Geosci. Model Dev., 16, 4083–4112, 2023
https://doi.org/10.5194/gmd-16-4083-2023

Geoscientific Model Development
Open Access
EGU

Modeling river water temperature with limiting forcing data: Air2stream v1.0.0, machine learning and multiple regression

Manuel C. Almeida and Pedro S. Coelho

MARE – Marine and Environmental Sciences Centre, ARNET – Aquatic Research Network Associate Laboratory, NOVA School of Science and Technology, NOVA University Lisbon, Caparica, Portugal

Correspondence: Manuel Almeida (mcvta@fct.unl.pt)

Received: 18 August 2022 – Discussion started: 21 November 2022
Revised: 18 May 2023 – Accepted: 25 June 2023 – Published: 20 July 2023

Abstract. The prediction of river water temperature is of key importance in the field of environmental science. Water temperature datasets for low-order rivers are often in short supply, leaving environmental modelers with the challenge of extracting as much information as possible from existing datasets. Therefore, identifying a suitable modeling solution for the prediction of river water temperature with a large scarcity of forcing datasets is of great importance. In this study, five models, forced with the meteorological datasets obtained from the fifth-generation atmospheric reanalysis, ERA5-Land, are used to predict the water temperature of 83 rivers (with 98 % missing data): three machine learning algorithms (random forest, artificial neural network and support vector regression), the hybrid Air2stream model with all available parameterizations and a multiple regression. The machine learning hyperparameters were optimized with a tree-structured Parzen estimator, and an oversampling–undersampling technique was used to generate synthetic training datasets. In general terms, the results of the study demonstrate the vital importance of hyperparameter optimization and suggest that, from a practical modeling perspective, when the number of predictor variables and observed river water temperature values are limited, the application of all the models considered in this study is crucial. Basically, all the models tested proved to be the best for at least one station. The root mean square error (RMSE) and the Nash–Sutcliffe efficiency (NSE) values obtained for the ensemble of all model results were 2.75 ± 1.00 and 0.56 ± 0.48 °C, respectively. The model that performed the best overall was random forest (annual mean – RMSE: 3.18 ± 1.06 °C; NSE: 0.52 ± 0.23). With the application of the oversampling–undersampling technique, the RMSE values obtained with the random forest model were reduced from 0.00 % to 21.89 % ($\mu = 8.57$ %; $\sigma = 8.21$ %) and the NSE values increased from 1.1 % to 217.0 % ($\mu = 40$ %; $\sigma = 63$ %). These results suggest that the solution proposed has the potential to significantly improve the modeling of water temperature in rivers with machine learning methods, as well as providing increased scope for its application to larger training datasets and the prediction of other types of dependent variables. The results also revealed the existence of a logarithmic correlation among the RMSE between the observed and predicted river water temperature and the watershed time of concentration. The RMSE increases by an average of 0.1 °C with a 1 h increase in the watershed time of concentration (watershed area: $\mu = 106$ km^2; $\sigma = 153$).

1 Introduction

Water temperature (WT) is recognized as a key parameter in aquatic systems due to its influence on water quality (e.g., chemical reaction rate, oxygen solubility), as well as the distribution and growth rate of aquatic organisms (e.g., primary production; fish growth and habitat) (Smith, 1972; Webb et al., 2003; Caissie, 2006). As such, the accurate prediction and assessment of river WT are crucial parts of many Earth science applications. The thermal dynamics in rivers are quite complex as they depend on an array of physical and chemical factors (Smith and Lavis, 1975; Jeppesen and Iversen, 1987). River WT follows a seasonal and a diurnal cycle, driven by heat input and losses at the boundary conditions of a river section (upstream and downstream transfer; air–water and sediment–water interface; lateral contribution from trib-

utaries and groundwater) under specific meteorological and hydrological conditions (Walling and Webb, 1993; Wetzel, 2001). The complexity of river WT estimation is often more pronounced for sub-daily temporal and spatial scales (Toffolon and Piccolroaz, 2015), and it is therefore common practice to average out sub-daily effects and to consider a daily discretization for modeling purposes. This assumption can have a significant impact on lake and reservoir water quality modeling results, namely when lake and/or reservoir inflows are large. The fall and spring turnover onset, stratification strength and length, and the overall heat budget can be affected; therefore, some caution is needed regarding this type of approach. Air temperature correlates with the equilibrium temperature of a river and is therefore frequently used as the independent variable; hence, it is not unusual to find a strong linear correlation between daily air temperature and stream and river WT with a time lag (Smith, 1981; Crisp and Howson, 1982). The existing body of literature includes a number of examples of the successful implementation of linear regression models correlating air and WT using data relating to different time periods, mostly weekly and/or monthly, as the serial dependency for these timescales is generally small (e.g., Mackey and Berrie, 1991; Webb and Nobilis, 1997). That said, several studies have shown departures from linearity, showing that the rate of evaporative cooling increases at peak air temperatures, which means that the river WT will therefore not increase linearly with the mean air temperature (Mohseni et al., 1998, 2002), thereby demonstrating the need for more complex models and sampling of an increased number of independent variables. There are many sources of error in the modeling of river WT, including those associated with the definition of the input data and boundary conditions or with the river WT measurements used in model calibration or related to the model's structure. The predictor variables can represent a significant source of uncertainty, as river WT is affected not only by local environmental conditions, but also by upstream conditions (Moore et al., 2005). In order to minimize this source of uncertainty, some authors use a space-averaging approach in which the predictor variables consider a variety of buffer zones with different lengths and widths (e.g., Macedo et al., 2013; Segura et al., 2014). However, the extent of the area affecting the river energy balance at a certain point is still unclear (Moore et al., 2005; Gallice et al., 2015).

In the past decades, different types of models have been successfully used to model river WT under different spatial and temporal scales. In general, the model selection depends not only on the study's requirements, namely the output timescale, but also on the availability of the input data. These include statistical models, such as linear regression (e.g., Neumann et al., 2003; Rehana and Mujumdar, 2011), multiple regression (e.g., Jeppesen and Iversen, 1987; Jourdonnais et al., 1992), nonlinear regression (e.g., Mohseni et al., 1998) and stochastic regression models (e.g., Ahmadi-Nedushan et al., 2007; Rabi et al., 2015) as well as hybrid models (statistics methods combined with a physically based process, e.g., Gallice et al., 2015; Toffolon and Piccolroaz, 2015). Process-based models, based on the concepts of heat advection, transportation and equilibrium temperature, are quite accurate when the boundary conditions are well characterized (e.g., Sinokrot and Stefan, 1993; Younus et al., 2000; Du et al., 2018), although they do require a large amount of forcing data, including stream geometry, air temperature, dew point temperature (or relative humidity), cloud cover and shortwave solar radiation, degree of shading, and wind direction and velocity. Machine learning (ML) models, such as artificial neural networks (ANNs), have also proved to be a robust option for river WT prediction (e.g., Piotrowski et al., 2015; Temizyurek and Dadaser-Celik, 2018; Zhu et al., 2019c). In general, results show the performance of ML models to be comparable (Feigl et al., 2021; Zhu et al., 2018). Multi-layer perception neural network models are, in most cases, not outperformed by more complex and advanced neural network models (Piotrowski et al., 2015; Zhu et al., 2019b). ML outperformed standard modeling approaches, such as multiple regression, the hybrid Air2stream model developed by Toffolon and Piccolroaz (2015) (Feigl et al., 2021), linear regression, nonlinear regression, and stochastic models (Zhu et al., 2018). This is not a prevailing rule as the Air2stream model was also able to outperform ML, clearly indicating its potential as a valid solution in certain conditions (Zhu et al., 2019d). Table 1 describes the RMSE between observed and predicted river WT obtained from several studies and using different models. Overall, the results are quite impressive, varying from 0.42 to 2.30 °C in the case of the ML models. The worst results, as expected, correspond to the classical statistical models, namely multiple regression. Oversampling–undersampling techniques are useful where regression is applicable, but the values of interest are rare or uncommon, producing an imbalanced dataset. Several available strategies exist, such as random undersampling (Torgo et al., 2015), the Synthetic Minority Oversampling Technique for Regression (SMOTER) (Torgo et al., 2013) and the introduction of Gaussian noise (Branco et al., 2016). The the Synthetic Minority Oversampling Technique for Regression with Gaussian Noise (SMOGN) Python package combines random undersampling with the two previously mentioned oversampling techniques (SMOTER and the introduction of Gaussian noise) as a function of K-nearest-neighbor (KNN) distances underlying an observation. SMOGN was successfully implemented by Wang et al. (2021) to improve the quantification of nonlinear relationships between monthly burned area and biophysical factors in southeastern Australian forests. SMOGN was applied to resample the proportion of burned area. This algorithm was also successfully applied by Agrawal and Petersen (2021) to increase a satellite imagery dataset required to identify arsenic contamination and increase the performance of ML algorithms. Although this type of solution is still not widely used, it should be considered as it has the potential to improve ML performance,

particularly in cases in which the forcing datasets are small and inconsistent. From an environmental science perspective, accurate time-varying boundary conditions are vital in order to calibrate models or evaluate system evolution. For WT calibration, this ideally means using continuous inflow temperatures, although this is complicated by the fact that WT measurements are often in short supply or completely unavailable, particularly for low-order streams. Therefore, the main objective of this study is to identify a suitable WT modeling solution for rivers with limiting forcing data. Improving this type of solution would deliver potential benefits for a wide range of environmental modeling applications, such as the analysis of seasonal and diurnal trends as well as biogeochemical processes in rivers based on observation datasets and the improvement of lake and reservoir water quality model boundary conditions.

