JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 32 (2026), pp. 1–35 doi:10.5311/JOSIS.2026.32.449

RESEARCH ARTICLE

Evaluation of gridded precipitation and temperature datasets in Spain. A proposal for improving their accuracy using random forest multi-model ensembles

Francisco Gomariz-Castillo[1], Francisco Alonso-Sarría[1], Carmen Valdivieso-Ros[1], Francisco Pellicer-Martínez[2], Gabriel Molina-Pérez[3], and José Molina-Ruíz[4]

[1]Departamento de Geografía / Instituto Universitario del Agua y del Medio Ambiente, Universidad de Murcia, Spain
[2]Escuela Politécnica Superior, Universidad Católica de Murcia (UCAM), Spain
[3]Departamento de Geografía / Instituto Universitario del Agua y del Medio Ambiente, Universidad de Murcia, Spain
[4]Departamento de Geografía, Universidad de Murcia, Spain

Received: March 15, 2025; returned: May 21, 2025; revised: November 2, 2025; accepted: January 13, 2026.

Abstract: The generation of uniform, gridded data from spatially discontinuous station values and the assessment of their accuracy are essential for water resources assessment and management, and for climate change studies, especially in semi-arid environments. Spatial and temporal grids have been generated in recent years as a basis for several studies. This work has two objectives: a) to evaluate the accuracy of the grids available for Spain by comparing their monthly values with stations not used in their estimation or prediction, and b) to verify the improvement in accuracy using Multi-Model Ensembles based on machine learning. A dual ensemble approach is presented: (i) multiple individual Random Forest (RF) ensembles per weather station, using only the information from the station, and (ii) spatially distributed grid prediction using a single ensemble model that incorporates all the information from the nearest stations and their distances using Random Forest Spatial Interpolation -RFSI-). Both models were used to generate monthly data grids of maximum, minimum and mean temperature, and total precipitation, with high spatial resolution (5

km). Seven datasets: Iberia01, STEAD, AEMET, SIMPA, EOBSv27 and STEAD, were used as predictors. Accuracy was estimated using the root mean square error, the percentage bias and the Nash-Sutcliffe efficiency index obtained using block cross-validation buffering (LOOBUF-CV), robust to spatial autocorrelation. The significance of the differences was assessed using ANOVA with heteroscedasticity correction in the residuals. Preliminary results indicate that multi-model ensembles using RF outperform individual grids. Among other reasons, ensembles aggregate the different representations of meteorological processes included in each grid and reduce the uncertainty associated with each individual grids.

Keywords: Climatic datasets, Random forest, Random forest spatial interpolation, Multi-model ensembles, ANOVA with correction errors, Iberian Peninsula

1 Introduction

Climate variables are highly relevant in agricultural engineering and environmental or hydrological studies aimed at planning and management of water resources [17]; as well as in the analysis of climate risks or impacts derived from climate change [5]. The greater availability of data and the increase in computing capacity have made it possible to generate spatio-temporal data grids that provide easy access to already refined estimates [56], including the spatial patterns of variables [19]. This type of product allows access to data that are (i) already cleaned and spatially distributed, (ii) obtained from a large number of observations, (iii) generated by specialists using complex forecasting algorithms, and (iv) with a wide temporal range and different time steps. In Spain, nine projects have produced grids for climate variables. This line of research is being followed in several studies around the world. For example, eight grids have been generated in Spain (Table 1 shows which are used in this study), five of them for Spain or the Iberian Peninsula, another for Europe and two for the whole world. [40] produce grids with a resolution of 1 km in China, while [32] provide an exhaustive review of 29 gridded climate products and their use in hydrological analyses. With the proliferation of such products, a research line has emerged with the objective of comparing them on a local or regional scale; For example, [52] compared global precipitation grids and their regionalized versions in an Ethiopian basin, and [47] compared temperature and precipitation grids in the Chilean Andes. In the same vein, [32] stress the need to evaluate the different sources of data available in any area of interest, concluding that no single source of data is better than the others. Their study includes products generated using the most common strategies:

- Ground-based datasets (20 products under their study), especially interpolation methods using weather station networks. The advantage is starting from data with well-defined biases and uncertainties inherited from the instrumentation.
- Grids generated by remote sensing techniques (including 20 studies), derived from various sensors aboard satellites. Their advantage over the above is that they provide spatially homogeneous coverage and temporally continuous records, but are limited to temporal coverage at the start of data collection.
- Reanalysis-derived products (including 23 products), synthesized from process-based climate models, usually used in conjunction with the two previous types to obtain homogeneous grids in time and space. They are synthesized from process-

> based climate models through complex interactions between a priori predictions from a physically based, dynamical process model and observational data. The data and observations are incorporated into a physically based, dynamical process model using multiple models, resulting in a large number of climate variables with latencies of hours to months.

[16] distinguish two types of errors in these grids: (a) uncertainties associated with the initial and boundary conditions, and (b) errors in the model itself. In this sense, Multi-Model Ensembles (MME) make it possible to reduce the errors of individual models by integrating different representations of the physical processes of the models from which they are derived, reducing their implicit uncertainty by combining their results [23,39,46].

In climatology, MMEs are often used to obtain series from models associated with climate change scenarios, so that predictions for future periods can be made from historical data. Traditionally, simple methods based on simple averages have been used for this purpose. [55] uses different grids of climate data as input to the GR4J hydrological model to assess the impact on water resources; its main advantage is that it is very simple to implement and can outperform at least half of the ensemble members in terms of mean square error [28,45]. One of its main problems is that it is unable to correct for generalized biases, requiring correction by local observations to be used at regional or local scales. Various proposals have been made to overcome this problem, the most important of which is the use of weights to average the different variables, such as weighted averages or Bayesian Model Averaging (BMA) [35]. Other studies use MMEs from multiple linear regression-based methods, or that of [29], which is based on an MME framework using L-moments to improve consistency in the assessment of changes in precipitation extremes under different climate change scenarios. [38] propose ensembles of climate projections for Europe based on a calibration with observational data using Gaussian regression. More recently, machine learning algorithms have been used. [3,7,30]. In Spain, [46] use Random Forest (RF) to improve daily Reference Potential Evapotranspiration (ET0) series from models associated with climate change scenarios, comparing their results with three methods based on weighted averages (including BMA), six linear-based combination MME (four regression-based methods and two geometric-based methods) and Support Vector Regression (SVR), concluding that RF is the best fit to the observed data and best represents climate variability in the climate variables studied.

However, there are few works that propose the use of different products and their ensemble to improve the results, among them the study by [58] for daily precipitation, although they use only 12 stations. Perhaps the most important in Europe is the ensemble-based project for observational grids E-OBS [14], although not from the perspective of our study, but from conditional simulations. Therefore, in this study we try to answer some questions that arise when approaching a study integrating this type of products for climatic variables.

- Which of the available grids best fits the observed data on average?
- Is it possible to improve the representation of temperature and precipitation grids in a relatively simple way using MME techniques?
- Is there a real benefit in investing in multiple grid generation projects?

This study has a twofold objective: a) to evaluate the accuracy of the available grids for Spain at monthly scale with observations from weather stations not used in their gen-

eration, and b) to use MMEs based on the RF algorithm to improve the estimation accuracy of monthly weather data for minimum (MinT), maximum (MaxT) and average (AvgT) temperature, and precipitation (TotP). To this end, the following hypotheses are made: a) ensemble techniques can improve the fit of individual grids, and b) the use of additional climatic data to those used in the grids can improve the results obtained. In addition, we have evaluated the results of two validation strategies, (i) validation of the prediction at a single point in space (by partitioning the dataset into calibration-validation vs. test) and (ii) validation of spatio-temporal predictions (one-to-one spatial cross-validation with distance buffer -LOOBUF-CV-); these strategies are derived from the prediction strategy used in studies using climate data, implemented in this study: (i) point prediction at a single climate station (per-station prediction, using only the information associated with individual stations and generating MMEs with multiple RF models), and (ii) spatially distributed grid prediction (using a single model that incorporates all the information of the nearest stations and their distance using Random Forest Spatial Interpolation -RFSI-).

2 Material and methods

The proposed framework has been carried out by integrating seven networks of meteorological stations and eight datasets of spatio-temporal grids available for Spain (excluding the archipelagos), all generated with heterogeneous methodologies. The variables studied were maximum, minimum and average temperature, and precipitation, with a monthly time step. The time interval ranges from January 2000 to December 2014 (180 months) and was chosen based on the objectives of the study, for which it was necessary to have information from all the grids and a sufficiently large number of observations. As a result, an RFSI-based spatio-temporal grid has been generated from the ensemble.

Figure 1 summarises the work process carried out to achieve the objectives. (A) summarises the process carried out to download and process the information described in subsection 2.2, using the observed data from 961 stations belonging to seven official climate and agroclimatic networks of weather stations with daily information available for download, as well as the climate data grids from eight projects in Spain. Once the information was processed and debugged in a Quality Controlled data (QC data) step, the MMEs were generated and their validation was carried out (subsections 2.3 and 2.4), for which the series have been associated at the station level and the two MMEs strategies and their validation strategies are generated. (B) spatio-temporal ensembles whose objective is to generate grids (RFSI MMEs and LOOBUF-CV cross-validation strategy) and (C) temporal ensembles per-station (SPLIT-V validation strategy). Finally, in section (D) we analyse the accuracy of fit of the MMEs and the two previous validation strategies (subsection 2.4).

The spatial information was processed using Quantum GIS [43]. The data analysis programme R-CRAN version 4.3 [44] was used to debug and process the data and to implement the MMEs. The author's experimental environment was as follows OS: Linux Debian 12; CPU: Intel(R) Core (TM) i9-13900K 5.8GHz; RAM: 128GB; GPU: NVIDIA GeForce RTX-4080.

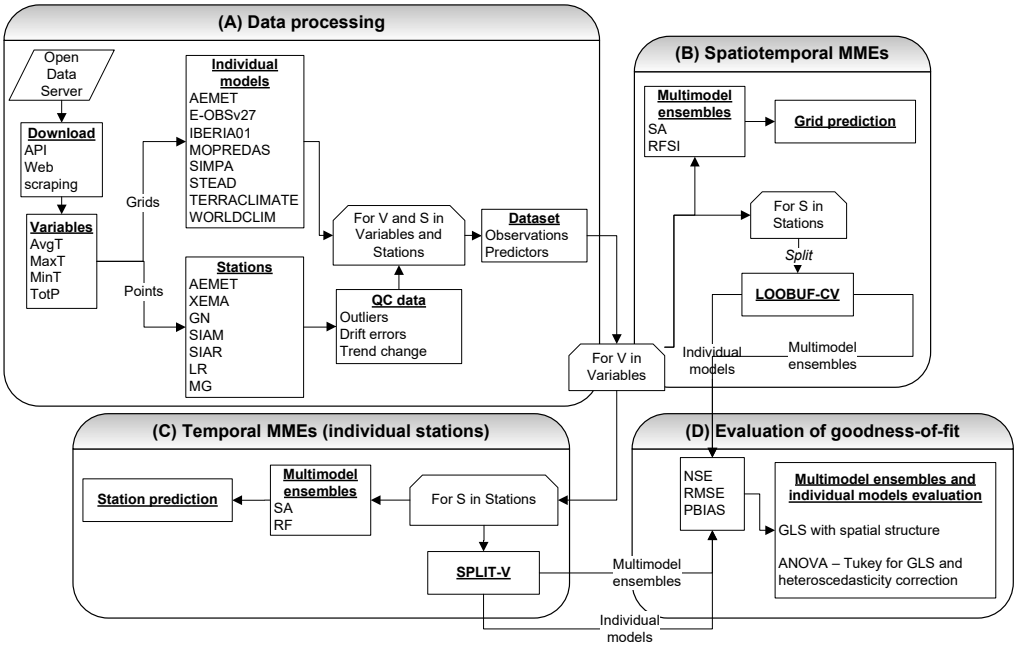

Figure 1: Conceptual summary of the work process.

2.1 Study area

The study area covers the Spanish territory of the Iberian Peninsula, it is approximately located between 36ºN and 43ºN in latitude, the orographic distribution and altitudes of the mountain systems are irregular. In addition, its location between the Atlantic Ocean and the Mediterranean Sea and its territorial extension of approximately 580,000 km^2, including Portugal, introduces the continental factor [13]. This geographical context generates large climatic variations in terms of the territorial distribution of temperature and precipitation.

Average annual temperatures in the north and mountainous areas are generally low, especially in the Cantabrian Mountains and the Pyrenees, with values below 10°C; in the central plateau and southwest of the peninsula they are moderate, between 10 and 15°C, with cold winters and hot summers [12]. Towards the Mediterranean and the southwest of the peninsula, temperatures are milder, between 15 and 20°C, with mild winters and warm summers, and the coastal regions of the south and the Guadalquivir valley have the highest average temperatures in the country, exceeding 20°C in many areas, with very hot summers and mild winters [2]. In terms of rainfall, the peninsula is divided into four regions according to its annual distribution [33]: (i) the western peninsula, with abundant rainfall in autumn and winter and very little in summer, (ii) the centre of the peninsula, with less abundant rainfall but more regular throughout the year, except in summer, when

it decreases significantly, (iii) the Mediterranean coasts, with little rainfall for a large part of the year, except in autumn, and (iv) the northeastern interior, where the maximum rainfall occurs in the transitional seasons and the minimum in summer and winter.

