPSG 4326).

Technical Validation

Validation of the identified unaltered basins. Analysis of the distribution of the basins estimated to be in a natural regime revealed that most with a lower degree of anthropisation were located in Northern Spain, especially in Galicia, Asturias, Cantabria, the Basque Country, and the Western Pyrenees, as shown in Fig. 3. The basins with less altered hydrological regimes were abundant in areas influenced by oceanic climates. In contrast, those with more altered regimes were found in regions characterised by continental and Mediterranean climates. Radinger *et al.*[56] detected variable flow patterns between and within geographical regions greatly influenced by climatic conditions. As expected, most of the basins analysed in the Mediterranean area exhibited high anthropisation. Leduc *et al.*[57] repeteadly emphasised the anthropisation of water resources in the Mediterranean area. This analysis allowed validation of the information provided by ROEA that assesses whether basins are altered based on the presence of upstream reservoirs. The approach undertaken in this study confirmed that some of the basins the ROEA estimated to be unaltered were instead altered, because agricultural land use in these basins has significantly altered their hydrological regime. Thus, 149 unaltered basins were identified, which are of great utility for future large-scale hydrological studies in the Iberian Peninsula.

Validation of reanalysis products using hydrological modelling. Spearman correlation analysis, shown in Fig. 4a, illustrated a spatial gradient from the west-central regions (with the highest values) to the east (with the lowest values) for ERA5-Land. The highest RMSE values (Fig. 4c) were observed for ERA5-Land and EMO1 in the northern region. Gomis-Cebolla *et al.*[39] found similar results indicating that the northern coast was one of the most critical regions in reanalysis modelling due to its performance. However, a more complez spatial pattern was observed for the rest of the statistics, which complicated straightforward spatial regionalisation. Regarding the correlation coefficient (Fig. 4a), the performance of ERA5-Land and EMO1 was similar, indicating that both were equally correlated with AEMET. However, ERA5-Land demonstrated better performance for RBIAS and RMSE compared to EMO1. In contrast, KGE showed a better performance for EMO1. Regarding monthly temperatures (i.e. the maximum, minimum, and mean), the correlation between all data sources was very high, with a median value higher than 0.98 for all temperatures. In addition, for RMSE performance, the median values were less than 1 °C.

Validation of these meteorological time series was conducted in a hydrological framework. Precipitation and temperature were fundamental inputs to hydrological models, where precipitation significantly influenced model accuracy[33]. In this study, the performance of three meteorological products was evaluated as input in a