It is also important to note that the studies defined to evaluate the performance of different modeling approaches are normally restricted to a very small number of test sites and usually contain a reasonable amount of forcing data (Table 1) – hence, the vital importance of increasing the number of test sites and using a limited amount of forcing data to model river WT. The methodological approach was therefore defined to attempt to answer the following questions.

1. What is the best modeling solution to predict river WT with limited forcing data?

2. How do the length of the calibration period and percentage of missing data affect model performance?

3. Can the performance of an ML model be improved through the modification of the raw training dataset with an oversampling–undersampling technique?

4. Is it possible to relate the modeling error to river and watershed geomorphological and hydrological variables (e.g., time of concentration; wet and dry season)?

To that end, 83 river sections with different geomorphological, meteorological and hydrological conditions were modeled. These stations correspond to all the sections for which the Portuguese Water Resources Information System (SNIRH) holds WT and discharge datasets, which are also, coincidentally, characterized by 98 % missing data. The modeling ensemble includes five different models, three of which use ML algorithms optimized with a sequential model-based optimization approach: random forest (RF), artificial neural network (ANN) and support vector regression (SVR). The remaining models include the hybrid Air2stream model (using all model parametrization variations: three, four, five, seven and eight parameters) (Toffolon and Piccolroaz, 2015) and multiple regression (MR). The SMOGN algorithm was also used to generate 100 synthetic samples from raw training datasets. These modified datasets were then considered to force the best model.

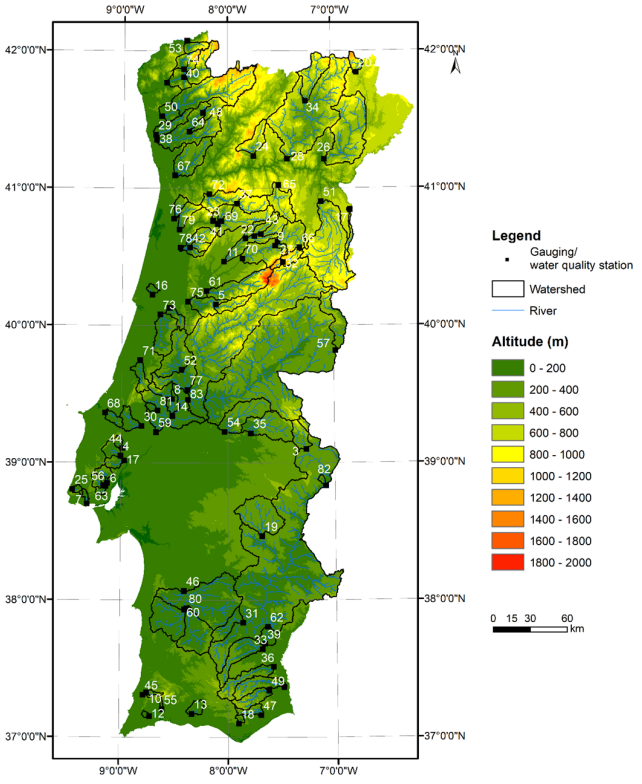

Figure 1. Location of the watersheds considered in the study (from a DEM, Shuttle Radar Topography Mission; Farr et al., 2007.)

The results of this study will hopefully prove useful from a practical perspective by helping to improve the quality and consistency of river WT datasets.

2 Study area and data

The watersheds considered in this study are located in Portugal (Fig. 1). This southern European country has a typical Mediterranean climate. Maximum daily mean air temperature ranges from 13 °C in the central highlands to 25 °C in the southeastern region. The minimum daily mean air temperature ranges from 5 °C in the northern and central regions to 18 °C in the south (Soares et al., 2012). The spatial and temporal heterogeneity of precipitation, which differs from a relatively wet annual maximum of over 2500 mm yr^{-1} in the mountainous northwest to a much drier 400 mm yr^{-1} in the flat southeast, is defined by complex topography and coastal processes (Cardoso et al., 2013; Soares et al., 2015).

The models used in this study were forced with daily mean, maximum and minimum air temperature and global radiation values obtained from the fifth-generation atmospheric reanalysis, ERA5-Land, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5-Land is the ECMWF's most advanced reanalysis dataset for land applications (Muñoz-Sabater, 2019, 2021).

Table 1. List of reviewed publications on river WT modeling and the corresponding RMSE between observed and modeled WT values.

Reference	Geographic location	Number of sites	Temporal scale	Model type	RMSE (°C)
Chenard and Caissie (2008)	Canada	1	day	ANN	0.96
DeWeber and Wagner (2014)	Eastern US	96	day	ANN	1.82; 1.93
Rabi et al. (2015)	Croatia	3	day	ANN	$\mu = 1.70$ $\sigma = 0.49$; $\mu = 2.06$ σ =0.35; $\mu = 2.30$ $\sigma = 0.76$
Zhu et al. (2019c)	US	3	day	ANN	0.768; 0.948; 1.242
Feigl et al. (2021)	Austria, Germany, Switzerland	10	day	ANN	Best results: 0.45; 0.42; 0.43
Zhu et al. (2019a)	Croatia	2	day	ANN	1.35; 1.70
Zhu et al. (2019d)	Europe, US	8	day	ANN	[0.46,1.69]
Rehana (2019)	India	1	day	SVR	1.69
Rajesh and Rehana (2021)	India	1	day	SVR	0.99
Lu and Ma (2020)	US	1	hour	RF	1.04
Feigl et al. (2021)	Austria, Germany, Switzerland	10	day	RF	0.58
Rajesh and Rehana (2021)	India	1	day	RF	1.03
Rehana (2019)	India	1	day	MR	1.85
Moore et al. (2013)	Western Canada	418	year	MR	2.1
Ducharne (2008)	France	88	month	MR	[1.4,1.9]
Zhu et al. (2019a)	Croatia	2	day	MR	2.33; 2.74
Toffolon and Piccolroaz (2015)	Switzerland	3	day	Air2stream	3 par [0.88, 1.05]; 4 par [0.87,1.04] 5 par [0.70, 1.05]; 7 par [0.65,0.78]; 8 par [0.75,0.62]*
Zhu et al. (2019d)	Europe, US	8	day	Air2stream	3 par [0.64, 1.25]; 5 par [1.31, 0.76]; 8 par [1.37;0.93]*
Feigl et al. (2021)	Austria Germany, Switzerland	10	day	Air2stream	8 par [0.74,1.17]*

* The model can be applied with three, four, five, seven or eight parameters (3 par, 4 par, 5 par, 7 par or 8 par).