2.2 Datasets

The study was carried out on observed data from 961 stations belonging to seven networks of meteorological and agroclimatic stations with downloadable daily information (Figure 2):

1. 218 weather stations from the spanish *Agencia Estatal de Meteorología* (AEMET[1]) (State Weather Agency), data was downloaded using the AEMET OpenData API. These stations are spread throughout Spain, with an average altitude of 395 m and $SD = 426$ m.
2. 155 agro-climatic stations from the *Generalitat de Catalunya* (XEMA[2]) (Catalan Government), data was downloaded from the web site of the *Generalitat de Catalunya* and using its API. This network is concentrated in the Autonomous Community of Catalonia, a region near the coast located in the northeast of Spain, with an average altitude of 529 m and $SD = 614$ m.
3. 27 automatic agro-climatic stations from the network of meteorological stations of the *Comunidad Autónoma de Navarra* (GN[3]) (Navarre Government) with information available on its web site. It is a network located in the Autonomous Community of Navarre, in northern Spain, with an average altitude of 682 m. and $SD = 322$ m.
4. 47 agro-climatic stations from *Sistema de Información Agroclimático de la Región de Murcia* (Agricultural Information System of Murcia) (SIAM[4]) downloadable from its web site. This network is located in the Region of Murcia, a semi-arid area of the Mediterranean in south-eastern Spain, with an average altitude of 262 m and $SE = 238$ m.
5. 393 agroclimatic stations fro *Sistema de Información Agroclimático para Riegos* (Agroclimatic Information System for Irrigation) (SIAR[5]) available through its API. It is a network spread throughout Spain like the AEMET network, but with agrometeorological objectives, its average altitude is 416 m and $SD = 286$ m.
6. 13 weather stations from the *Comunidad Autónoma de La Rioja* (Autonomous Community of La Rioja) (LR[6]) downloadable from their web site. It is located in northern Spain, with an average altitude of 894 m and $SD = 415$ m.
7. 109 weather stations from the *Xunta de Galicia* (Galician Government) meteorological network (*MeteoGalicia*) (MG[7]) downloadable from its web site. It is located in northwestern Spain, with an average altitude of 529 m and $SD = 614$ m.

Although there are many stations in all networks, the data made available to end users comes from stations that have undergone rigorous quality control. We performed a second check using the R-CRAN *climatol* package [22], described in [21], and time series with

[1] https://www.aemet.es/es/datos_abiertos/AEMET_OpenData
[2] https://ruralcat.gencat.cat/agrometeo.estacions
[3] http://meteo.navarra.es/estaciones/descargardatos.cfm
[4] http://siam.imida.es/apex/f?p=101:46:4024577541991326
[5] https://servicio.mapa.gob.es/websiar
[6] https://www.larioja.org/agricultura/es/informacion-agroclimatica/red-estaciones-agroclimaticas-siar
[7] https://www.meteogalicia.gal/observacion/estacions/estacions.action?request_locale=en

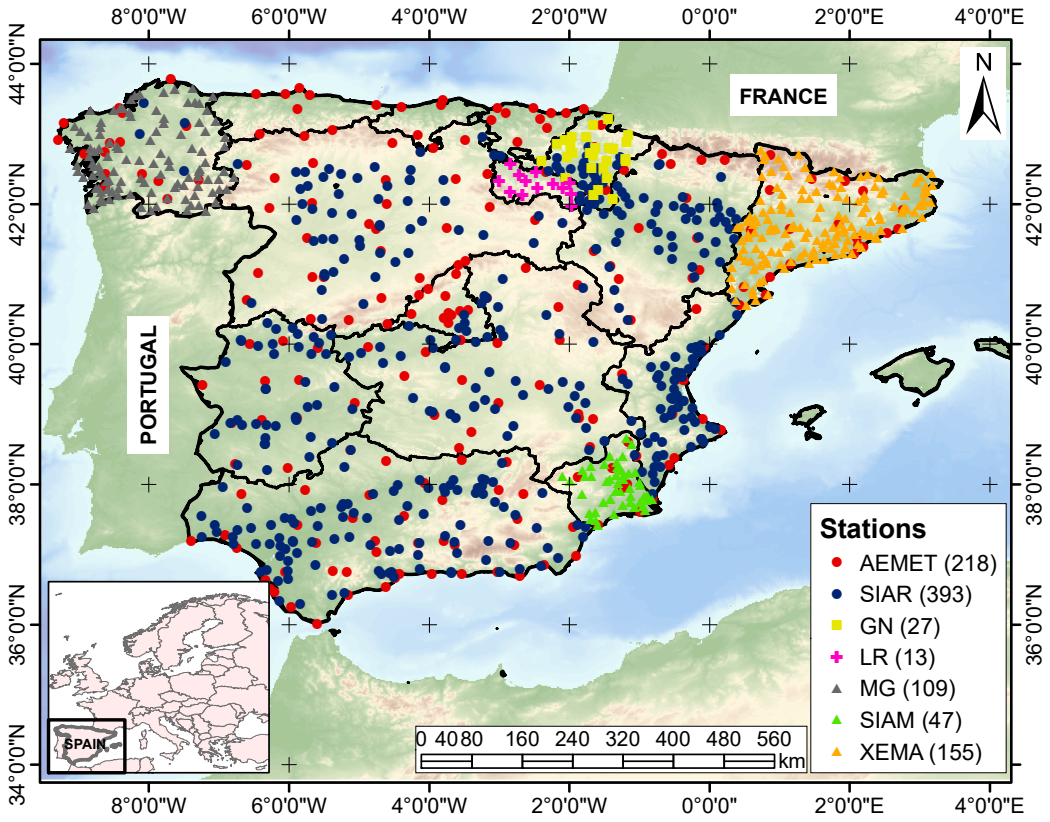

Figure 2: Spatial distribution of the meteorological and agroclimatic stations used in this study.

a daily time step to ensure there were no errors. In this process, the observed series at each station are compared with a reference series estimated as the inverse of the distance-weighted average of the nearest stations. Once the reference series has been generated, differences are obtained and standardisation is performed. Initially, data was collected from 1,029 stations. First, a preliminary analysis of inhomogeneities was performed using standardised differences and the Standard Normal Homogeneity Threshold (SNHT) climatological contrast, as described in [4, 21]. In the event of inhomogeneities, it was decided not to correct the series, but rather to remove the stations from the remainder of the work process. As a result, 962 stations were included in the analysis (Figure 2). Second, outliers can be identified based on an extreme value threshold (conservative values of $sd = \pm 25$ for maximum temperature, $sd = \pm 20$ for minimum temperature, $sd = \pm 30$ for average temperature and $sd = \pm 50$ for total precipitation); As a result, 486 precipitation records, 141 maximum and minimum temperature records, and 40 average temperature records were identified and set to zero so as not to include estimates based on spatial interpolation. Finally, data from stations with more than 60 months of data throughout the period

from 2000 to 2015 was selected. As a result, 936 stations were selected for monthly average temperature, 934 stations for monthly maximum temperature, 935 stations for monthly minimum temperature and 914 stations for monthly total precipitation.

Eight climate data grids (Table 1) were used as covariates. These grids were produced using heterogeneous spatial interpolation methods and data from stations with strict quality controls. The Spanish Meteorological Agency (AEMET) grid [5] is the official reference dataset in Spain. It uses an optimal interpolation algorithm based on historical analyses of the HIRLAM numerical prediction model, which takes orography and seasonal variation into account. This is then corrected using observed data from around 2,300 precipitation stations. The IBERIA01 dataset [27] was generated for the Iberian Peninsula using 3,486 precipitation stations and 275 temperature stations. For spatial interpolation, it uses a combined method based on thin plate splines (3D-TPS) together with ordinary kriging (OK) for daily anomalies. SIMPA [36] is a grid generated by kriging interpolation with external drift based on elevations to incorporate the influence of topography. It uses data from over 4,000 precipitation stations and over 1,000 temperature stations from regional networks and its own network. MOPREDAS [8] is a specific precipitation dataset that incorporates up to 5,234 stations in a variable manner. It uses a two-stage interpolation algorithm: Indicator Kriging (IK) is used to predict the probability of dry months, while Universal Kriging (UK) estimates the amount of precipitation using covariates such as coordinates, altitude, and distance to the coast. STEAD [50] is a specific dataset for daily temperatures that uses 5,520 weather stations in a variable manner. It employs an interpolation algorithm based on generalised linear mixed models (GLMMs) and generalised linear models (GLMs), using the same covariates as MOPREDAS. EOBSv27 [14] is a European dataset incorporating 218 temperature and precipitation stations in Spain from the AEMET and XEMA networks. It performs a two-stage interpolation: a generalised additive model (GAM) with altitude, followed by an interpolation of the residuals using random Gaussian field (GRF) simulation. WORLDCLIM or CRU-Ts4.0 [24] is a global dataset that uses Angular-Distance Weighting (ADW) as its interpolation algorithm. TERRACLIMATE [1] is a global dataset that uses climatologically aided interpolation to superimpose monthly climate anomalies from CRU-Ts4.0 and JRA-55 with monthly climate normals from WORLDCLIM.

The daily time step grids have been aggregated by month and assigned as a time series to the selected stations. The data used cover the time interval from January 2000 to December 2014, which is the time interval in which it is ensured that all grids have information simultaneously. As detailed in subsection 2.4, the AEMET network was used to calibrate the models, but the final validation was carried out without this network in order to make a more honest accuracy estimation, given that eight climate data grids use at least this network for their data collection.

2.3 Multi-model ensembles

Three MME methods were used in this study: (i) simple average (SA), (ii) R and (iii) RFSI, a method based on RF but incorporating the spatial component in the models. RF was used to construct the ensembles in the per-station approach, so that one model was obtained for each station (961 models per variable), whereas the second approach (RFSI) generates a single model per variable (Figure 1).

Simple Average (SA) is the most used method in this type of study because of its simplicity and ease to interpret, it has been used as a base model for assessing the ability of

Dataset	Extent	Spatial Res.	Temporal Res.
AEMET [5]	Spain	0.05º	1951-2021 (daily)
EOBSv27 [14]	Europe	0.1º	1950-2022 (daily)
IBERIA01 [27]	Iberian Peninsula	0.1º	1971-2015 (daily)
MOPREDAS [8]	Spain	0.1º	1915-2019 (monthly)
SIMPA [36]	Spain	500 m	1950-2020 (monthly)
STEAD [50]	Spain	0.05º	1901-2014 (daily)
TERRACLIMATE [1]	World	0.04167º	1958-2022 (monthly)
WORLDCLIM [24]	World	0.04167º	1960-2021 (monthly)

Table 1: Climatic grids used in this study. MOPREDAS only includes precipitation variables. STEAD only includes temperature variables. SIMPA only includes monthly total precipitation and monthly average temperature.

more complex methods. Its main disadvantages are that it does not weight the importance of each grid in the combination and does not correct for problems of systematic bias of all members.

RF [11] is a non-parametric algorithm based on ensembles of decision trees and bagging (bootstrap aggregation). Each tree is calibrated using a bootstrapped subsample of cases, and the features to perform each split in the trees are selected from a random subsample of the total feature set. When all trees have been trained without pruning, each new case is analysed by all trees and the final prediction is obtained by averaging the results. RF uses two parameters: the number of regression trees to grow (n_{tree}), and the number of features randomly sampled in each split (m_{try}). Once the model is trained, the prediction can be obtained as the equation 2.

Its advantages over traditional methods include estimating the importance of each predictor in the model, not requiring a priori statistical assumptions, being less sensitive to outliers, and not overfitting the models. Moreover, authors such as [25] or [20] conclude that by setting default parameters, the results obtained are very similar to the calibrated model; therefore, in this study we have used the default value for regression problems $m_{try} = p/3$, where p is the number of predictors, and $m_tree = 1000$, as higher values generally do not produce any improvement [42]. Another advantage of RF is its ability to infer the importance of each grid in the final ensemble model. In the case of this study, we used the percentage increase in the mean standard error, which estimates the increase in the mean squared error when a variable is removed. For each tree that makes up the forest, the MSE is calculated with the Out-of-Bag (OOB) as an internal estimate of the model error using observations that were not selected in the bootstrap of each tree; then a random permutation is performed on one of the input variables, which produces a change in the MSE; this process is repeated for all the input variables and all the trees, and for each permuted variable the MSE and the difference with respect to the initial value are obtained; at the end of the process, the estimated values are averaged. In this study, the fast implementation included in the ranger package [57] was used.