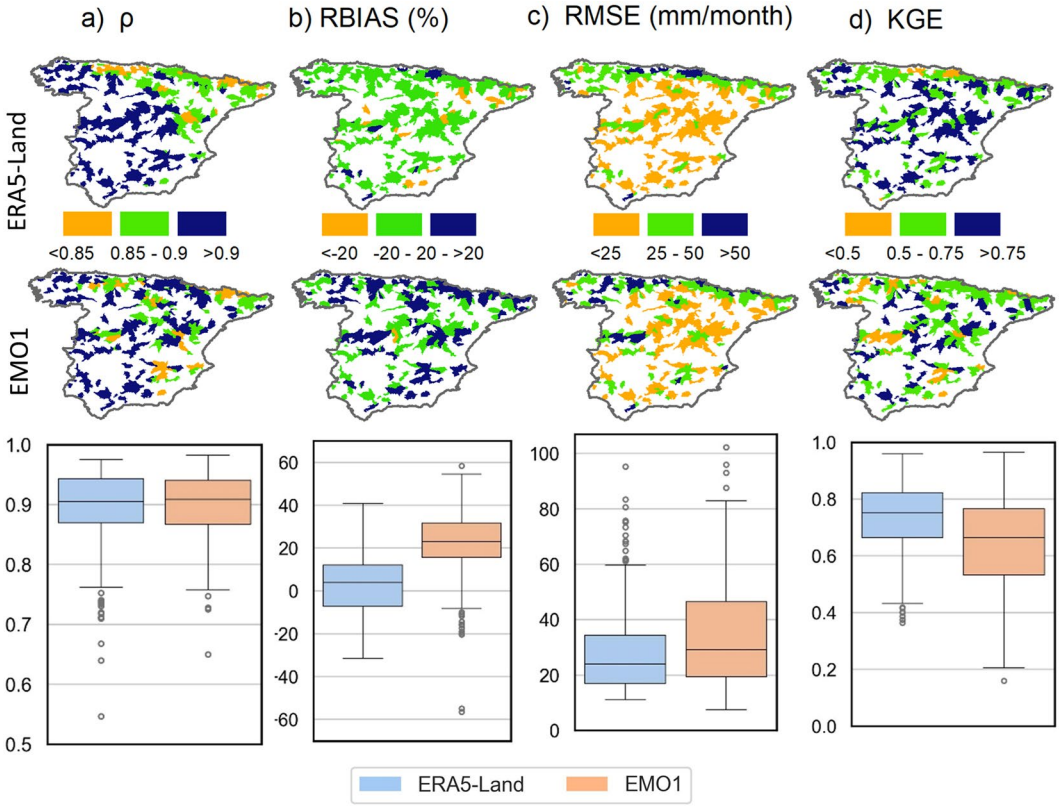

Fig. 4 Spatial distribution and boxplot of ERA5-Land and EMO1 monthly continuous statistics: Spearman correlation (**a**), RBIAS (**b**), RMSE (**c**) and KGE (**d**) statistics.

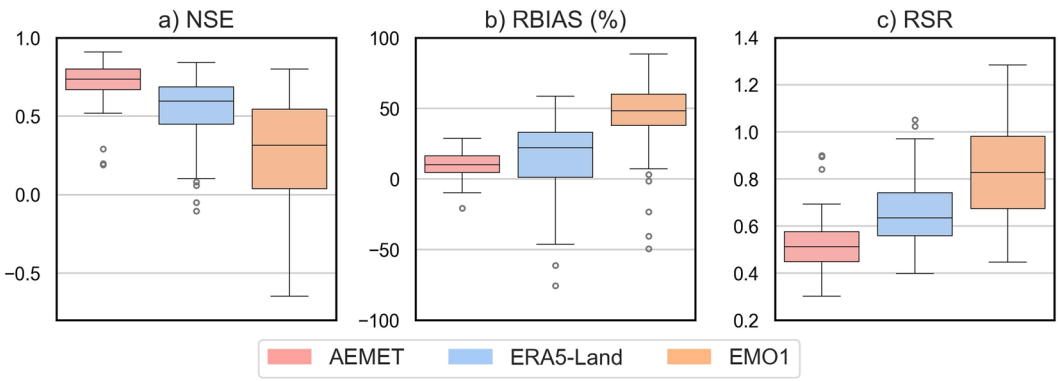

Fig. 5 Boxplot of the monthly statistics from the streamflow simulation with the hydrological model using AEMET, ERA5-Land and EMO1 meteorological data: NSE (**a**), RBIAS (**b**) and RSR (**c**) statistics.

hydrological model for streamflow simulation. For this purpose, a common time period was selected during which climatological data from the three data sources (i.e. EMO-1, ERA5-Land, and AEMET) were available (1990–2020). This new time period considered which stations were identified as unaltered. The missing monthly streamflow data was less than 10%, identifying 87 stations in which the Témez hydrological model was applied.

The model was calibrated using AEMET data, and the simulated streamflow was obtained using three precipitation and temperature products. The performance ofeach product (Fig. 5) was assessed by comparing it with the observed streamflows at the outlet of each basin. According to Moriasi *et al.*[54], AEMET demonstrated performance exceeding satisfactory levels in nearly all basins. The worst models were observed with EMO1. The variability of the results obtained with ERA5-Land and EMO1 was much higher than with AEMET, with the highest dispersion observed with EMO1 data. The poorest performance was observed with EMO1, as depicted in Fig. 5, where the median values for NSE/RBIAS/RSR were 0.6/22/0.6 for ERA5 and 0.3/48/0.8 for EMO1.