The horizontal resolution of this dataset ($0.1° \times 0.1°$; native resolution is 9 km) is higher than that corresponding to ERA-Interim and ERA5 ($0.28° \times 0.28°$; native resolution 31 km grid). The vertical coverage ranges from 2 m above surface level to a soil depth of 289 cm. The Carbon Hydrology-Tiled ECMWF Scheme for Surface Exchanges Over Land (CHTESSEL) forced with atmospheric forcing derived from ERA5 near-surface meteorology state and flux fields (10 m above ground level) is central to ERA5-Land. The surface fluxes are linearly interpolated from the ERA5 resolution of approximately 31 km to the ERA5-Land resolution of 9 km. Land characteristics, such as soil and vegetation type and vegetation cover, are described by time-invariant fields (Muñoz-Sabater et al., 2021a). The air temperature reanalysis dataset (hourly data) covering a period of 42 years (1 January 1980 to 31 December 2021) was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store (Muñoz-Sabater, 2019, 2021). The watershed discharge data used to force the models and the WT considered for the model's validation are also available from SNIRH (http://snirh.apambiente.pt, last access: 17 July 2023). The SNIRH provides data and WT values for 2471 water quality stations, only 98 of which have gauging stations with discharge values, one of the conditions required to implement the Air2stream model. The missing discharge data were replaced with the corresponding climatological year value; hence, only the gauging stations with data spanning at least a full year (365 or 366 values) were kept. Following this initial analysis, the number of stations considered was reduced to 83. Data availability varies from station to station but generally covers a period of 42 years (1980–2021). However, a significant number of daily river WT values are missing, ranging from 96.9 % to 99.9 % ($\mu = 98.8$ %; $\sigma = 0.68$).

Table 2 shows the number of WT values for the annual data series for the dry season (April to September) and for the wet season (October to March) separated into training and test datasets, considering all stations.

3 Methodology

The definition of the methodological approach was supported by the following.

1. It is important to model a significant number of watersheds to reduce the degree of uncertainty in the results. This was minimized by modeling all the watersheds located in Portugal for which river WT and discharge values were available.

2. The number and type of models are also key to gaining a comprehensive understanding of the structural differences between the models and their performance. The five models considered in this study include state-of-the-art algorithms, with one classic modeling approach (MR) included to establish a benchmark.

3. Generally speaking, there are no available sources of observed meteorological data for either the watershed or the area surrounding the lowest part of low-order rivers;

Table 2. WT for the annual, dry and wet season, training, and test data series.

Temporal scale	Phase	Total number	Mean	Standard deviation	Maximum	Minimum
Annual	Train	8384	101	60	237	11
Annual	Test	3593	43	26	102	5
Dry season	Train	4161	50	32	116	4
Dry season	Test	1783	21	14	50	2
Wet season	Train	4223	51	29	124	4
Wet season	Test	1810	22	13	53	2

as such, the forcing meteorological datasets considered in this study were obtained from the ERA5-Land reanalysis.

The modeling reference is the watershed main gauging station or water quality station. Therefore, the hourly air temperature (°C) and global radiation (shortwave) ($J\,m^{-2}$) input datasets of the nearest ERA5-Land grid point were initially downloaded before the air temperature datasets were corrected according to the gauging station and the ERA5-Land grid point altitude. This correction was achieved by considering a linear variation of air temperature with altitude: $\frac{dT}{dz} = -6.0\,°C\,km^{-1}$ (Fahrer and Harris, 2004). After this correction, the mean, maximum and minimum daily air temperature values and the mean global radiation values were computed from the hourly meteorological datasets. Initially the model predictors were selected on the basis of their availability and the results obtained with other studies (e.g., Zhu et al., 2019c; Feigl et al., 2021). These included mean, maximum and minimum daily air temperature (°C); mean daily total radiation (shortwave) ($J\,m^{-2}$); discharge ($m^3\,s^{-1}$); and two temporal features, the month (0–12) and the day (1–365) of the year (MOY and DOY, respectively) (Table 3).

The Results section starts with the evaluation of the ERA5-Land mean daily air temperature datasets. These datasets were compared with ground measurements of mean daily air temperature considering all the meteorological datasets located within a 5 km radius of the stations considered in this study. Following this initial analysis, the models (see Sect. 3.1 to 3.6) were applied to each of the 83 input datasets, divided between a training (70 % of the entire dataset) and testing dataset (the remaining 30 %). The validation phase was not considered due to the size of the available datasets. It should be noted that, in the case of the Air2stream model, 70 % of the initial dataset corresponds to the calibration dataset and the remaining 30 % to the validation dataset. Hyperparameter optimization was achieved for the ML models through the application of the Tree-structured Parzen Estimator (TPE) algorithm (see Sect. 3.6). Given the large number of input datasets and the fact that the optimization process can be very time-consuming, the following approach was implemented (Fig. 2).

1. The 83 stations were ordered as a function of the number of samples (lowest to highest) and were divided into four different classes ($L \leq 50$, $50 < L \leq 100$, $100 < L \leq 200$, $L > 200$). Three stations were selected within each class: (1) the station with the fewest samples, (2) the station with the most samples and (3) the station with the number of samples that most closely corresponded to the average sample number for each class. The 12 datasets selected corresponded to stations 1, 7, 12, 13, 22, 29, 30, 46, 59, 60, 73 and 83.

2. The ML and TPE algorithms were applied to the 12 datasets. At this stage there were 12 optimized model structures computed with the TPE algorithm for each ML model.

3. The 12 optimized models obtained for each ML were subsequently applied to the 83 datasets, and the best-performing model at each station was calculated on the basis of the computed root mean square error (RMSE). Hence, the ensemble of the best results obtained across the 12 different models for the 83 stations defines the overall ML results.

To evaluate the possibility of further improving the results obtained with the best model, 100 different training datasets were then derived for each of the 12 datasets selected in step (1) through the application of the Synthetic Minority Oversampling Technique for Regression with Gaussian Noise (SMOGN) (Branco et al., 2017) (Sect. 3.7). The five SMOGN parameters that drive the algorithm were randomly derived within each model run considering a predefined parameter space (Table A2). A description of the model parameters is included in Table A2. The best ML model obtained in step (3) was then forced with the modified training datasets (100 for each station) and optimized with TPE.

Following this analysis, and in order to further investigate the relevance of the predictor variables, the input feature importance was estimated for all stations by considering the best-performing model. Additionally, the best model was used to evaluate the differences between observed and model river WT considering the sequential increase in the models' predictors: (1) mean air temperature, (2) mean air temperature + discharge, (3) mean air temperature + discharge + radiation, (4) mean air temperature + discharge + radiation + maximum air temperature, (5) mean air temperature + discharge + radiation + maximum air temperature + minimum

Table 3. Model predictor variables.

Model	Predictor variables	Output variable
RF	Mean, max. and min. daily air temperature (°C) Mean daily total radiation (shortwave) ($J\,m^{-2}$) Mean daily discharge ($m^3\,s^{-1}$) MOY and DOY	Water temperature
ANN	Mean, max. and min. daily air temperature (°C) Mean daily total radiation (shortwave) ($J\,m^{-2}$) Mean daily discharge ($m^3\,s^{-1}$) MOY and DOY	
SVR	Mean, max. and min. daily air temperature (°C) Mean daily total radiation (shortwave) ($J\,m^{-2}$) Mean daily discharge ($m^3\,s^{-1}$) MOY and DOY	
Air2stream	Mean daily air temperature (°C) Mean daily discharge ($m^3\,s^{-1}$)	
MR	Mean, max. and min. daily air temperature (°C) Mean daily total radiation (shortwave) ($J\,m^{-2}$) Discharge ($m^3\,s^{-1}$) MOY and DOY	

air temperature, (6) mean air temperature + discharge + radiation + maximum air temperature + minimum air temperature + MOY, and (7) mean air temperature + discharge + radiation + maximum air temperature + minimum air temperature + MOY + DOY.

The effect of the watershed geomorphological and hydrological variables was addressed with the analysis of the watershed time of concentration, a variable that encapsulates some of the main watershed characteristics that affect the river WT. The well-known Temez equation (Temez, 1978) (Sect. 3.8) initially defined for small-scale Mediterranean watersheds was selected for this analysis. Additionally, the Gaussian mixture model algorithm implemented with the machine learning Python package, scikit-learn (Pedregosa et al., 2011), was used for cluster analysis. The algorithm assumes that the data points belong to a mixture of normal distributions. The covariance structure of the data and the center of the distributions are used to compute probabilistic cluster assignments.

The results from the various models were evaluated with six metrics considering the observed and predicted daily datasets of river WT. During the evaluation of results three types of datasets were considered.

- *Annual datasets*. All available daily averages of WT are compared to field data.
- *Wet season*. Only the daily averages of WT corresponding to the wet season are compared to field data (October to March).
- *Dry season*. Only the daily averages of WT corresponding to the dry season are compared to field data (April to September).

The metrics were selected in order to not only provide a consistent interpretation of the models' results, but also to facilitate comparison with the results obtained in other studies (Sect. 3.9). The following sections describe each of the models and outline their relevant advantages and disadvantages.