The problem with RF in the case of spatial data is that, as a global regression model, it does not reproduce well local patterns in spatially distributed models. Several proposals have been made to address this aspect, such as Random Forest Regression Kriging (RFRK), a version of regression kriging methods that incorporates as a spatial component the spatial interpolation of an ordinary kriging of the residuals. In other cases, it has been proposed

to include the distance matrix as a spatial predictor in the model in order to minimise autocorrelation in the residuals. [26] propose the Random Forest for spatial predictions framework (RFsp) which includes buffer distance maps to all observation locations as covariates, and [10] spatialRF uses the distance matrix as a predictor. The advantage of this type of framework over RFRK is that it does not require the assumptions of ordinary kriging or a prior study of spatial autocorrelation, yet it achieves similar results [26, 48]. In this study, we used the proposed RFSI proposed by [48], where the added covariates are defined as the observations at the n nearest locations and the distances from these locations to the predicted location. The RFSI final prediction can be expressed as:

$$\hat{z}\left(s_0\right) = f\left(x_1\left(s_0\right), ..., x_m\left(s_0\right), z\left(s_1\right), d_1, ..., z\left(s_n\right), d_n\right) \tag{1}$$

whereas the simpler RF predictive equation would be:

$$\hat{z}\left(s_0\right) = f\left(x_1\left(s_0\right), ..., x_m\left(s_0\right)\right) \tag{2}$$

In these last two equations, $x_i\left(s_0\right)\left(i = 1, ..., m\right)$ are covariates (individual grids) at location s_0, $s_i\left(i = 1, ..., n\right)$ are the $i - th$ nearest observation location from s_0, and d_i are the distances between s_i and s_0. The number of nearest weather stations to include in each estimation is a model hyperparameter to be set. In this study, after several tests, we have used $n = 5$. In [48] the RFSI algorithm is explained in detail.

Once the models had been estimated, the final predictions were made using a compromise solution: predictions were generated at 5 km in an attempt to minimise the impact of the scale mismatch and change of support problem (COSP) [15, 18] in accordance with the original resolutions 1. For this purpose, a dual strategy was employed for the simple members used as covariates: the lower-resolution climatic grids were aggregated using a simple average, and reinterpolation using Thin-Plate Splines was employed for the lower-resolution grids.

2.4 Accuracy estimation

A dual validation strategy was used to validate the results. The first one is based on (a) station-by-station temporal validation with simple temporal partitioning (SPLIT-V), which is applied for point prediction at a single climate station, and the second is based on (b) buffered/spatial leave-one-out cross-validation (LOOBUF-CV), used for spatially distributed grid prediction using RFSI.

The first type of validation is based on a simple partitioning of the alternate years of the time series, from January 2000 to December 2014 (180 months), for each station, using the months of the even years (96 months) for calibration and the other half of the months corresponding to the odd years for testing. In this way, we have tried to reduce the effect of different time patterns that may occur in the $2/3$ division strategy for calibration in the first part of the time series versus the remaining $1/3$ at the end of the time series.

In the second case, the LOOBUF-CV validation strategy is based on a leave-one-out cross-validation (LOO-CV), but discards for calibration all stations farther than a threshold (15 km in this work) from the station used for validation in each iteration. This procedure tries to reduce the problems of accuracy overestimation due to the spatial autocorrelation of the weather stations. In this work, we have used the version included in the R-CRAN package *blockCV* [54].

However, analysing the validation results using all the stations is not representative for the purposes of this study, since most of the stations used may have been involved in calibrating the individual grids. An example of this situation can be seen in most of the AEMET network stations, where the NSE is close to 1. Ideally, the validation results would be analysed using reference stations that were not used for calibration in any of the grids. However, this information is not available. Therefore, a more realistic approach was to remove all AEMET stations from the analysis set. However, some of the remaining stations may still have been used, which could lead to an overestimation of the goodness of fit of the individual grids. While other stations are used by the grids (e.g. the EOBS27 grid uses many of the XEMA network stations, and the Iberia01 grid uses some of the SIAR network stations), the final validation results are more realistic without the AEMET weather stations. Ultimately, 743 stations were used for AvgT and MaxT, 739 for MinT, and 732 for TotP.

The accuracy of the models was estimated by three statistics: (i) Root Mean Square Error (RMSE), (ii) Percent Bias (Pbias) and (iii) Nash Efficiency Index (NSE). The RMSE (equation 3) is a measure of absolute error (in the same units as the variable under consideration) that measures the difference between the values predicted by a model and the actual observed values; it ranges from 0 (perfect fit) to $+\infty$ (high error).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{obs,i} - y_{sim,i}\right)^2}{n}} \tag{3}$$

where n is the number of data points, $y_{obs,i}$ is the $i-th$ observation for the variable, and $y_{sim,i}$ is the corresponding prediction.

Pbias (Equation 4) measures the average tendency of the simulated data to be larger or smaller than its reference values. Positive values indicate underestimation, and negative values indicate overestimation.

$$Pbias = \frac{\sum_{i=1}^{n} \left(y_{obs,i} - y_{sim,i}\right)}{\sum_{i=1}^{n} y_{obs,i}} \cdot 100 \tag{4}$$

NSE (equation 5) is a dimensionless index with a similar interpretation to R^2, but more robust to the problems of the latter, such as sensitivity to extreme biases and insensitivity to constant biases (additive or multiplicative).

$$NSE = 1 - \frac{\sum_{i=1}^{n} \left(y_{obs,i} - y_{sim,i}\right)^2}{\sum_{i=1}^{n} \left(y_{obs,i} - \overline{y}_{obs}\right)^2} \tag{5}$$

Table 2 includes recommended performance evaluation criteria proposed by [37] for NSE and Pbias.

Classification	Pbias	NSE
Very Good	$Pbias < \pm 5.0$	$0.80 < NSE < 1.00$
Good	$\pm 5 < Pbias < \pm 10$	$0.70 < NSE < 0.80$
Satisfactory	$\pm 10 < Pbias < \pm 15$	$0.50 < NSE < 0.70$
Unsatisfactory	$Pbias > \pm 15$	$NSE < 0.50$

Table 2: Performance evaluation criteria for models evaluation based on [37] at monthly temporal scale.

The significance of differences between models was assessed by one-way analysis of variance (ANOVA). When ANOVA discovers significant differences, the effects of the differences between the different models were assessed using Tukey-Kramer pairwise contrasts, identifying homogeneous subsets of significance; in Figures 3 to 5 these groupings are included as letters at the top of the boxes, indicating whether their means are significantly different (boxes in subgroups with different letters) or not (subgroups with the same letter). Normality and homoscedasticity were assessed using the Kolmogov-Smirnov and Levene tests. In cases where heteroscedasticity was present, a heteroscedasticity-consistent covariance matrix of the parameters (HC3) [31] was used in ANOVA and the Tukey-Kramer contrast to correct for heteroscedasticity.

In order to estimate the internal uncertainty of RFSI, we used the standard deviation of the predictions of the 1000 individual trees for each month and year. This is a widely accepted method that captures the internal variability of the model [26, 48].

3 Results

As mentioned in subsection 2.4, in order to obtain a more honest accuracy estimate, the AEMET stations, used to generate the grids included in the ensembles, are not used in the validation. Nevertheless, it is likely that some of the stations used to generate the grids are still used in the validation, meaning the accuracy may be somewhat overestimated. Therefore, as a preliminary step to evaluating the results, all stations (including AEMET) were validated, and those with $RMSE < 0.05$ in the variables $Avg.T.$ and $Tot.P.$ for SPLIT-V validation (with a few exceptions all of them belonged to the AEMET network), were removed prior to calculate the accuracy statistics.

In all cases, and for both types of validation, the ANOVA results (Table 3 and 4) show the existence of at least one significant difference between grids, MMEs, or grids and MMEs.

Variable	NSE		RMSE		Pbias (absolute)	
	F value	Pr	F value	Pr	F value	Pr
AvgT	142.910	< 0.0001	333.530	< 0.0001	187.300	< 0.0001
MaxT	45.591	< 0.0001	188.090	< 0.0001	121.560	< 0.0001
MinT	136.720	< 0.0001	312.870	< 0.0001	153.570	< 0.0001
TotP	73.119	< 0.0001	50.651	< 0.0001	48.950	< 0.0001

Table 3: Summary of the one-way ANOVAs for goodness of fit statistics in validation. SPLIT-V (per-station single temporal partition).

Subsections 3.1 and 3.2 include detailed results for RF (per-station) and RFSI strategies. The results seem to indicate that both strategies improve the fit obtained compared to individual models.

3.1 RF (per-station) results

Figures 3 to 5 summarize the per-station accuracy results for the validation (single temporal partition -SPLIT-V- summarized for all stations) used to evaluate the first of the ensemble strategies: one model per station to generate point-wise ensemble series. These

Variable	NSE		RMSE		Pbias (absolute)	
	F value	Pr	F value	Pr	F value	Pr
AvgT	51.658	< 0.0001	114.090	< 0.0001	15.810	< 0.0001
MaxT	14.534	< 0.0001	57.210	< 0.0001	3.451	0.001
MinT	65.712	< 0.0001	115.060	< 0.0001	17.766	< 0.0001
TotP	98.351	< 0.0001	51.369	< 0.0001	38.816	< 0.0001

Table 4: Summary of the one-way ANOVAs for goodness of fit statistics in validation. LOOBUF-CV (buffered/spatial leave-one-out cross-validation).

figures show the distribution of values per station, with boxes ordered by accuracy and distributed into groups of non-significantly different models, according to a posteriori contrast (letters at the top). For example, Figure 3 (A-C) shows that RF is only included in group (a). This indicates that, according to the a posteriori contrast, the mean obtained from NSE in validation is significantly better than the mean of the other MMEs with which is compared. However, in Figure 3 (D), RF is again grouped as (a), indicating that its validation fit is better, but SA, SIMPA, and AEMET belong also to group (a), indicating that RF is not significantly better than them despite also belonging to group (b). Table 5 contains the remaining accuracy summary table.

The NSE values are high for the four variables (Figure 3), although with some outliers (usually stations not used to estimate the grids). This may be because if all the simple members have not been estimated correctly at a particular station, the ensemble may not be good. However, it may be that, even if some of the simple members do not fit well, if the rest of the members fit well, abnormally low values may be obtained in a specific member but high values in the rest and in the MMEs. For this reason, for example, Figure 3 (A) shows how in RF the number of abnormally low values is significantly reduced (in all stations $NSE > 0.55$. In the four cases, RF obtains the best average NSE and a smaller dispersion in the results; this model fits, on average, the observations significantly better than the individual grids for the three temperature variables (group a in Figure 3 A to C) and the monthly total precipitation (Figure 3 D), although in this latter case it does not seem to be significantly different from SA, SIMPA or AEMET. In the case of temperature variables, STEAD accuracy appears in second place. This is a dataset generated specifically for temperature using data from more than 5000 stations [51]. SA is ranked third because the fit of the individual grids is already very good and no general biases or a large number of outliers are observed in them; in those cases SA is accurate enough [6]. In the case of precipitation, SIMPA stands out after SA, mainly because it is a dataset specifically created for variables related to hydrology and a monthly time step, using a large number of rain gauges from the Automatic Hydrological Information System (*Sistema Automático de Información Hidrológica* - SAIH, the Spanish Government's main weather and gauging information network). The same behaviour can be observed in RMSE (Figures 4 and 5), with groupings very similar to the previous figure: RF is better in all four variables, being significantly better in the case of the temperature related variables and having a lower dispersion; it is also noteworthy the low dispersion around 0 of Pbias.

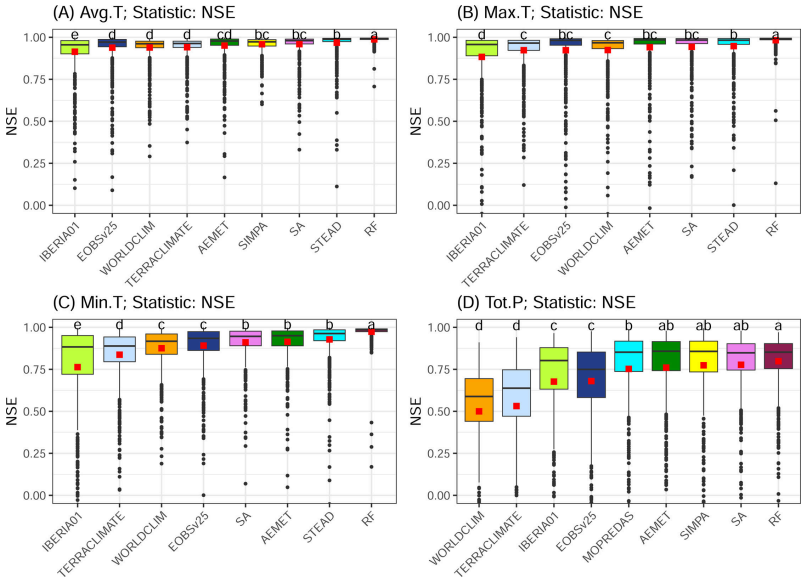

Figure 3: MMEs by station (SPLIT-V) using RF approach. Results of NSE in validation (Table 5). Significantly different groups (Tukey–Kramer contrast using HC3, alpha = 0.05) are represented by different letters at the top of the figure. (A) AvgT, (B) MaxT, (C) MinT, (D) TotP.