Code availability

The code developed by Kratzert *et al*. in the CARAVAN initiative has been used for the calculation of the basin attributes (available at https://zenodo.org/records/6578598). The code used for the validation of the meteorological data including the coding of the Témez hydrological model are written in Python and are available at https://zenodo.org/records/10605646.

Received: 18 March 2024; Accepted: 1 July 2024;
Published online: 06 July 2024

References

1. Addor, N. *et al*. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal* **65**, 712–725 (2020).
2. Jaffrés, J. B. D., Cuff, B., Cuff, C., Knott, M. & Rasmussen, C. Hydrological characteristics of Australia: national catchment classification and regional relationships. *Journal of Hydrology* **612**, 127969 (2022).
3. Stein, L., Clark, M. P., Knoben, W. J. M., Pianosi, F. & Woods, R. A. How Do Climate and Catchment Attributes Influence Flood Generating Processes? A Large-Sample Study for 671 Catchments Across the Contiguous USA. *Water Resources Research* **57**, e2020WR028300 (2021).
4. Mathevet, T., Gupta, H., Perrin, C., Andréassian, V. & Le Moine, N. Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. *Journal of Hydrology* **585**, 124698 (2020).
5. Towler, E. *et al*. Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States. *Hydrol. Earth Syst. Sci.* **27**, 1809–1825 (2023).
6. Liu, H., Tolson, B. A., Newman, A. J. & Wood, A. W. Leveraging ensemble meteorological forcing data to improve parameter estimation of hydrologic models. *Hydrological Processes* **35**, e14410 (2021).
7. Mehrvand, S., Boucher, M.-A., Kornelsen, K. & Amani, A. Comparing three machine learning algorithms with existing methods for natural streamflow estimation. *Hydrological Sciences Journal* **69**, 79–94 (2024).
8. Ouyang, W. *et al*. Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of Hydrology* **599**, 126455 (2021).
9. Ma, K. *et al*. Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions. *Water Resources Research* **57**, e2020WR028600 (2021).
10. Gupta, A., Carroll, R. W. H. & McKenna, S. A. Changes in streamflow statistical structure across the United States due to recent climate change. *Journal of Hydrology* **620**, 129474 (2023).
11. Ayzel, G. & Heistermann, M. The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset. *Computers & Geosciences* **149**, 104708 (2021).
12. Addor, N., Newman, A. J., Mizukami, N. & Clark, M. P. The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci*. (2017).
13. Newman, A. J. *et al*. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.* **19**, 209–223 (2015).
14. Alvarez-Garreton, C. *et al*. The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrol. Earth Syst. Sci.* **22**, 5817–5846 (2018).
15. Coxon, G. *et al*. CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth Syst. Sci. Data* **12**, 2459–2483 (2020).
16. Chagas, V. B. P. *et al*. CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth Syst. Sci. Data* **12**, 2075–2096 (2020).
17. Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T. & Troch, P. CABra: a novel large-sample dataset for Brazilian catchments. *Hydrol. Earth Syst. Sci.* **25**, 3105–3135 (2021).
18. Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C. & Peel, M. C. CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia. *Earth Syst. Sci. Data* **13**, 3847–3867 (2021).
19. Klingler, C., Schulz, K. & Herrnegger, M. LamaH-CE: LArge-SaMple DAta for Hydrology and Environmental Sciences for Central Europe. *Earth Syst. Sci. Data* **13**, 4529–4565 (2021).
20. Hao, Z. *et al*. CCAM: China Catchment Attributes and Meteorology dataset. *Earth Syst. Sci. Data* **13**, 5591–5616 (2021).
21. Helgason, H. B. & Nijssen, B. LamaH-Ice: LArge-SaMple Data for Hydrology and Environmental Sciences for Iceland. *Earth Syst. Sci. Data* **16**, 2741–2771 (2024).
22. Höge, M. *et al*. CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland. *Earth Syst. Sci. Data* **15**, 5755–5784 (2023).
23. Kratzert, F. *et al*. Caravan - A global community dataset for large-sample hydrology. *Sci Data* **10**, 61 (2023).
24. Morin, E. *Caravan extension Israel - Israel dataset for large-sample hydrology (Version_1.0)*. https://doi.org/10.5281/zenodo.7758516 (2023).
25. Mälicke, M. *CAMELS-DE/CAMELS-DE.github.io: v0.3.0*. https://doi.org/10.5281/zenodo.7611830 (2023).
26. Koch, J. *Caravan extension Denmark - Danish dataset for large-sample hydrology (v_05)*. https://doi.org/10.5281/zenodo.7962379 (2022).
27. Casado-Rodríguez, J. *CAMELS-ES: Catchment Attributes and Meteorology for Large-Sample Studies - Spain*. https://doi.org/10.5281/zenodo.8428374 (2023).
28. Berbel, J., Expósito, A., Gutiérrez-Martín, C. & Mateos, L. Effects of the Irrigation Modernization in Spain 2002–2015. *Water Resour Manage* **33**, 1835–1849 (2019).
29. Mezger, G., González del Tánago, M. & De Stefano, L. Environmental flows and the mitigation of hydrological alteration downstream from dams: The Spanish case. *Journal of Hydrology* **598**, 125732 (2021).
30. Mulligan, M., van Soesbergen, A. & Sáenz, L. GOODD, a global dataset of more than 38,000 georeferenced dams. *Sci Data* **7**, 31 (2020).
31. Lehner, B. *et al*. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecol & Environ* **9**, 494–502 (2011).
32. MITECO. Modelo SIMPA. https://www.miteco.gob.es/en/agua/temas/evaluacion-de-los-recursos-hidricos/evaluacion-recursos-hidricos-regimen-natural.html.
33. Sun, Q. *et al*. A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics* **56**, 79–107 (2018).
34. Senent-Aparicio, J., López-Ballesteros, A., Pérez-Sánchez, J., Segura-Méndez, F. & Pulido-Velazquez, D. Using Multiple Monthly Water Balance Models to Evaluate Gridded Precipitation Products over Peninsular Spain. *Remote Sensing* **10**, 922 (2018).
35. Liu, X., Yang, T., Hsu, K., Liu, C. & Sorooshian, S. Evaluating the streamflow simulation capability of PERSIANN-CDR daily rainfall products in two river basins on the Tibetan Plateau. *Hydrol. Earth Syst. Sci.* **21**, 169–181 (2017).