3.1 Random forest

The RF algorithm (random forest regressor) was implemented with the machine learning Python package, scikit-learn (Pedregosa et al., 2011). This model fits classifying decision trees on various subsamples of the datasets and then combines the predictions. Decision trees can model complex nonlinear relations. The algorithm uses averaging to control overfitting and improve the algorithm predictive accuracy, thus effectively balancing the bias–variance trade-off. They are robust to outliers, missing values, and irrelevant or noisy variables because the model implicitly performs feature selection and generates uncorrelated decision trees. Beyond these advantages, there is one major drawback common to all the ML methods, with results difficult to interpret due to the intrinsically black-box nature of the algorithm. More details about RF can be found in the literature (Breiman, 2001; Louppe, 2014).

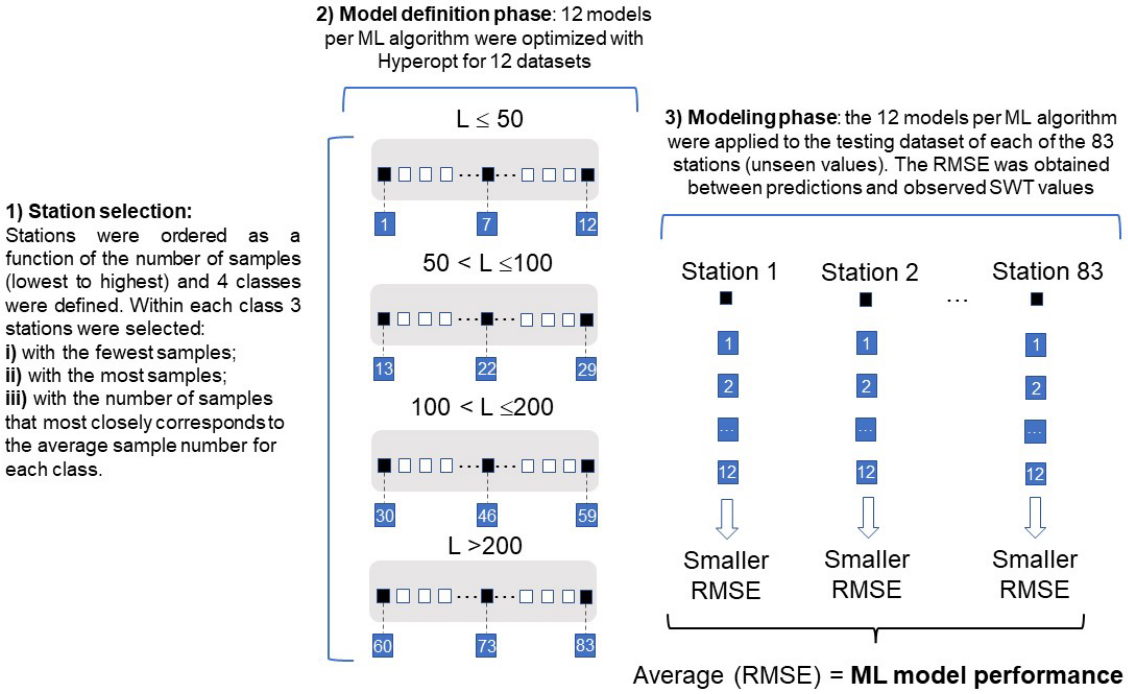

Figure 2. Schematic and simplified representation of the modeling process. Initially, 12 stations were selected as a function of the number of samples they contained. The ML models were trained and optimized for the 12 station datasets (model definition phase). The ML models were then applied to the 83 stations (modeling phase). The ensemble of the best results as a function of the RMSE describes the final ML results per station.

3.2 Artificial neural network

The ANN prototyping and building were achieved with the NeuPy Python library (Shevchuk, 2022). This library uses Tensorflow (an open-source platform for machine learning) as a computational back end for deep learning models (Abadi et al., 2016). The momentum algorithm was selected for the ANN implementation because of the improved control it provides with regard to overfitting. This is an iterative first-order optimization method that uses the gradient calculated from the average loss of a neural network. This algorithm promotes a gradual transition in the balance between stability and rate of change (Qian, 1999); the result is faster convergence and reduced oscillation. ANN has been successfully used to model river WT (Chenard and Caissie, 2008; DeWeber and Wagner, 2014; Piotrowski et al., 2015). This type of model is reasonably accurate and does not require a large number of input variables but does have two significant drawbacks. The model has no capacity to provide information on energy flux mechanisms within the river and has a tendency to overfit the training dataset, thereby considerably diminishing the model's ability to generalize the features or patterns present in the training dataset (Srivastava et al., 2014). For the implementation of the model, the training data were shuffled before training and the weights were randomly initiated. The loss function included the MSE to measure the accuracy of the results, as well as L2 regularization and dropout layers to minimize overfitting. The step decay algorithm was used to regularize the learning rate.

3.3 Support vector regression

The epsilon support vector regression algorithm was also implemented using the machine learning Python package, scikit-learn (Pedregosa et al., 2011). This type of algorithm is generally characterized by the use of kernels functions, sparseness of the solution and the absence of a local minimum (Platt, 1998; Smola and Schölkopf, 2004). The algorithm searches for a line or hyperplane in multidimensional space that divides two or more variables. The hyperplane with the optimum number of points is the best fit (Awad and Khanna, 2015). The SVR training relays on the use of a symmetrical loss function, which penalizes high and low errors. The algorithm also ignores errors that are lower than a certain threshold, ε. According to Awad and Khanna (2015), the computational complexity of the algorithm does not depend on the dimensionality of the input space, which is a relevant advantage. It also offers good prediction accuracy and excellent generalization capability. Regardless of the advantages, this algorithm can be computationally expensive, which can be a significant drawback.

3.4 Air2stream

The Air2stream model solves a lumped heat-exchange budget between an unknown river section volume, its tributaries, groundwater and the atmosphere (Toffolon and Piccolroaz, 2015). The river WT variation is described by the following equation:

$$\rho C_p V \frac{\mathrm{d}T_W}{\mathrm{d}t} = AH + \rho C_p \left(\sum_i Q_i T_{W,i} - Q T_w \right), \quad (1)$$

where T_w is the WT of a river section with a volume V and surface area A, and ρ and C_p are the water density and the specific heat capacity, respectively. H is the net heat flux at the air–water interface, and $T_{\mathrm{W},i}$ is the ith WT of the discharge Q_i tributary or groundwater. The model assumes that air temperature can be used as a proxy for all surface heat fluxes. A Taylor series expansion is used to include the overall effect of air temperature. Q is the discharge downstream of the river section and t is time. Equation (2) is the simplified form of Eq. (1) (Toffolon and Piccolroaz, 2015). This equation, with eight parameters, forms the basis of the Air2stream model:

$$\frac{\mathrm{d}T_\mathrm{W}}{\mathrm{d}t} = \frac{1}{\theta^{a_4}} (a_1 + a_2 T_\mathrm{a} - a_3 T_\mathrm{w} + \theta \left(a_5 + a_6 \cos \left(2\pi \left(\frac{t}{t_y} - a_7 \right) \right) - a_8 T_\mathrm{w} \right)), \quad (2)$$

where T_a is the air temperature, θ is the dimensionless discharge ($\theta = Q/\overline{Q}$)(3) and $\overline{Q}$ is the mean discharge. The parameter a_4 is related to the exponent of the rating curve. The model is fitted to the entire input dataset (air temperature, WT and discharge), and the value of a_4 and the value of all others model parameters are estimated during the model optimization process (calibration phase). In this study the Crank–Nicolson scheme was used to solve the model equation. Following Toffolon and Piccolroaz (2015), the model parameters were estimated using the particle swarm optimization method with inertia weight (Shi and Eberhart, 1998) with a population size of 2000 particles and 2000 iterations. In this study five versions of this model were considered to model WT: the three-, four-, five-, seven- and eight-parameter versions. Please refer to Toffolon and Piccolroaz (2015) for a full description of each one of the models' parameterizations.

3.5 Multiple regression

This model was implemented using the machine learning Python package, scikit-learn (Pedregosa et al., 2011). In this model the predicted value is expected to result in a linear combination of the input features:

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p, \quad (3)$$

where $\hat{y}$ is the predicted value, w_0 is the $\hat{y}$ intercept (constant term), w_1 to w_p are the model coefficients, and x_1 to x_p are the model input features. The model fits a linear model with coefficients w_1 to w_p to minimize the residual sum of squares between the observed and predicted values.