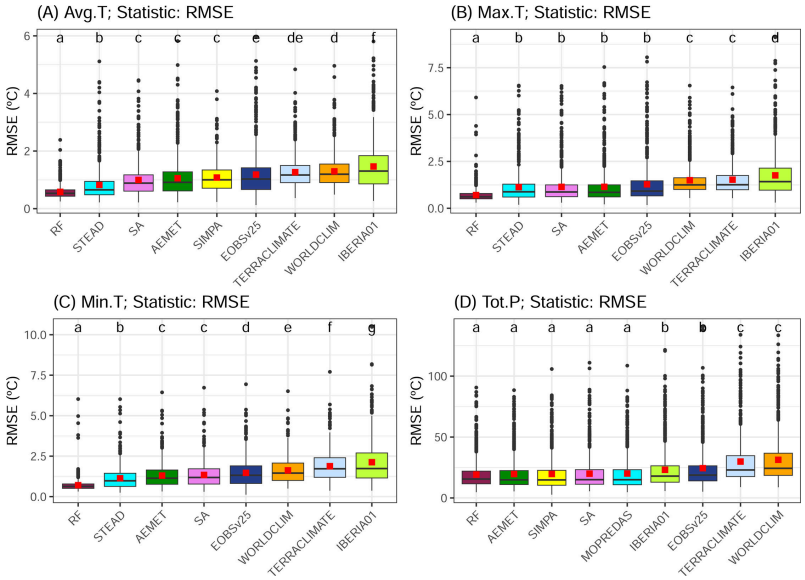

Figure 4: MMEs by station (SPLIT-V) using RF approach. Results of RMSE in validation (Table 5). Significantly different groups (Tukey–Kramer contrast using HC3, alpha = 0.05) are represented by different letters at the top of the figure. (A) AvgT, (B) MaxT , (C) MinT , (D) TotP.

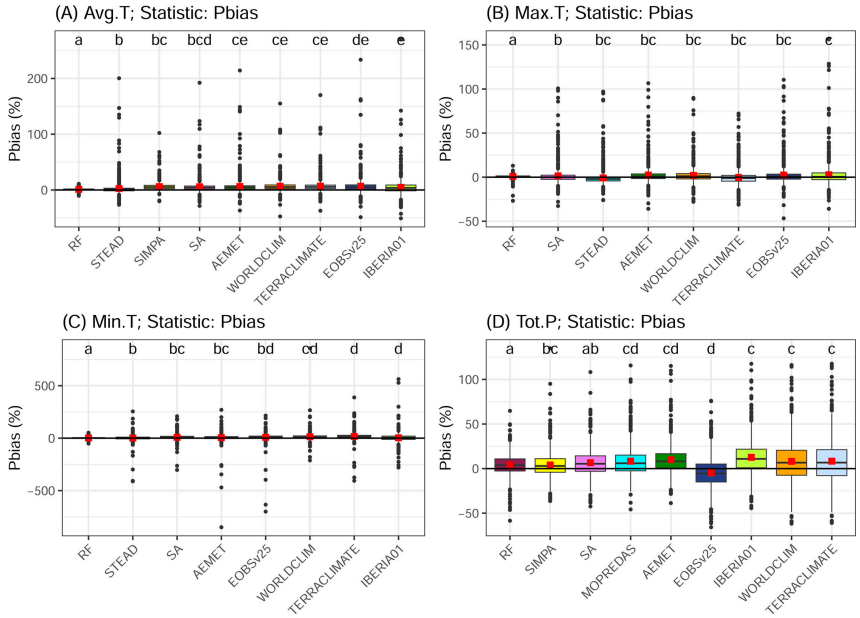

Figure 5: MMEs by station (SPLIT-V) using RF approach. Results of Pbias in validation (Table 5). Significantly different groups (Tukey–Kramer contrast using HC3, alpha = 0.05) are represented by different letters in top of figure; The ordering and a posteriori contrasts were performed for the absolute Pbias, while the graphical representation was performed for Pbias to better interpret the underestimation or overestimation of MMEs. (A) AvgT, (B) MaxT , (C) MinT , (D) TotP.

Figure 6 shows the per-station (mean variable importance across all stations) importance for the RF model of the individual grids used as predictors. Figure 6 (A-C) shows the importance for the models of temperature variables. The behaviour in terms of importance is similar to that observed in RFSI: in this case, STEAD is the most important predictor, followed by AEMET and EOBSv27. For AvgT, SIMPA moves into fourth place, albeit with values similar to those of the top three. In all three cases, the remaining grids are some distance behind the top three. For the monthly total precipitation (Figure 6 D), SIMPA is the most important individual grid, followed by AEMET and MOPREDAS. This behaviour, together with the goodness of fit observed in the simple members Figure (3 and Figure 5), could be due to the difference in the initial spatial resolution of the grids. In general, the best simple members are those with higher spatial resolution; this may be due to the effects derived from the COPS (Change of Support Problem) [15], where grids with lower spatial resolution tend to lose information derived from spatial variability and generalise the variable studied [18].

AvgT	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.951	0.007	1.057	0.048	6.307	1.105
EOBSv27	0.938	0.008	1.184	0.053	6.677	1.210
IBERIA01	0.914	0.009	1.464	0.061	4.677	1.236
SIMPA	0.959	0.003	1.081	0.036	6.007	0.552
STEAD	0.968	0.005	0.820	0.042	2.507	1.031
TERRACLIMATE	0.941	0.006	1.268	0.039	6.997	0.892
WORLDCLIM	0.939	0.006	1.295	0.040	6.667	0.830
RF	**0.988**	0.001	**0.583**	0.017	**1.076**	0.100
SA	0.959	0.005	0.996	0.041	5.692	0.927
MaxT	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.941	0.011	1.145	0.068	2.429	0.791
EOBSv27	0.923	0.014	1.279	0.078	2.448	0.899
IBERIA01	0.883	0.017	1.757	0.086	2.745	1.061
STEAD	0.948	0.009	1.128	0.063	**-0.836**	0.747
TERRACLIMATE	0.923	0.009	1.525	0.060	-0.293	0.702
WORLDCLIM	0.924	0.010	1.495	0.058	2.073	0.707
RF	**0.982**	0.005	**0.704**	0.029	0.860	0.136
SA	0.943	0.010	1.142	0.065	1.428	0.783
MinT	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.914	0.008	1.301	0.054	4.972	3.374
EOBSv27	0.891	0.010	1.469	0.062	8.416	3.347
IBERIA01	0.763	0.024	2.127	0.098	3.635	3.456
STEAD	0.928	0.008	1.137	0.050	**1.180**	2.371
TERRACLIMATE	0.837	0.012	1.880	0.064	16.727	2.890
WORLDCLIM	0.876	0.009	1.623	0.057	13.431	3.048
RF	**0.974**	0.005	**0.707**	0.029	1.573	0.344
SA	0.911	0.008	1.340	0.054	8.058	2.024
TotP	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.760	0.027	19.654	1.004	10.018	1.227
EOBSv27	0.680	0.019	24.442	1.279	-4.722	1.386
IBERIA01	0.677	0.034	23.104	1.173	12.618	1.437
MOPREDAS	0.753	0.026	20.056	1.071	8.274	1.303
SIMPA	0.774	0.023	19.704	1.076	**4.011**	1.116
TERRACLIMATE	0.531	0.032	29.870	1.405	8.304	1.814
WORLDCLIM	0.500	0.030	31.332	1.431	8.156	1.812
RF	**0.798**	0.012	**19.244**	0.900	4.189	0.894
SA	0.777	0.019	19.897	1.029	6.668	1.149

Table 5: Summary of validation goodness-of-fit metrics. Temporal validation per-station with simple temporal partition using RF approach, excluding the AEMET network. ci is a 95 % confidence interval of the average of the metric.

3.2 RFSI results

We used buffered/spatial leave-one-out cross-validation (LOOBUF-CV) to evaluate the accuracy of RFSI. Figures 7 to 9 and Table 6 show that the accuracy values are slightly lower than with RF. In this case, each weather station validates a model that has been calibrated

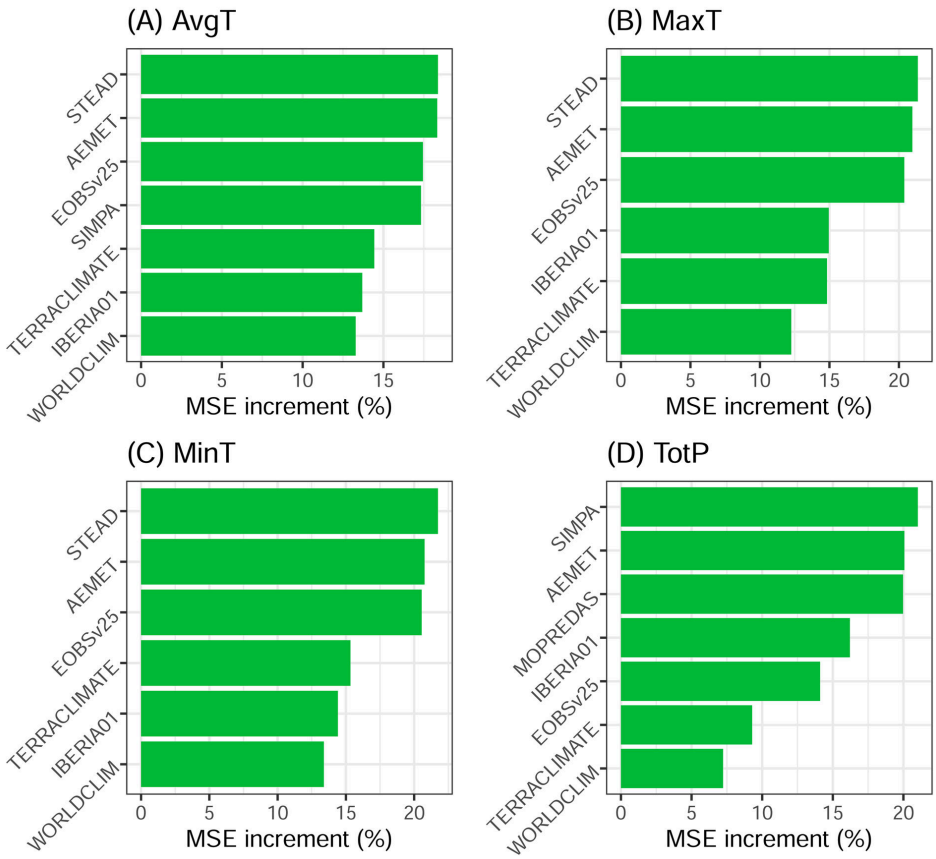

Figure 6: RF (per-station approach) predictor importance of MMEs based on the average increase, across all stations, of the Mean Standard Error in percent.

with weather stations located at least 15 km away. Nevertheless, the accuracy is still higher than that of individual grids.

It is also noteworthy that RFSI accuracy values have less dispersion, fewer outliers, and the number of stations with values considered abnormally low is significantly reduced; this is because in the first MME approach they were calibrated with the same data of the station series (which in many of them were not used to generate the grids of the simple elements), whereas now RF uses other nearby stations for this generation. In the case of precipitation for example, the number of low outliers decreases substantially with RFSI due to the use of neighbouring stations to generate the estimate. Nevertheless, in the case of NSE (Figure 7), groups of non-significantly different models show that RFSI is significantly better than the others for mean temperature and total precipitation (Figure 7 A and D), and its accuracy in predicting maximum and minimum temperatures is better than that of the others (Figure 7 B-C), although there are no significant differences compared with STEAD and AEMET.

As in Subsection 3.1, the best simple members are those with the largest spatial resolution, which may be due to the same COPS effect. This must also be taken into account

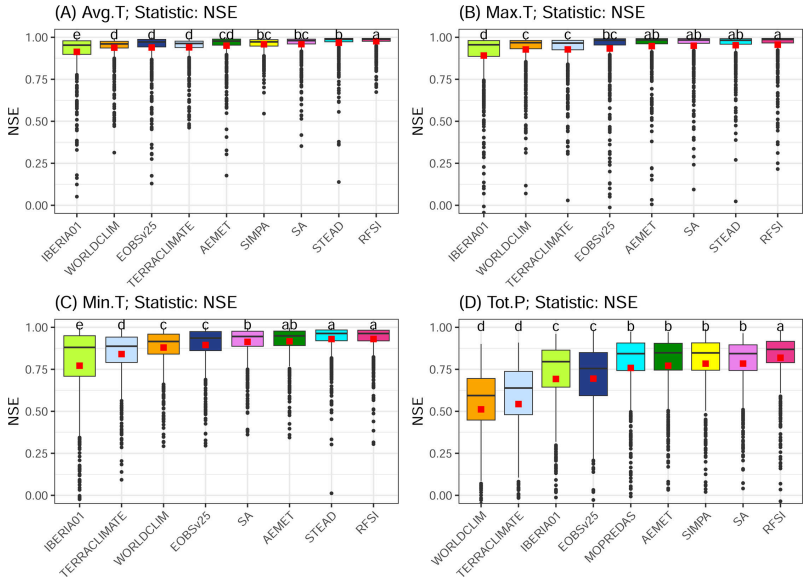

Figure 7: MMEs by station (LOOBUF-CV) using RFSI approach. Results of NSE in validation (Table 6). Significantly different groups (Tukey–Kramer contrast using HC3, alpha = 0.05) are represented by different letters at the top of the figure. (A) AvgT, (B) MaxT , (C) MinT , (D) TotP.