36. Nogueira, M. Inter-comparison of ERA-5, ERA-interim and GPCP rainfall over the last 40 years: Process-based analysis of systematic and random differences. *Journal of Hydrology* **583**, 124632 (2020).
37. Hu, Q. *et al*. Rainfall Spatial Estimations: A Review from Spatial Interpolation to Multi-Source Data Merging. *Water* **11**, 579 (2019).
38. Muñoz-Sabater, J. *et al*. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
39. Gomis-Cebolla, J., Rattayova, V., Salazar-Galán, S. & Francés, F. Evaluation of ERA5 and ERA5-Land reanalysis precipitation datasets over Spain (1951–2020). *Atmospheric Research* **284**, 106606 (2023).
40. Thiemig, V. *et al*. EMO-5: a high-resolution multi-variable gridded meteorological dataset for Europe. *Earth Syst. Sci. Data* **14**, 3249–3272 (2022).
41. Pérez-Sánchez, J., Senent-Aparicio, J. & Jimeno-Sáez, P. The application of spreadsheets for teaching hydrological modeling and climate change impacts on streamflow. *Computer Applications in Engineering Education* **30**, 1510–1525 (2022).
42. Heberger, M. *delineator.py. Zenodo* https://doi.org/10.5281/zenodo.10143149 (2023).
43. Senent-Aparicio, J., López-Ballesteros, A., Cabezas, F., Pérez-Sánchez, J. & Molina-Navarro, E. A Modelling Approach to Forecast the Effect of Climate Change on the Tagus-Segura Interbasin Water Transfer. *Water Resour Manage* **35**, 3791–3808 (2021).
44. Vicente, D. J., Rodríguez-Sinobas, L., Garrote, L. & Sánchez, R. Application of the system of environmental economic accounting for water SEEAW to the Spanish part of the Duero basin: Lessons learned. *Science of The Total Environment* **563–564**, 611–622 (2016).
45. Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models Part I - A discussion of principles. *Journal of Hydrology* **10**, 282–290 (1970).
46. Moriasi, D. N., Gitau, M. W., Pai, N. & Daggupati, P. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. *Trans. ASABE* **58**, 1763–1785 (2015).
47. CEDEX. *Evaluación de Recursos Hídricos En Régimen Natural En España (1940/41 - 2017/18).* https://www.miteco.gob.es/content/dam/miteco/es/agua/temas/evaluacion-de-los-recursos-hidricos/cedex-informeerh2019_tcm30-518171.pdf (2020).
48. Senent-Aparicio, J. *et al*. Impacts of swat weather generator statistics from high-resolution datasets on monthly streamflow simulation over Peninsular Spain. *Journal of Hydrology: Regional Studies* **35**, 100826 (2021).
49. Jimeno-Sáez, P. *et al*. A Preliminary Assessment of the "Undercatching" and the Precipitation Pattern in an Alpine Basin. *Water* **12**, 1061 (2020).
50. Pérez-Martín, M. A., Estrela, T., Andreu, J. & Ferrer, J. Modeling Water Resources and River-Aquifer Interaction in the Júcar River Basin, Spain. *Water Resour Manage* **28**, 4337–4358 (2014).
51. Jódar, J. *et al*. Groundwater discharge in high-mountain watersheds: A valuable resource for downstream semi-arid zones. The case of the Bérchules River in Sierra Nevada (Southern Spain). *Science of The Total Environment* **593–594**, 760–772 (2017).
52. Rivadeneira Vera, J. F., Zambrano Mera, Y. E. & Pérez-Martín, M. Á. Adapting water resources systems to climate change in tropical areas: Ecuadorian coast. *Science of The Total Environment* **703**, 135554 (2020).
53. Virtanen, P. *et al*. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).
54. Moriasi, D. N. *et al*. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE* **50**, 885–900 (2007).
55. Senent-Aparicio, J. *et al*. BULL Database – Spanish Basin attributes for Unravelling Learning in Large-sample hydrology. *Zenodo* https://doi.org/10.5281/zenodo.10605646 (2024).
56. Radinger, J., Alcaraz-Hernández, J. D. & García-Berthou, E. Environmental and spatial correlates of hydrologic alteration in a large Mediterranean river catchment. *Science of The Total Environment* **639**, 1138–1147 (2018).
57. Leduc, C., Pulido-Bosch, A. & Remini, B. Anthropization of groundwater resources in the Mediterranean region: processes and challenges. *Hydrogeol J* **25**, 1529–1547 (2017).

Acknowledgements

This research was supported by the project TwinTagus from the Spanish Ministry of Science and Innovation under grant PID2021-128126OA-I00. Gerardo Castellanos-Osorio was supported by the Ministry of Science, Innovation and Universities of Spain under an FPI grant (PRE2022-101437).

Author contributions

J.S.-A. developed the conceptualization of the study, drafted the first version of the manuscript, and was the lead author. G.C.-O., P.J.-S., A.L.-B., F.S.-M. and J.P.-S. handled the data extraction, processing and technical validation of the BULL dataset. All co-authors contributed to organizing and editing the drafts and the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.S.-A.