3.6 Hyperparameter optimization

Hyperparameter optimization was achieved using the Tree-structured Parzen Estimator (TPE) algorithm implemented with the Hyperopt library (Bergstra et al., 2013). The optimization process is initiated with the selection of a prior distribution (e.g., uniformly distributed); then, for the first iterations, the TPE algorithm is warmed up with some random iterations (random search). After this initial setup the algorithm collects new observations, and on completion of the iterations it selects the set of parameters that it will try during the next iteration. The algorithm scores and divides the collected observations into two groups. The first group includes the best observations and the second group all the others. The main objective is to identify a set of parameters most likely to be in the first group. The TPE algorithm can serve as a good alternative to the Gaussian process (GP) with expected improvement (EI) as it fixes some of the disadvantages associated with the latter. It can be difficult to select the right hyperparameters for GP with EI due to the many different kernel types associated with this process. TPE uses simpler kernels as a building block, which facilitates hyperparameter selection. Furthermore, TPE is faster than GP with EI when the number of hyperparameters increases. One notable drawback, however, is that the TPE algorithm selects parameters independently from each other. It is a well-known fact that the number of epochs of an ANN and regularization are related and that these two parameters influence the overfitting to a significant degree. To overcome this problem two different choices for the epochs, with and without regularization, were constructed. TPE hyperparameter optimization consists of 20 random parameter samples and 200 iterations. The Hyperopt algorithm samples 1000 candidates and selects the candidate that has the highest expected improvement (n_EI_candidates = 1000). The coefficient of determination (R^2) was considered to be the algorithm score. The algorithm uses 20 % of best observations to estimate the next set of parameters (gamma = 0.2). Table A1 shows the model parameters and the corresponding optimization range.

3.7 Synthetic Minority Oversampling Technique for Regression with Gaussian Noise (SMOGN)

SMOGN (Branco et al., 2017) is highly effective when working with imbalanced regression datasets. The algorithm applied with the Python implementation obtained from the SMOGN GitHub repository (SMOGN, 2022) combines random undersampling with two oversampling approaches: the Synthetic Minority Oversampling Technique for Regression (SMOTER) (Torgo et al., 2013) and SMOTER with Gaussian Noise (SMOTER-GN) (Branco et al., 2016). The al-

gorithm selects between two sampling techniques, considering the K-nearest-neighbor (KNN) distances underlying an observation: if the distance is too great, SMOTER-GN is applied; otherwise SMOTER is applied. By combining the two approaches to generate synthetic samples the authors made the decision to apply SMOGN, a more conservative approach which would minimize the potential risks incurred with SMOTER (Branco et al., 2017). Table A2 describes the parameter search space considered to derive the model datasets.

3.8 Time of concentration

The time of concentration was estimated using the Temez equation (Temez, 1978), which was defined for small natural watersheds located in Spain. In this equation, T_C is the time of concentration in hours, L is the length of the main water line (km) and J is the mean steepness (ratio between the mean fall and the L length of the water line) (m m^{-1}).

$$T_C = 0.3\left(\frac{L}{J^{1/4}}\right)^{0.76} \tag{4}$$

3.9 Evaluation metrics

Model assessment was performed with six different metrics: the mean absolute error (MAE), the root mean square error (RMSE), the Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), the Kling–Gupta efficiency (KGE) (Kling et al., 2012), the bias and the coefficient of determination (R^2). The metrics were computed using the following equations, where m_i and o_i are the modeled and observed values, $\overline{m}$ and $\overline{o}$ are their means, σ_m is the standard deviation of the modeled values, σ_o is the standard deviation of the observed values, and r is the Pearson coefficient:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|m_i - o_i|, \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2}, \tag{6}$$

$$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{N}(o_i - m_i)^2}{\sum_{i=1}^{N}(o_i - \bar{o})^2}\right], \tag{7}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_m}{\sigma_o} - 1\right)^2 + \left(\frac{\bar{m}}{\bar{o}} - 1\right)^2}, \tag{8}$$

$$\mathrm{bias} = \bar{m} - \bar{o}, \tag{9}$$

$$R^2 = \frac{\sum_{i=1}^{N}(m_i - \bar{o})^2}{\sum_{i=1}^{N}(o_i - \bar{o})^2} \times 100. \tag{10}$$

The RF and ANN algorithms use the mean square error to measure the results accuracy:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2. \tag{11}$$

4 Results

4.1 Air temperature – ERA5-Land versus ground-observed datasets

In this analysis the observed air temperature datasets of a total of 11 meteorological stations were considered. These are all the available air temperature datasets observed within a 5 km radius of the stations considered in this study. The results show that the mean RMSE obtained between the two datasets considering all stations varied from 1.26 to 2.05 °C ($\mu = 1.54$ °C; $\sigma = 0.24$ °C) and that, according to the mean bias values, the ERA-Land datasets tend to overestimate the observed air temperature datasets at 91 % of the stations. Overall, a mean RMSE value of 1.54 °C ($\sigma = 0.24$ °C) and a mean NSE value of 0.90 ($\sigma = 0.07$) are indicative of a good performance. This conclusion corresponds to the results obtained in other studies, namely Vannela et al. (2022) (Italy, three regions – RMSE: 1.76, 1.82 and 1.97 °C), Araújo et al. (2022) (Brazil, three regions – RMSE: 0.60, 1.11 and 0.41 °C) and Zhao and He (2022) (China, one region, −2.2 °C). However, as shown in Fig. 3, several significant sporadic discrepancies were produced between the two datasets. The results also show a nationwide distribution of stations with an RMSE of over 2 °C. Generally, these results suggest that the consideration of the ERA5-Land air temperature datasets for WT modeling can, sporadically, induce some significant discrepancies between the two datasets.

4.2 Model intercomparison – annual datasets

The results obtained from all the models for the testing phase and the annual datasets showed the RF model ensemble, with a mean RMSE of 3.18 °C ($\sigma = 1.06$), to be the top-performing model. The ANN model ensemble, with a mean RMSE of 3.22 °C ($\sigma = 1.05$), and the SVR model ensemble, with a mean RMSE of 3.37 °C ($\sigma = 0.96$), ranked second and third, respectively (Table A3). The SVR model produced the lowest RMSE of all the simulations run: 1.34 °C for station 8 with a training dataset of 20 values (SVR parameters: kernel, "sigmoid"; degree, 3; $C = 1000$, gamma = 0.0001, epsilon = 0.005). The RF was also the best-performing model based on a single model run (RF parameters: n_estimators, 50; max_depth, 485; min_samples_split, 5; max_features, "auto"; bootstrap, true), with a mean RMSE of 3.37 °C ($\sigma = 0.96$).

The Air2stream model with three parameters is the best of the hybrid model parameterizations, with a mean RMSE of 4.06 °C ($\sigma = 1.17$), followed by the MR, with an annual

Table 4. Evaluation of ERA5-Land daily air temperature datasets – MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and ERA5-Land values.

Station	Number of dataset values	MAE, °C	RMSE, °C	NSE	KGE	Bias, °C	R^2
st4	80	1.10 ± 0.26	1.39 ± 0.28	0.91 ± 0.04	0.94 ± 0.05	0.74 ± 0.47	0.94 ± 0.02
st6	120	1.10 ± 0.37	1.34 ± 0.38	0.90 ± 0.17	0.92 ± 0.09	-0.15 ± 0.79	0.90 ± 0.11
st30	98	1.31 ± 0.29	1.72 ± 0.40	0.91 ± 0.07	0.95 ± 0.06	-0.48 ± 0.70	0.92 ± 0.05
st32	67	1.16 ± 0.52	1.43 ± 0.58	0.96 ± 0.04	0.94 ± 0.05	-0.75 ± 0.90	0.97 ± 0.02
st38	110	0.88 ± 0.34	1.26 ± 0.48	0.94 ± 0.09	0.96 ± 0.06	-0.46 ± 0.57	0.95 ± 0.04
st42	21	1.19 ± 0.47	1.53 ± 0.58	0.93 ± 45.49	0.87 ± 2.22	-0.42 ± 0.75	0.94 ± 0.03
st50	90	1.08 ± 0.30	1.45 ± 0.48	0.91 ± 0.06	0.89 ± 0.11	-0.14 ± 0.39	0.92 ± 0.04
st62	24	1.30 ± 0.74	1.67 ± 0.80	0.89 ± 6.28	0.94 ± 0.92	-0.17 ± 1.36	0.90 ± 0.04
st68	47	1.60 ± 1.18	2.05 ± 1.09	0.71 ± 9.81	0.86 ± 0.23	-1.47 ± 1.24	0.88 ± 0.2
st83	137	1.49 ± 0.40	1.79 ± 0.39	0.92 ± 0.03	0.94 ± 0.04	-0.60 ± 0.88	0.93 ± 0.02
st91	51	1.04 ± 0.13	1.33 ± 0.16	0.93 ± 0.04	0.96 ± 0.09	-0.46 ± 0.47	0.94 ± 0.04