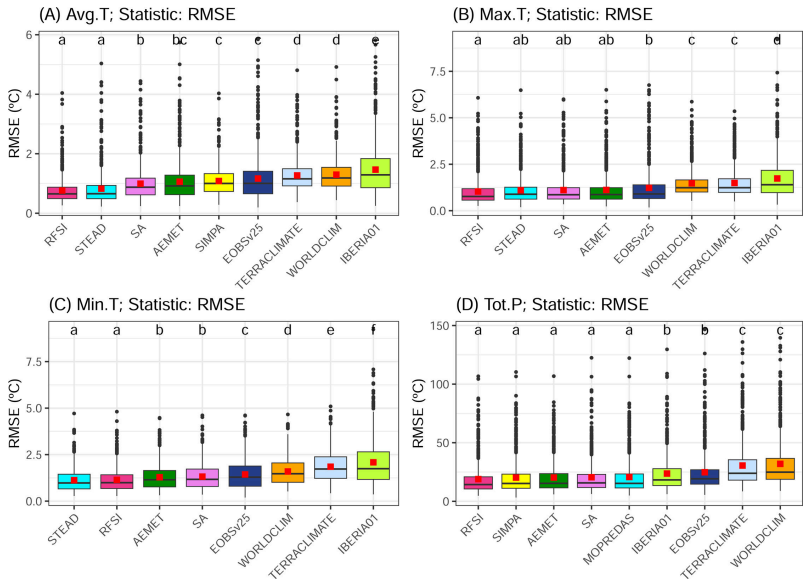

Figure 8: MMEs by station (LOOBUF-CV) using RFSI approach. Results of RMSE in validation (Table 6). Significantly different groups (Tukey–Kramer contrast using HC3, alpha = 0.05) are represented by different letters at the top of the figure. (A) AvgT, (B) MaxT , (C) MinT , (D) TotP.

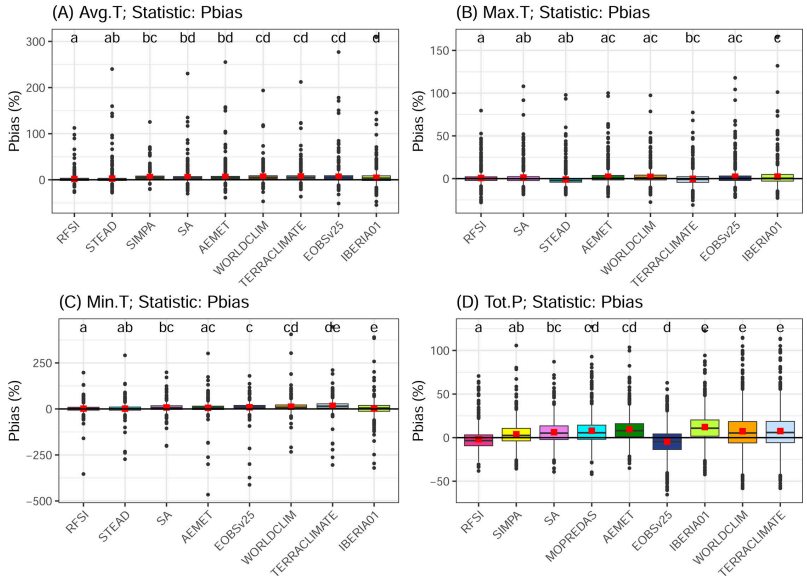

Figure 9: MMEs by station (LOOBUF-CV) using RFSI approach. Results of Pbias in validation (Table 6). Significantly different groups (Tukey–Kramer contrast using HC3, alpha = 0.05) are represented by different letters at the top of the figure; The ordering and a posteriori contrasts were performed for the absolute Pbias, while the graphical representation was performed for Pbias to better interpret the underestimation or overestimation of MMEs. (A) AvgT, (B) MaxT , (C) MinT , (D) TotP.

during the prediction process, in which the resolution has been resampled to 5 km. The three higher-resolution grids were aggregated using simple box averages, while three others have undergone downscaling by reinterpolation with thin-plate splines, which may produce artificial autocorrelation effects [18, 53], estimating the parameter λ using Generalised Cross-Validation (GCV).

It is also better than the others for RMSE (Figure 8) and Pbias (Figure 9), although with less significant differences than NSE (the results obtained in RF stand out in the case of Pbias) and following the same pattern as in the case of the individual models with simple validation: STEAD is the best simple component for temperatures and SIMPA for precipitation.

The results obtained are in line with other studies using RF with a spatial component [26, 48], probably due to the incorporation of geographical proximity effects in the prediction process. Both authors also emphasise that this type of strategy obtains better fits than other interpolation techniques such as Regression Kriging, Inverse Distance Weighting, or RF. Furthermore, the RFSI and RF per-station strategies are robust to the multicollinearity inherent in this type of study based on ensembles with the same variable from different sources. This is because they use node division with subsets of random variables and handle nonlinear data and interactions well [9]. Conversely, LOOBUF-CV seems an adequate validation strategy for evaluating the predictive performance of models calibrated with spatial data, as it enables the validation process to be spatially independent [34, 41].

AvgT	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.951	0.007	1.060	0.048	6.391	1.203
EOBSv27	0.939	0.008	1.165	0.054	6.652	1.325
IBERIA01	0.914	0.009	1.465	0.061	4.470	1.311
SIMPA	0.959	0.003	1.082	0.035	6.025	0.603
STEAD	0.968	0.005	0.825	0.042	2.594	1.129
TERRACLIMATE	0.940	0.006	1.270	0.039	7.272	1.001
WORLDCLIM	0.938	0.006	1.297	0.040	6.901	0.920
RFSI	**0.977**	0.003	**0.763**	0.031	**1.738**	0.643
SA	0.959	0.005	0.994	0.041	5.757	1.020
MaxT	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.948	0.009	1.116	0.060	2.176	0.694
EOBSv27	0.934	0.011	1.226	0.070	2.206	0.811
IBERIA01	0.891	0.015	1.734	0.080	2.339	0.958
STEAD	0.953	0.007	1.103	0.056	-1.092	0.652
TERRACLIMATE	0.928	0.008	1.493	0.056	-0.196	0.660
WORLDCLIM	0.928	0.008	1.477	0.055	2.141	0.643
RFSI	**0.956**	0.007	**1.034**	0.057	**0.688**	0.577
SA	0.950	0.008	1.109	0.058	1.262	0.698
MinT	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.917	0.007	1.286	0.049	6.011	2.419
EOBSv27	0.895	0.009	1.435	0.057	9.009	2.382
IBERIA01	0.772	0.021	2.086	0.089	1.845	3.107
STEAD	0.930	0.007	**1.126**	0.045	**1.640**	1.899
TERRACLIMATE	0.841	0.011	1.851	0.059	16.763	2.527
WORLDCLIM	0.880	0.008	1.594	0.052	12.187	2.237
RFSI	**0.931**	0.006	1.143	0.045	1.664	1.699
SA	0.914	0.007	1.318	0.049	7.981	1.728
TotP	NSE		RMSE		Pbias	
	Avg.NSE	ci	Avg.RMSE	ci	Avg.Pbias	ci
AEMET	0.772	0.021	20.330	1.003	9.543	1.143
EOBSv27	0.695	0.016	24.808	1.288	-4.712	1.294
IBERIA01	0.693	0.026	23.703	1.160	12.166	1.366
MOPREDAS	0.758	0.024	20.737	1.075	7.917	1.238
SIMPA	0.784	0.016	20.270	1.075	3.899	1.015
TERRACLIMATE	0.543	0.027	30.519	1.415	7.553	1.728
WORLDCLIM	0.512	0.026	31.975	1.449	7.322	1.720
RFSI	**0.819**	0.012	**18.954**	1.006	**-1.779**	0.954
SA	0.784	0.014	20.565	1.037	6.242	1.070

Table 6: Summary of goodness-of-fit and validation error metrics. Buffered/spatial leave-one-out cross-validation (LOOBUF-CV) using RFSI approach, excluding the AEMET network. ci is a 95 % confidence interval of the average of the metric.

Similarly to Figure 6, Figure 10 shows the importance of the variables in RFSI including as predictors both the individual grids and the observed data at the nearest stations and their distance (see equation 1). For temperature variables (Figure 10 A to C), the most important predictors are the data observed at the nearest stations, although the distances

of these with respect to the predicted data are not particularly relevant, being the least important predictors (pink bars in the figure). The difference between the individual grids is not very large; STEAD stands out for MaxT and MinT and is the second most important for AvgT (in this case, the most important is SIMPA, a grid that has no data for MaxT and MinT).

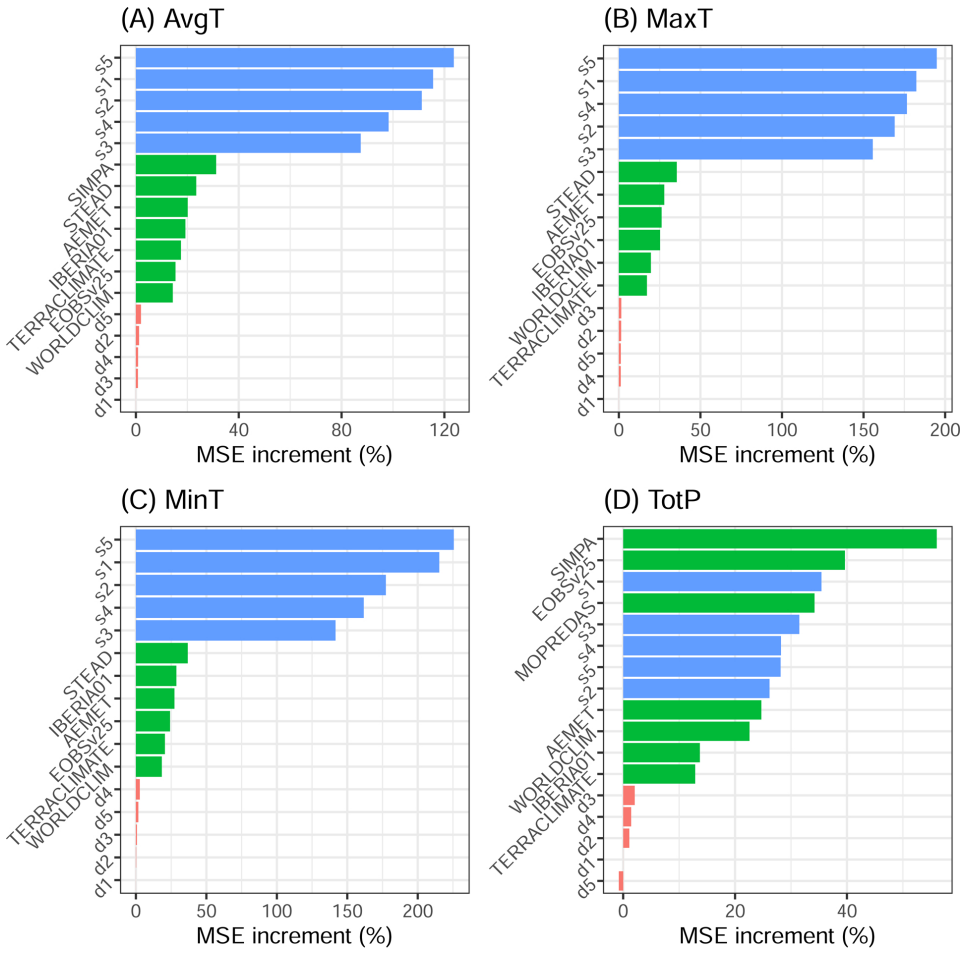

Figure 10: RFSI covariate importance of MMEs based on the increase in the Mean Standard Error in percent. s_i and d_i represent observations (blue color) and distances (pink color) to the $i-th$ nearest observation location; Green color represent individual model members.

For the monthly total precipitation (Figure 10 D), as well as in per-station RF (Figure 6 D), it is noteworthy that single grids are more important than in case of temperature variables compared to the nearest stations. SIMPA, estimated with a very large number of stations, monthly data and a spatial resolution of only 500 m, is the most important single grid followed by EOBSv27 and MOPREDAS (grid created specifically for precipitation); Therefore, the first most important station (s1, the closest station) is the third most important variable. It is likely that the grids are more important than the observations because

of the greater uncertainty in their spatial distribution and the strong dependence on local effects such as altitude or relief exposure. In addition, the distance between stations is often large, with an average inter-station distance of 10,300 meters and a maximum inter-station distance of 46.880 meters. (see distribution in Figure 2). Thus, spatial patterns included in the grid interpolation models become relevant.

Figures 11 and 12 show the spatial distribution of per-station values after the LOOBUF-CV validation. For variables AvgT (Figure 11 A), MaxT and MinT (Figure 12 A-B) the distribution is similar. The very high number of stations with NSE greater than 0.8 (very good performance) in AvgT indicates that this ensemble method achieves very good results. For precipitation TotP (Figure 11 B), 534 stations have an NSE greater than 0.80, 104 stations have an NSE between 0.7 and 0.8, and only 38 stations have an unsatisfactory performance. It is interesting to analyse the difference in accuracy between the two variables. As [32] point out, the results obtained depend to a large extent on the spatial density of weather stations, especially in the case of TotP, due to its greater uncertainty, which is partly caused by local factors. Conversely, AvgT achieves high or very high accuracy even in areas with a low density of weather stations.