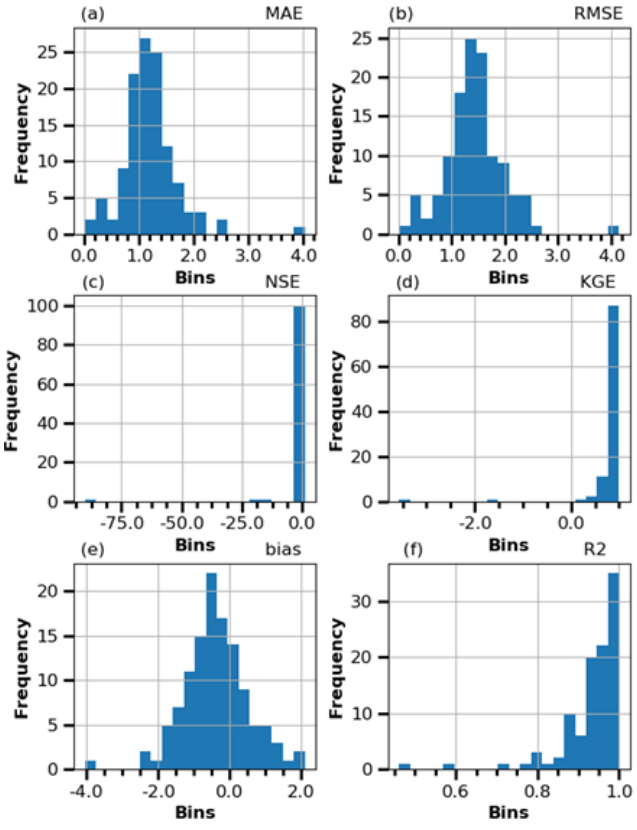

Figure 3. Metric histograms of daily air temperature – ERA5-Land versus ground-observed datasets.

mean RMSE of 4.28 °C ($\sigma = 1.17$). The NSE, KGE and R^2 values are closely aligned with the RMSE variation among the different models. Considering the performing ratings defined by Moriasi et al. (2007), the results obtained with the RF model ensemble, as described by the mean annual NSE value ($\mu = 0.52$; $\sigma = 0.23$), can be considered satisfactory ($0.50 < \text{NSE} < 0.65$). According to the same classification, the ANN and the SVR, with a mean annual NSE value of 0.48 ($\sigma = 0.28$) and 0.47 ($\sigma = 0.19$), produce an unsatisfactory modeling performance ($\text{NSE} \leq 0.50$). The same classification was obtained with all the parameterizations of the Air2stream model and the MR, but with a significantly reduced NSE value. The mean annual RMSE for the ensemble of all model results for the testing phase was 2.75 °C ($\sigma = 1.00$), varying from 1.34 to 6.03 °C. Therefore, according to the mean NSE value ($\mu = 0.56$; $\sigma = 0.48$), the model ensemble can be considered satisfactory. The contribution of the individual models to the results ensemble considering the stations with the lowest mean annual RMSE was as follows – RF: 35; ANN: 17; SVR: 14; Air2stream (3 par): 1; Air2stream (8 par): 2; MR: 14. It is important to mention that these results are not correlated with the number of values in the training or testing datasets but are a consequence of the dataset's quality and of the model's performance.

Figures 4 and 5 show the RMSE obtained with each model during the training and testing phases, respectively. The interannual variability is described by the standard deviation. The stations are ordered as a function of the number of training and testing datasets, from the smallest to the largest.

The results help to explain the performance of the models during the testing phase by showing the following.

1. During the training phase, all models exhibited a very low mean RMSE and interannual variability, except the Air2stream (three parameters) and the MR.

2. The RF underfitted the training datasets with fewer than 30 values, and consequently the predicted WT values exhibited a high RMSE and interannual variability during the testing phase ($\sigma = 1.28$) (Figs. 4 and 5).

3. During the training phase, the ANN exhibited the lowest mean annual RMSE ($\mu = 0.44$ °C; $\sigma = 0.40$) (Table 4). This model clearly overfitted the training datasets, with fewer than 30 values, which increased the RMSE obtained for stations 1 to 11 (Figs. 4 and 5). The model

mean RMSE variability during the testing phase is equal to that obtained for the RF, which exhibited the lowest variability during the testing phase ($\sigma = 1.28$).

4. Like the ANN, the SVR overfitted the training datasets of the first 10 stations, although the model had the lowest mean RMSE interannual variability during the testing phase ($\sigma = 1.25$), including for stations 1 to 10.

5. The Air2stream (three-parameter) model and the MR exhibited the highest mean RMSE and interannual variability during both phases. In fact, the MR exhibited a significant degree of interannual variability ($\sigma = 4.10$) for the datasets with fewer than 30 values (stations 1 to 10), which was reflected in the results obtained during the testing phase.

Figure 6 was included to provide greater insight into the underfitting and overfitting associated with the ML models. The training datasets with fewer than 30 values are clearly underfitted by the RF model (Fig. 6a) and overfitted by the ANN and SVR (Fig. 6c and e). In the case of the ANN and the SVR, the overfitting is stronger and more closely correlated with the number of training datasets (RF: $R^2 = 0.13$; ANN: $R^2 = 0.52$; SVR: $R^2 = 0.58$).

It is also interesting to look at the results obtained from the models with regard to levels of performance. Figure 7 shows the temporal evolution of the WT values obtained during the training and testing datasets for stations 59 (138 training values) and 2 (11 training values). Based on the RF model results, these are the stations with the best and worst mean annual RMSE. There are clear, fundamental differences between the ML models and the Air2stream and MR models. The ML models are highly effective. They describe a large number of spurious observed values in the WT values that can be associated with the sub-daily variation of the river WT, underground inflows or a monitoring error, and, by doing so, the predicted temporal evolution of the river WT oscillates widely (Fig. 7a, c and e). This was not the case with the Air2stream or MR models. The results obtained from these two models demonstrate the fact that, in the absence of quality input training information (quantity plus quality), their predictive performance is significantly lower than that of the ML models. This is illustrated by the less oscillating sinusoidal evolution of the river WT (Fig. 7g and i). When considering very small training datasets, such as the dataset corresponding to station 2, with 11 training values and 5 testing values, ML models tend to have a very unrealistic response as they either overfit or underfit the training datasets (Fig. 7b, d and f). In this example, the Air2stream (five-parameter) model has a delayed but more realistic response. The MR performed the worst, with the model unable to describe the correlation between the predictor variables and the observed river WT (Fig. 7j).

4.3 Model intercomparison – seasonal datasets

The results obtained for the dry and wet season testing datasets, considering all metrics, suggest that model performance is better for the dry season, with the exception of the results obtained with the Air2stream model using three, four and five parameters (Tables A4 and A5). The model using three and four parameters does not consider the effect of river discharge, and the five-parameter version assumes that the effect of the discharge can be retained using only a constant value. This suggests that the inclusion of discharge data increased the error in the wet season simulation for the Air2stream model with seven and eight parameters. Following the initial selection of the gauging and water quality stations, the missing discharge values were replaced by the corresponding climatological year value. Missing discharge data replacement varied from 0.0 % to 82.6 % ($\mu = 30.0$; $\sigma = 22.3$). Approximately 28 % of the stations have missing discharge values of over 50 %, which represents an important source of uncertainty that probably affected the Air2stream model performance.

The results obtained with the best-performing model (RF) considering the annual datasets are in line with the previous conclusion that model performance is better for the dry season, but only when the DOY predictor is excluded. The inclusion of the DOY predictor modified the correlation among the different variables and the performance of the models over the wet and dry season, enhancing the importance of this variable in relation to the overall modeling performance.