These patterns do not appear to be random, but are mainly related to some of the autonomous weather station networks (GN, LR, MG and XEMA) in areas with rough topography and a low density of stations. For average temperature, only a few stations have an NSE value below 0.8, and no spatial pattern can be discerned in their distribution. For minimum temperature, lower NSE values appear in a cluster of stations on the border between the Aragonese and Castilian communities, in a mountainous area with a low density of AEMET stations. Therefore, the corresponding 5 km grid cell values may average a few very different situations. Similar problems are probably occurring in northern Catalonia. In the GN network (Galicia), problems appear along the community border, both on the coast and on the boundary with the rest of the country, in an area with low AEMET station density. Finally, relatively close stations in the SIAM network have NSE values between 0.5 and 0.7. For maximum temperature, the lowest NSE values appear in some stations in the Pyrenees area in the north of Catalonia and north of Navarre (XEMA and GN networks). The problems with the LR network are similar to the previous case. A pattern similar to that encountered in the SIAM network appears now on the eastern coast of Spain, with a few close stations having high NSE values. For precipitation, the NSE values are lower, with NSE values below 0.5 in several stations, as precipitation is more difficult to model than temperature. The spatial pattern shows that the lowest values are related to the LR and MG networks.

Table 7 summarizes the number of stations falling below the Good" NSE threshold (NSE < 0.7) for the four variables studied. In the case of temperature-related variables, the percentage of stations below this threshold is low, with RFSI being the method with the lowest number of stations and IBERIA01 the grid with the worst results. TotP is the variable with the highest number of stations below this threshold, and RFSI standing out as the method with the lowest percentage of stations in this range. The grids that used a smaller data set for predictions are those with a larger number of stations below this threshold. This is the case for TERRACLIMATE and WORLDCLIM in TotP.

It is very interesting to observe such a high accuracy in RFSI given the type of cross-validation used, which probably penalises RFSI compared to the individual grids. In the LOOBUF-CV with a threshold distance, the model to predict at a given station is calibrated only with other stations located farther than the threshold. Individual grids such

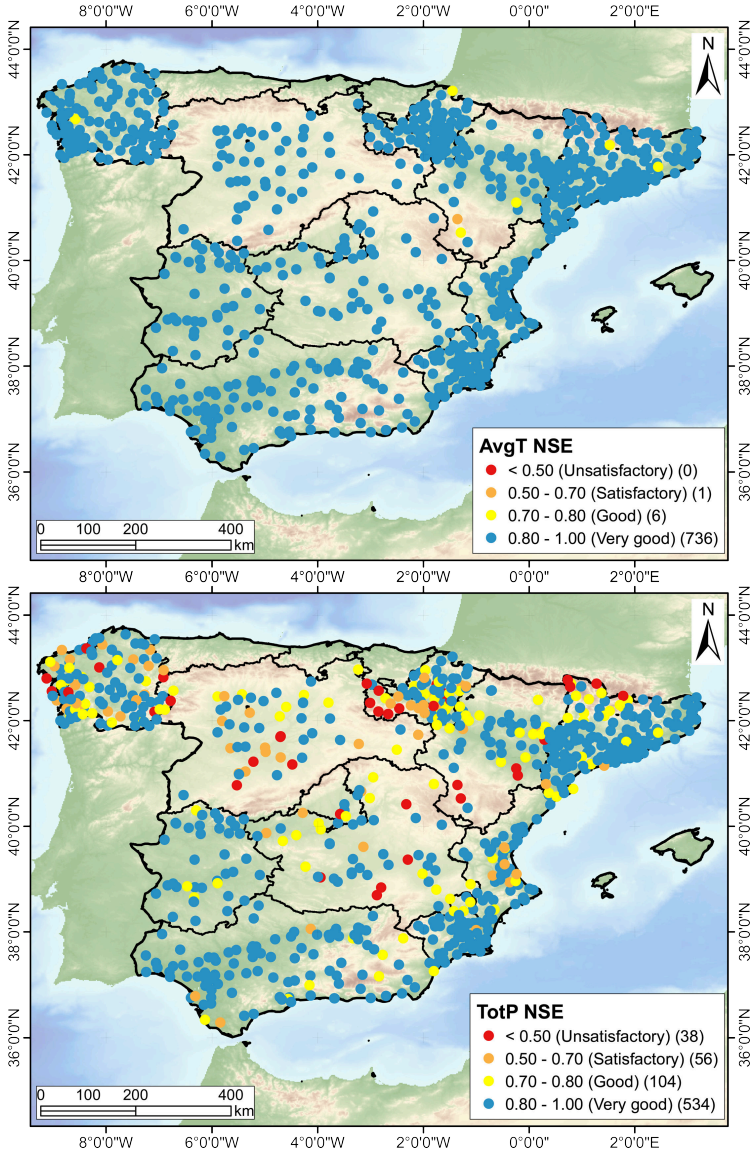

Figure 11: Spatial distribution of NSE in validation. (A) monthly mean temperature (743 stations) and (b) monthly total precipitation (732 stations). Classification in performance according to Table 2.

as STEAD, where more than 5000 stations were used, or SIMPA, which uses the network of SAIH stations, include all the stations in the generation of the network.

To complement the LOOBUF-CV goodness-of-fit analysis, an internal uncertainty analysis of RFSI was performed (Figure 13 for spatial distribution, Figure 14 to represent its temporal pattern, and Table 8 summarising its results). The analysis of the spatial distribution of uncertainty revealed a similar spatial distribution pattern for the temperature vari-

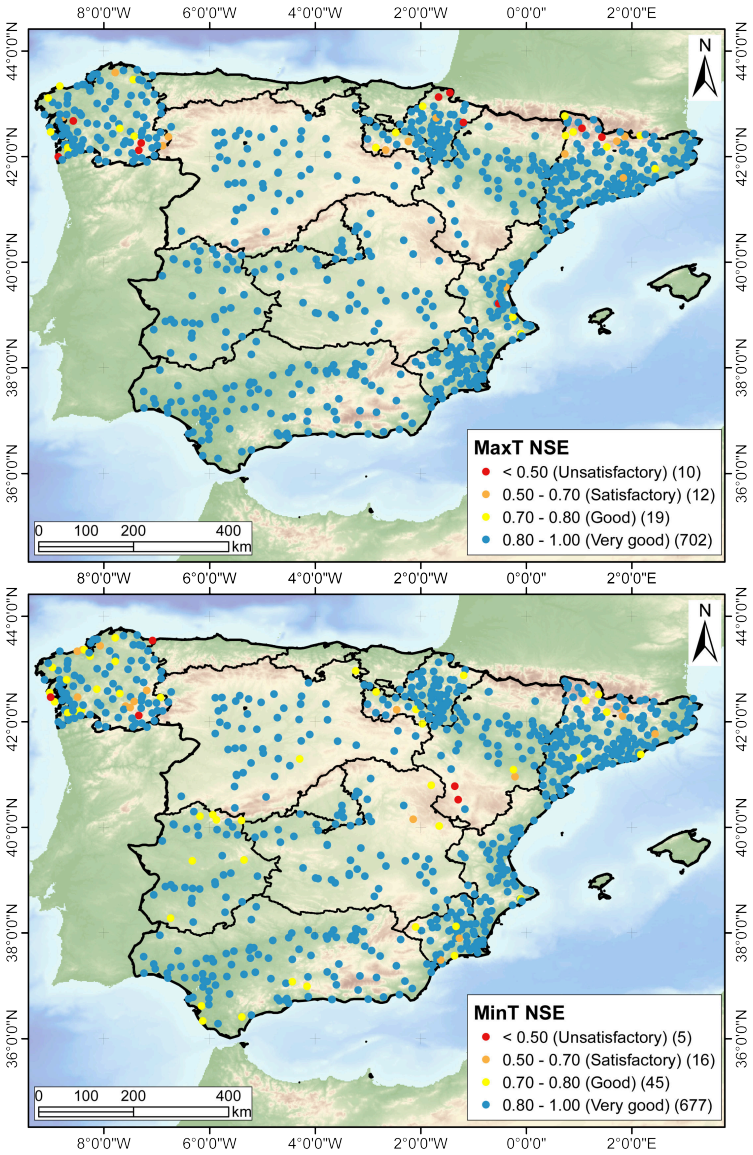

Figure 12: Spatial distribution of NSE in validation. (A) monthly maximum temperature (743 stations) and (b) monthly minimum temperature (732 stations). Classification in performance according to Table 2.

ables (Figure 13 A-C). These areas of high uncertainty are mainly located in high altitude and interior areas where thermal amplitude is higher. In any case, the standard deviation is not high, with average values below 1.5 ºC for the three variables (Table 8); furthermore, no trend is observed in the time series of average values in them (Figure 14 A-C).

With regard to precipitation (Figure 13 D), greater uncertainty is observed in the north of the Iberian Peninsula, with a distribution that practically coincides with the Eurosiberian

Model	AvgT	%	MaxT	%	MinT	%	TotP	%
AEMET	17	2.29	29	3.90	30	4.06	141	19.26
EOBSv27	30	4.04	40	5.38	50	6.77	286	39.07
IBERIA01	41	5.52	62	8.34	177	23.95	229	31.28
MOPREDAS	-	-	-	-	-	-	141	19.26
SIMPA	2	0.27	-	-	-	-	143	19.54
STEAD	11	1.48	22	2.96	25	3.38	-	-
TERRACLIMATE	20	2.69	35	4.71	109	14.75	476	65.03
WORLDCLIM	20	2.69	35	4.71	64	8.66	562	76.78
RFSI	1	0.13	22	2.96	21	2.84	94	12.84
SA	12	1.62	24	3.23	30	4.06	139	18.99
Total num. of stations	743	100	743	100	739	100	732	100

Table 7: Number of stations falling below the Good" NSE threshold (NSE < 0.7) ($NSE < 0.7$). LOOBUF-CV, excluding the AEMET network. The last row refers to the total number of stations used in each variable.

bioclimatic region (characterised by an oceanic climate with higher precipitation and no dry season), compared to the rest of the Iberian Peninsula, which belongs to the Mediterranean bioclimatic region. In this regard, [25,47] observed for this variable using RF, RFsp and RFSI that uncertainty is small for low precipitation amounts, whereas it is large in areas with a high precipitation amount. It is likely that this same behaviour is observed in the intra-annual pattern (Figure 14 D), where uncertainty tends to decrease in the summer months, characterised by a significant reduction in the amount of precipitation, especially in Mediterranean bioclimatic regions.

	AvgT (ºC/month)	MaxT (ºC/month)	MinT (ºC/month)	TotP (mm/month)
Min.	0.6408	0.8773	0.9492	6.8144
Avg.	1.0180	1.4226	1.3393	16.5004
Max.	2.2051	3.1438	2.2369	64.1180
Q25%	0.8918	1.2040	1.2310	11.4221
Q50%	0.9749	1.3403	1.3221	13.9368
Q75%	1.0924	1.5474	1.4272	18.5441
Q95%	1.3912	2.1271	1.6342	33.0430

Table 8: Summary statistics of RFSI internal uncertainty (SD obtained from the 1000 trees). The resulting values have been estimated as the average of the uncertainty associated with each month and year in each cell of the generated grid (available in the Available data section).