Overall, the results are, as expected, similar to those obtained for the annual datasets, showing that the ANN and the SVR models overfitted the training datasets, in particular during the wet season, which also contributed to the worst model performance during this season. The differences regarding the mean MAE and RMSE of the testing phase are very small among the ML models, with the results of the ANN ensemble coming out slightly ahead of those obtained through the RF and SVR ensemble for both seasons. This deviation in terms of the results obtained for the annual datasets is driven by the difference in the length of the annual versus seasonal datasets and, consequently, the computation of the metrics, namely the MAE and the RMSE, highlighting the similarity between the ML models results. This is further emphasized by the mean NSE and KGE values, which, in the case of the wet season testing datasets, provide a contradictory result. According to the mean NSE, the RF and SVR model ensembles produce the best results (NSE – RF: 0.13, $\pm$1.91; SVR – 0.13, $\pm$1.10; ANN – 0.10, $\pm$1.22); nonetheless, the mean KGE values favor the ANN ensemble over the other ML results (KGE – RF: 0.46, $\pm$0.26; SVR – 0.37, $\pm$0.26; ANN – 0.48, $\pm$0.36). The Air2stream model with three parameters is the best of the hybrid model parameterizations, followed by the MR (Tables A4 and A5).

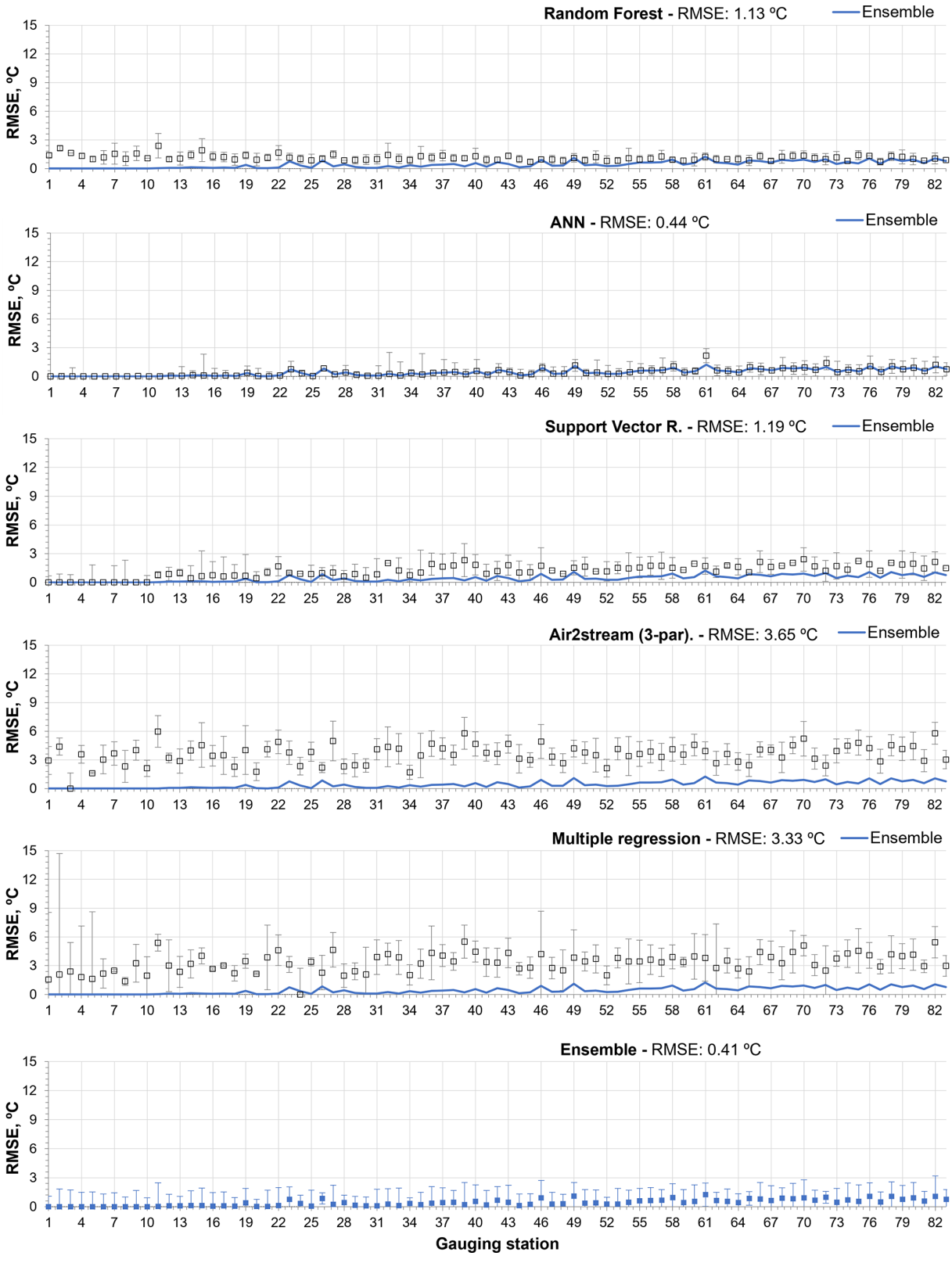

Figure 4. Root mean square error between observed and predicted WT values obtained during the training phase with all models (with standard deviation of interannual RMSE), considering the model results and the ensemble of all models results. Stations are ordered by the number of training dataset values, from smallest to largest.

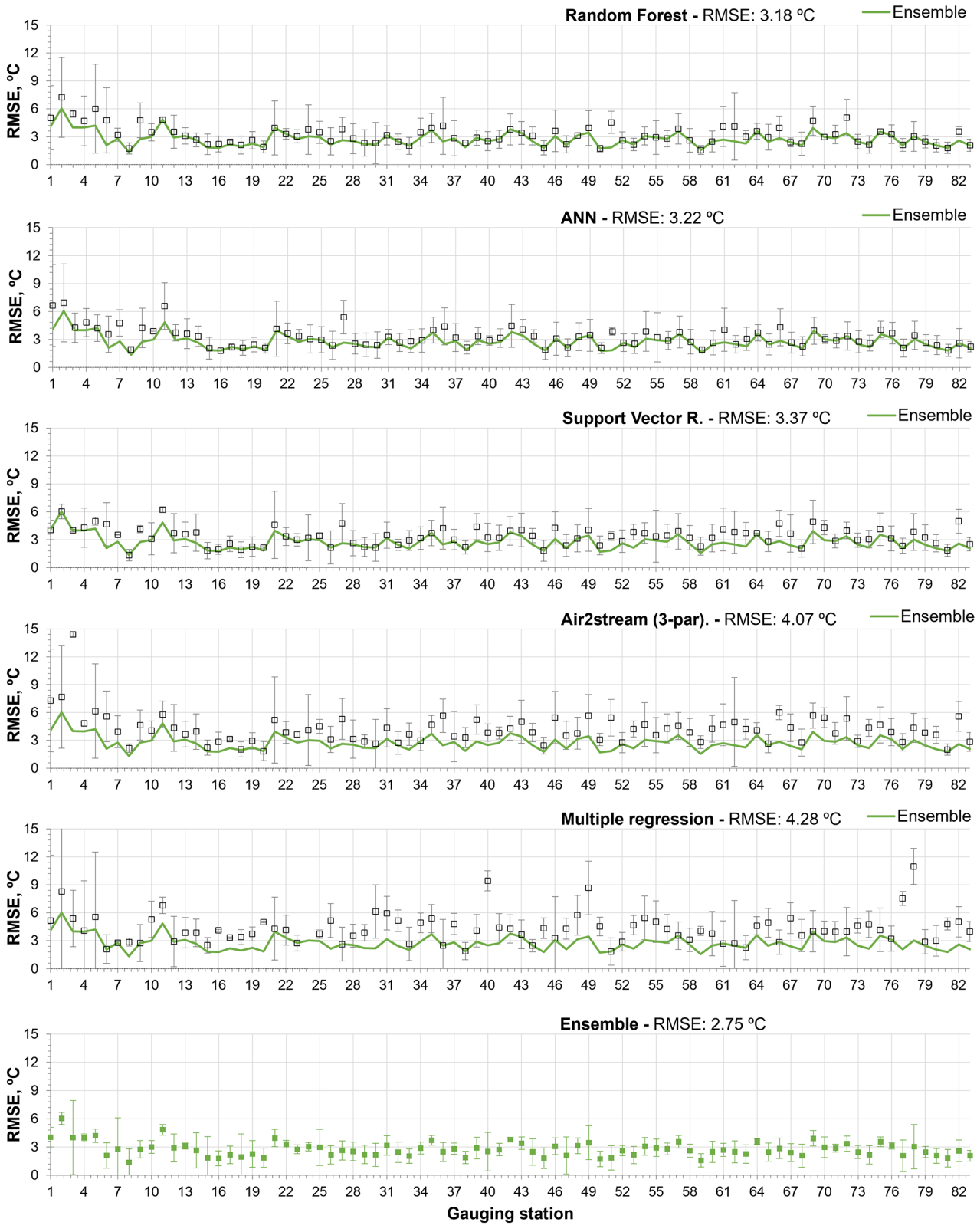

Figure 5. Root mean square error between observed and predicted WT values obtained during the testing phase with all models (with standard deviation of interannual RMSE), considering the model results and the ensemble of all models results. Stations are ordered by the number of testing dataset values, from smallest to largest.