RFSI stands out as a method for generating grids using ensembles, but may also be of interest as a relatively simple method for filling in gaps in station data, as well that RF per-station (for point prediction at a single climate station). Figure 15 and Table 9 show two examples for the ensembles obtained from the LOOBUF-CV validation (RFSI does not use the station data to generate the estimate), SPLIT-V for RF per-station (RF is calibrated with even years only) and the two best individual grids. In both cases, The RF per-station offers the best NSE value, since it uses part of the observed data for its calibration. However, RFSI obtains the second best NSE value, followed by SIMPA. The example of Figure 15 (A), belonging to station WQ of the XEMA network, is very interesting: in this case

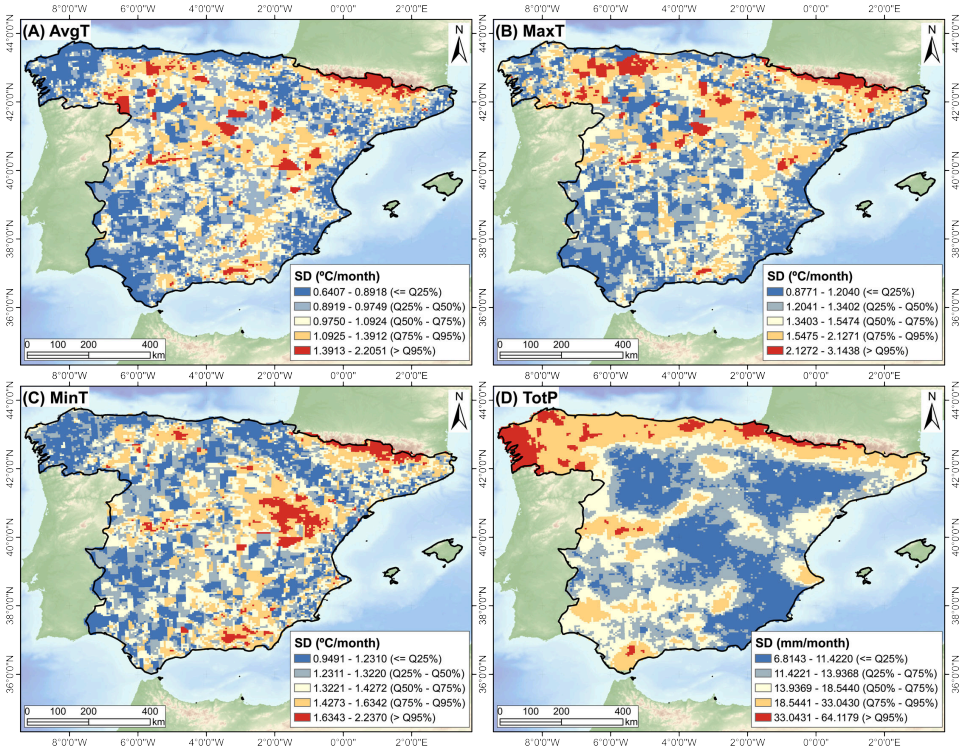

Figure 13: Internal uncertainty of RFSI (SD obtained from the 1000 trees) in the (A) AvgT, (B) MaxT, (C) MinT and (D) TotP models. The resulting values have been estimated as the average of the uncertainty associated with each month and year in each cell of the generated grid (available in the Available data section).

there are some months without information (from January to December 2006), which can be filled with the series generated in the ensemble by RFSI and RF per-station, highlighting again some interesting aspects such as a better fit to the data of months with maximum or minimum values of monthly mean temperature, emphasizing its high NSE value. The second of them (Figure 15 (B) represents the station NAV15 of the SIAR network, where it is noteworthy the ability of RFSI and RF per-station to reduce some errors in the maximum monthly precipitation values with respect to the individual grids.

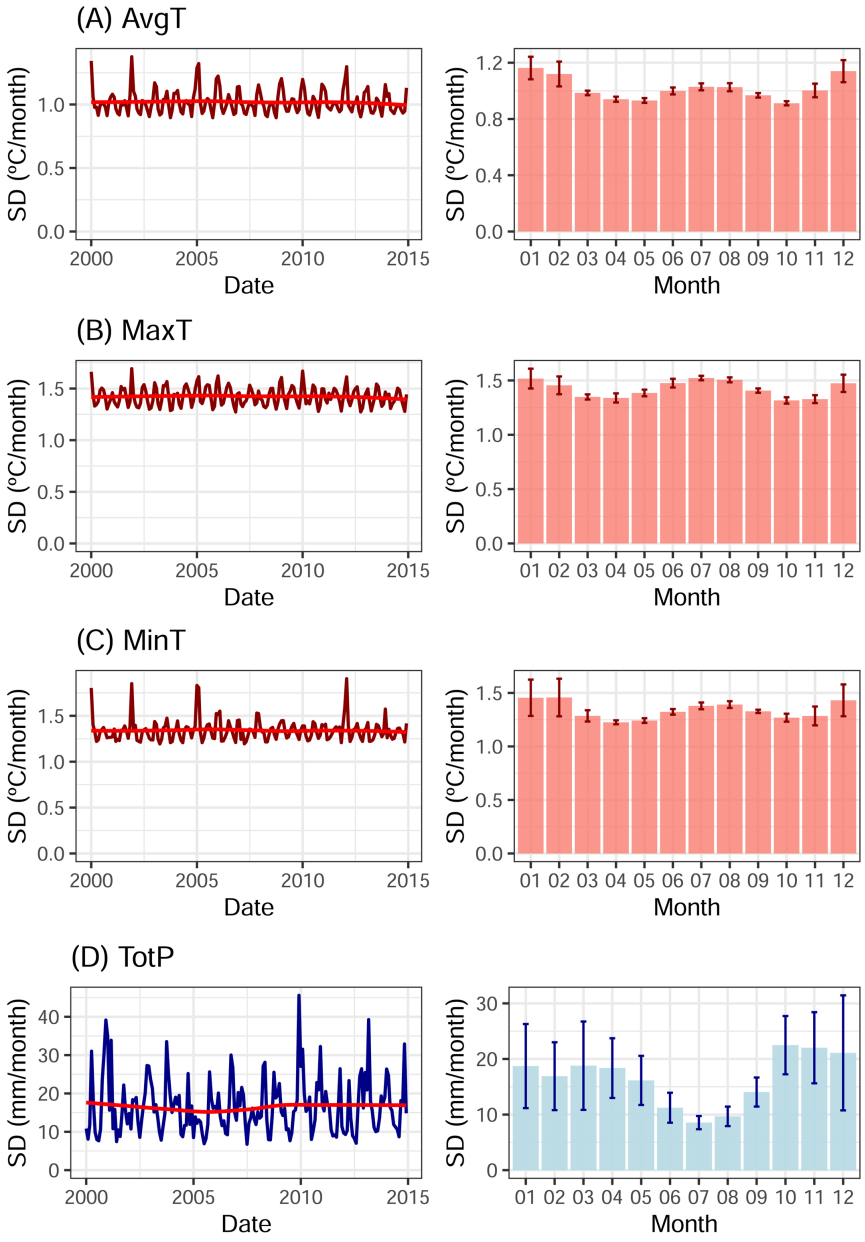

Figure 14: Average temporal distribution of RFSI internal uncertainty (SD obtained from the 1000 trees)) in the (A) AvgT, (B) MaxT, (C) MinT and (D) TotP models. The resulting values have been estimated as the average uncertainty of all grid cells (available in section Available data.

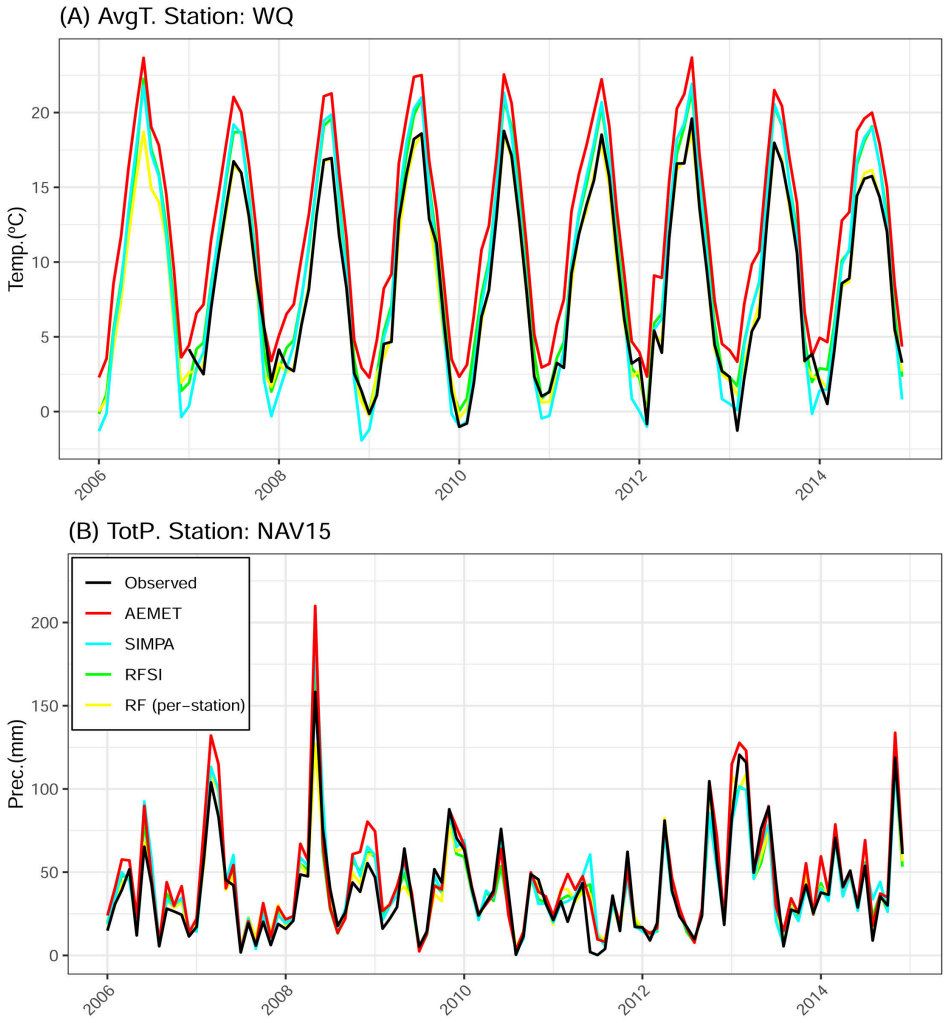

Figure 15: Example of two stations where the RFSI (green line) and RF per station (yellow) obtain a better fit than the individual members (AEMET and SIMPA) compared to the observed values (black line), and where their usefulness as a possible method for filling gaps is represented.

4 Conclusions

In this study, we present an agile methodological framework to evaluate the accuracy of gridded weather data in Spain and to obtain a more accurate result using an ensemble based on RF. In general, individual grids fit the observed data well, but with some issues caused by the way the interpolation process was carried out, such as the number of stations used and their distribution in space. For example, the two grids that seem to best fit the observed data are STEAD, a grid specifically created for the spatial distribution of tem-

[h!]

Model	Avg.T (est: WQ)	Tot.P (est: NAV15)
AEMET	0.6307	0.7936
IBERIA01	0.7532	0.4279
STEAD	0.7366	-
EOBSv27	0.6358	0.6072
TERRACLIMATE	0.7398	0.7339
WORLDCLIM	0.7353	0.7091
SIMPA	0.8999	0.8523
MOPREDAS	-	0.4712
SA	0.7630	0.7771
RFSI	0.9179	0.8824
RF (per station)	0.9861	0.9027

Table 9: NSE summary of series in Figure 15.

perature using data from more than 5000 stations, and SIMPA in the case of precipitation, which uses the SAIH network for its interpolation, with a spatial resolution of only 500 m.

However, the validation results of RF and RFSI are generally better because of the MME's ability to reduce the uncertainty of the grids used, the nature and differences in estimation by different methods, the use of covariates and different stations, as well as the RF ability to integrate different physical representations of the processes conveyed by the different grids. The results are particularly relevant when considering the validation method used for the spatial prediction, where the accuracy of the individual grids is likely to be overestimated compared to RFSI, which uses a buffer for the cross-validation that deselects stations more than 15 km away.

The framework presented is therefore an efficient alternative to improve individual grids, allowing also the incorporation of information from additional stations to those used in the generation of the individual grids (in a sense, it is a machine learning method that allows, in a simple way, the incorporation of new knowledge to that already present in the grids, derived from the point observations with which they have been calibrated). Furthermore, it can be used as a method to fill information gaps in climate series in a relatively simple way, and even to generate a reference series with which to check for possible errors (such as those related to drifts or outliers when using the series obtained with RFSI, where the data from the station itself are not used for its estimation). The usefulness of RF is therefore evident in both approaches tested: RF for the generation of time series from single station series or for the generation of spatio-temporal grids.

One limitation of this study is the use of the internal standard deviation of Random Forest as a proxy for uncertainty, rather than more computationally demanding methods such as quantile regression forests or residual kriging. However, previous studies have shown that this approach adequately captures model variability [48,49]. Future work could explore hybrid methods once computational limitations are overcome.

We therefore believe that the combined use of data from different grid projects can enrich scientific knowledge and produce data sets with lower uncertainty. In this regard, future lines of work may find it interesting to explore in greater depth how combining different grids with different resolutions—closely related to the COPS problem and the process of resolution homogenisation—affects the final adjustment of models and their predictions.

The disadvantage of this method is that predictions can only be made for the periods covered by the grids used as predictors. Therefore, it may be interesting to open a debate on

the problem of producing climate data grids in the framework of a research project without further support at the end of the project. The possible doubt as to whether the continuation of such projects to produce climate data grids could be considered a duplication of effort does not exist for other projects, such as those on global and/or regional climate change modelling. We believe that once the effort to start the project has been made, it should be continued because of the advantages of this type of grid products compared to traditional interpolation, such as the greater number of refined stations available to the projects, or the use of more complex and accurate interpolation methods performed by specialists with greater computing resources and access to input data, thus enriching scientific-technical knowledge.

5 Available data

Availability of grid datasets (RFSI predictions and internal uncertainty with a spatial resolution of 5 km) generated during the current study are available online at https://doi.org/10.6084/m9.figshare.30344053.

Acknowledgments

We thank the referees and editors for their constructive feedback regarding the initial version of the manuscript.

This work was supported by the Spanish Agencia Estatal de Investigación (Grant number TED2021-131131B-I00).

References

[1] ABATZOGLOU, J. T., DOBROWSKI, S. Z., PARKS, S. A., AND HEGEWISCH, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data 5*, 1 (2018), 170191. doi:10.1038/sdata.2017.191.

[2] AEMET. Valores climatológicos normales, 2025.

[3] AHMED, K., SACHINDRA, D. A., SHAHID, S., IQBAL, Z., NAWAZ, N., AND KHAN, N. Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. *Atmospheric Research 236* (2020), 104806. doi:10.1016/j.atmosres.2019.104806.

[4] ALEXANDERSON, H. A homogeneity test applied to precipitation data. *Journal of Climatology 6*, 6 (1986), 661–675. doi:10.1002/joc.3370060607.

[5] AMBLAR-FRANCÉS, M. P., RAMOS-CALZADO, P., SANCHIS-LLADÓ, J., HERNANZ-LÁZARO, A., PERAL-GARCÍA, M. C., NAVASCUÉS, B., DOMINGUEZ-ALONSO, M., PASTOR-SAAVEDRA, M. A., AND RODRÍGUEZ-CAMINO, E. High resolution climate change projections for the Pyrenees region. *Advances in Science and Research 17* (2020), 191–208. doi:10.5194/asr-17-191-2020.

[6] ARMSTRONG, J. S. *Combining Forecasts*. Springer US, Boston, MA, 2001, pp. 417–439. doi:10.1007/978-0-306-47630-3_19.

[7] BAÑO-MEDINA, J., MANZANAS, R., CIMADEVILLA, E., FERNÁNDEZ, J., GONZÁLEZ-ABAD, J., COFIÑO, A. S., AND GUTIÉRREZ, J. M. Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44. *Geoscientific Model Development 15*, 17 (2022), 6747–6758. doi:10.5194/gmd-15-6747-2022.

[8] BEGUERÍA, S., PEÑA-ANGULO, D., TRULLENQUE-BLANCO, V., AND GONZÁLEZ-HIDALGO, C. MOPREDAScentury: a long-term monthly precipitation grid for the Spanish mainland. *Earth System Science Data 15*, 6 (2023), 2547–2575. doi:10.5194/essd-15-2547-2023.

[9] BELGIU, M., AND DRˇ AGUT, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing 114* (2016), 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

[10] BENITO, B. M. *spatialRF: Easy Spatial Regression with Random Forest. R package version 1.1.3*, 2021. doi:10.5281/zenodo.4745208.

[11] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32. doi:10.1023/A:1010933404324.

[12] CAPEL MOLINA, J. J. Factores del clima de la península ibérica. *Paralelo 37*, 2 (1978), 5–13.

[13] CENTRO NACIONAL DE INFORMACIÓN GEOGRÁFICA, I. G. N., Ed. *España en mapas. Una síntesis geográfica. Serie Compendios del Atlas Nacional de España (ANE)* (2019). doi:10.7419/162.06.2018.

[14] CORNES, R. C., VAN DER SCHRIER, G., VAN DEN BESSELAAR, E. J. M., AND JONES, P. D. An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres 123*, 17 (2018), 9391–9409. doi:10.1029/2017JD028200.

[15] CRESSIE, N. Change of support and the modifiable areal unit problem.

[16] DEVINENI, N., SANKARASUBRAMANIAN, A., AND GHOSH, S. Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations. *Water Resources Research 44*, 9 (2008). doi:10.1029/2006WR005855.

[17] FICK, S. E., AND HIJMANS, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology 37*, 12 (2017), 4302–4315. doi:10.1002/joc.5086.

[18] GELFAND, A. E. On the change of support problem for spatio-temporal data. *Biostatistics 2*, 1 (mar 2001), 31–45. doi:10.1093/biostatistics/2.1.31.

[19] GOMARIZ CASTILLO, F., AND ALONSO SARRÍA, F. Effect of watershed subdivision and estimation of climatic variables on hydrological simulation with the swat model in semi-arid mediterranean basins. *Papeles de Geografía 64*, 64 (2018), 114–133. doi:10.6018/geografia/2018/331531.

[20] GOMARIZ-CASTILLO, F., ALONSO-SARRÍA, F., AND CÁNOVAS-GARCÍA, F. Improving Classification Accuracy of Multi-Temporal Landsat Images by Assessing the Use of Different Algorithms, Textural and Ancillary Information for a Mediterranean Semiarid Area from 2000 to 2015. *Remote Sensing 9*, 10 (2017), 1058. doi:10.3390/rs9101058.

[21] GUIJARRO, J. A. *Homogenization of climatic series with Climatol. Version 3.1.1*. State Meteorological Agency (AEMET), Balearic Islands Office, Spain, Madrid, Spain, 2019.

[22] GUIJARRO, J. A. *climatol: Climate Tools (Series Homogenization and Derived Products). R package version 4.3.1.*, 2023.

[23] HAGEDORN, R., DOBLAS-REYES, F. J., AND PALMER, T. The rationale behind the success of multi-model ensembles in seasonal forecasting — I. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography 57*, 3 (2005), 219–233. doi:10.3402/tellusa.v57i3.14657.

[24] HARRIS, I., OSBORN, T. J., JONES, P., AND LISTER, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data 7*, 1 (2020), 109. doi:10.1038/s41597-020-0453-3.

[25] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning*, second edi ed. Springer Series in Statistics. Springer New York, New York, NY, 2009. doi:10.1007/978-0-387-84858-7.

[26] HENGL, T., NUSSBAUM, M., WRIGHT, M. N., HEUVELINK, G. B., AND GRÄLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ 6* (2018), e5518. doi:10.7717/peerj.5518.

[27] HERRERA, S., CARDOSO, R. M., SOARES, P. M., ESPÍRITO-SANTO, F., VITERBO, P., AND GUTIÉRREZ, J. M. Iberia01: a new gridded dataset of daily precipitation and temperatures over Iberia. *Earth System Science Data 11*, 4 (2019), 1947–1956. doi:10.5194/essd-11-1947-2019.

[28] IPCC. *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, U.K., 2016.

[29] KIM, Y.-T., YU, J.-U., KIM, T.-W., AND KWON, H.-H. A novel approach to a multi-model ensemble for climate change models: Perspectives on the representation of natural variability and historical and future climate. *Weather and Climate Extremes 44* (2024), 100688. doi:10.1016/j.wace.2024.100688.

[30] KUMAR, A., MITRA, A. K., BOHRA, A. K., IYENGAR, G. R., AND DURAI, V. R. Multi-model ensemble (MME) prediction of rainfall using neural networks during monsoon season in India. *Meteorological Applications 19*, 2 (2012), 161–169. doi:10.1002/met.254.

[31] LONG, J. S., AND ERVIN, L. H. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician 54*, 3 (2000), 217–224. doi:10.1080/00031305.2000.10474549.

[32] MANKIN, K. R., MEHAN, S., GREEN, T. R., AND BARNARD, D. M. Review of gridded climate products and their use in hydrological analyses reveals overlaps, gaps, and the need for a more objective approach to selecting model forcing datasets. *Hydrology and Earth System Sciences 29*, 1 (2025), 85–108. doi:10.5194/hess-29-85-2025.

[33] MERINO, A., LÓPEZ, L., HERMIDA, L., SÁNCHEZ, J. L., GARCÍA-ORTEGA, E., GASCÓN, E., AND FERNÁNDEZ-GONZÁLEZ, S. Identification of drought phases in a 110-year record from western mediterranean basin: Trends, anomalies and periodicity analysis for iberian peninsula. *Global and Planetary Change 133* (2015), 96–108. doi:10.1016/j.gloplacha.2015.08.007.

[34] MEYER, H., AND PEBESMA, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution 12*, 9 (2021), 1620–1633. doi:10.1111/2041-210X.13650.

[35] MIN, S.-K., AND HENSE, A. A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophysical Research Letters 33*, 8 (2006). doi:10.1029/2006GL025779.

[36] MITECO. Modelo SIMPA. Periodo de simulación: 1940/41 a 2017/18, 2024.

[37] MORIASI, D. N., GITAU, M. W., PAI, N., AND DAGGUPATI, P. Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE 58*, 6 (2015), 885–900. doi:10.13031/trans.58.10715.

[38] O'REILLY, C. H., BEFORT, D. J., AND WEISHEIMER, A. Calibrating large-ensemble European climate projections using observational data. *Earth System Dynamics 11*, 4 (2020), 1033–1049. doi:10.5194/esd-11-1033-2020.

[39] OVIEDO TORRES, B. E., AND LEÓN ARISTIZÁBAL, G. *Guía de procedimiento para la generación de escenarios de cambio climático regional y local a partir de los modelos globales*. Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá, 2010.

[40] PENG, S., DING, Y., LIU, W., AND LI, Z. 1 km monthly temperature and precipitation dataset for China from 1901 to 2017. *Earth System Science Data 11*, 4 (2019), 1931–1946. doi:10.5194/essd-11-1931-2019.

[41] PLOTON, P., MORTIER, F., RÉJOU-MÉCHAIN, M., BARBIER, N., PICARD, N., ROSSI, V., DORMANN, C., CORNU, G., VIENNOIS, G., BAYOL, N., LYAPUSTIN, A., GOURLET-FLEURY, S., AND PÉLISSIER, R. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications 11*, 1 (2020), 4540. doi:10.1038/s41467-020-18321-y.

[42] PROBST, P., AND BOULESTEIX, A. L. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research 18* (2018), 1–8.

[43] QGIS DEVELOPMENT TEAM. *QGIS Geographic Information System*. QGIS Association, 2024.

[44] R CORE TEAM. R: A Language and Environment for Statistical Computing, 2024.

[45] ROUGIER, J. Ensemble Averaging and Mean Squared Error. *Journal of Climate 29*, 24 (2016), 8865–8870. doi:10.1175/JCLI-D-16-0012.1.

[46] RUIZ-AÍLVAREZ, M., GOMARIZ-CASTILLO, F., AND ALONSO-SARRÍA, F. Evapotranspiration response to climate change in semi-arid areas: Using random forest as multimodel ensemble method. *Water 13*, 2 (2021). doi:10.3390/w13020222.

[47] SCHUMACHER, V., JUSTINO, F., FERNÁNDEZ, A., MESEGUER-RUIZ, O., SARRICOLEA, P., COMIN, A., PERONI VENANCIO, L., AND ALTHOFF, D. Comparison between observations and gridded data sets over complex terrain in the Chilean Andes: Precipitation and temperature. *International Journal of Climatology 40*, 12 (2020), 5266–5288. doi:10.1002/joc.6518.

[48] SEKULIĆ, A., KILIBARDA, M., HEUVELINK, G. B., NIKOLIĆ, M., AND BAJAT, B. Random Forest Spatial Interpolation. *Remote Sensing 12*, 10 (2020), 1687. doi:10.3390/rs12101687.

[49] SEKULIĆ, A., KILIBARDA, M., PROTIĆ, D., AND BAJAT, B. A high-resolution daily gridded meteorological dataset for Serbia made by Random Forest Spatial Interpolation. *Scientific Data 8* (2020), 123. doi:10.1038/s41597-021-00901-2.

[50] SERRANO-NOTIVOLI, R., BEGUERÍA, S., AND DE LUIS, M. STEAD: a high-resolution daily gridded temperature dataset for Spain. *Earth System Science Data 11*, 3 (2019), 1171–1188. doi:10.5194/essd-11-1171-2019.

[51] SERRANO-NOTIVOLI, R., BEGUERÍA, S., SAZ, M. Á., LONGARES, L. A., AND 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*, 2 (2017), 721–738. doi:10.5194/essd-9-721-2017.

[52] TADESSE, K. E., MELESSE, A. M., ABEBE, A., LAKEW, H. B., AND PARON, P. Evaluation of Global Precipitation Products over Wabi Shebelle River Basin, Ethiopia. *Hydrology 9*, 5 (2022), 66. doi:10.3390/hydrology9050066.

[53] THENKABAIL, P., Ed. *Remote Sensing of Water Resources, Disasters, and Urban Studies*. CRC Press, oct 2015. doi:10.1201/b19321.

[54] VALAVI, R., ELITH, J., LAHOZ-MONFORT, J. J., AND GUILLERA-ARROITA, G. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution 10*, 2 (2019), 225–232. doi:10.1111/2041-210X.13107.

[55] VELÁZQUEZ-ZAPATA, J. A. Comparing Meteorological Data Sets in the Evaluation of Climate Change Impact on Hydrological Indicators: A Case Study on a Mexican Basin. *Water 11*, 10 (2019), 2110. doi:10.3390/w11102110.

[56] WANG, T., HAMANN, A., SPITTLEHOUSE, D. L., AND MURDOCK, T. Q. ClimateWNA—High-Resolution Spatial Climate Data for Western North America. *Journal of Applied Meteorology and Climatology 51*, 1 (2012), 16–29. doi:10.1175/JAMC-D-11-043.1.

[57] WRIGHT, M. N., AND ZIEGLER, A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software 77*, 1 (2017), 1–17. doi:10.18637/jss.v077.i01.

[58] YAZDANDOOST, F., MORADIAN, S., IZADI, A., AND BAVANI, A. M. A framework for developing a spatial high-resolution daily precipitation dataset over a data-sparse region. *Heliyon 6*, 9 (2020), e05091. doi:10.1016/j.heliyon.2020.e05091.