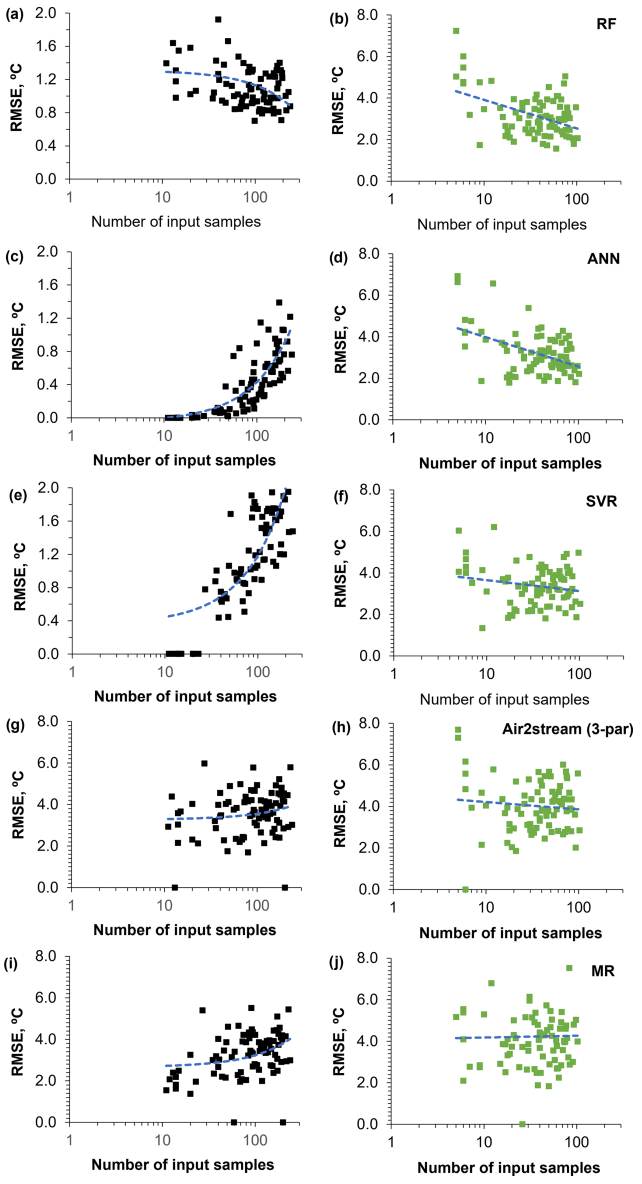

Figure 6. Root mean square error between observed and predicted WT values obtained with all models during the training (black dots) and testing (green dots) phases, ordered by the number of values in the training and testing datasets (from smallest to largest).

4.4 Modified training datasets – oversampling–undersampling technique

Based on the result obtained in Sect. 4.2, the best-performing model was the RF, and this model was therefore considered to evaluate the improvement of accuracy driven by the modified training datasets. The training datasets derived from the application of SMOGN and the ML optimization modeling approach have different characteristics due to the differences in the degree of oversampling–undersampling conducted ("extreme" versus "balanced") and the selection of the domain region of WT values considered to be rare ("high", "both", "low"). Table 5 shows the number of values in the raw training datasets and in the training datasets obtained with SMOGN corresponding to the best RF model performance. In general terms, considering all stations, oversampling had a more pronounced effect on 50 % of the stations (six stations), while undersampling influenced the sampling process in the case of 33 % of the stations. For the remaining two stations oversampling and undersampling had an identical effect on the total number of raw training datasets. The extreme–both parameterization was considered for the modeling of 58 % of the stations (seven stations), suggesting that more oversampling–undersampling was the best modeling solution for both, high and low regions. This parameterization was followed by extreme–high and balance-both, with 25 % (three stations) and 17 % (two stations), respectively (Table A8). The WT range affected by both the oversampling and undersampling process was similar for stations 46, 60 and 73, as described by the mean WT values (Table 5). These results suggest the tendency for stations with a lower number of values to be affected in different WT ranges, a fact mainly driven by the availability of samples within each region of the response variable WT.

The results obtained with the RF model forced with the modified training datasets had a significant effect on the modeling results. The RF model performance considering the raw training datasets and the modified training datasets is shown in Tables A6 and A7, respectively. The mean RMSE and MAE values obtained between the predicted and observed datasets were reduced from 0.0 % to 21.9 % ($\mu = 8.6 \pm 8.2$) and from 0.0 % to 29.9 % ($\mu = 10.3 \pm 9.2$), which can be considered a significant improvement of the RF model accuracy (Table 6). In fact, the RMSE and MAE values were reduced by more than 18 % and 15 %, respectively, for 50 % of the stations with fewer than 80 training samples.

4.5 Feature importance

Table 7 shows the mean feature importance obtained with the best-performing model (random forest regressor, Pedregosa et al., 2011) considering the mean annual RMSE and an RF with the following parameters, considering all station datasets: n_estimators, 50; max_depth, 485; min_samples_split, 5; max_features,"auto"; bootstrap, true; random_state, 42. The maximum importance values show that all features are relevant, at least for some stations, and that they should not be discarded. The mean importance values indicate that the mean air temperature and the DOY are of upmost importance in relation to the model training process, followed by the maximum and minimum air temperature. Discharge, global radiation and MOY clearly play a secondary role, as described by the mean and standard deviation values. Table A9 shows the evaluation of the RF model performance during the training and testing phases considering the annual datasets and the sequential increase in the

Table 5. Number of values in the raw training datasets and the training datasets obtained with SMOGN, corresponding to the best RF model performance.

Station	Train (raw datasets) (number of values)	Oversampling		Undersampling		Train (modified datasets) (number of values)	Test (number of values)
		Number of values	Water temperature range, °C (minimum; maximum; average with standard deviation)	Number of values	Water temperature range, °C (minimum; maximum; average with standard deviation)		
1	10	0	–	0	–	10	5
7	14	15	16.00; 28.00; 19.27 ± 3.39	4	19.00; 28.00; 22.25 ± 4.27	25	7
12	35	4	12.94; 15.75; 14.20 ± 1.40	10	17.60; 22.40; 20.10 ± 1.54	29	15
13	35	16	16.13; 19.80; 17.54 ± 1.15	10	8.30; 15.30; 13.16 ± 2.28	41	16
22	50	39	8.00; 26.00; 16.36 ± 4.51	17	15.00; 26.00; 19.76 ± 2.95	72	22
29	69	23	16.75; 22.56; 18.93 ± 1.57	23	6.50; 16.00; 12.42 ± 2.54	69	30
30	71	41	8.87; 22.00; 16.60 ± 4.47	9	7.00; 22.00; 14.79 ± 4.84	103	31
46	98	12	8.20; 36.00; 19.21 ± 8.58	33	14.60; 22.00; 18.58 ± 2.06	77	43
59	137	120	8.60; 26.00; 17.91 ± 3.53	40	9.00; 23.00; 15.77 ± 4.23	217	60
60	141	27	10.03; 25.03; 16.64 ± 5.76	31	14.10; 21.20; 17.16 ± 2.18	137	61
73	177	15	8.52; 30.00; 16.79 ± 6.41	53	13.50; 19.10; 17.04 ± 1.22	139	76
83	236	353	8.50; 28.00; 16.96 ± 3.41	32	17.00; 26.00; 20.18 ± 2.19	557	102

Table 6. Percent variation between the metrics obtained with the raw training datasets and the modified training datasets with RF model.

Annual	Train						Test					
Station/metric	MAE (%)	RMSE (%)	NSE (%)	KGE (%)	Bias (%)	R^2 (%)	MAE (%)	RMSE (%)	NSE (%)	KGE (%)	Bias (%)	R^2 (%)
1	−153.6	−129.0	19.7	4.5	100.0	22.1	2.5	0.7	−217.0	−10.1	19.5	1.2
7	7.3	−51.6	41.3	16.0	100.0	45.7	18.0	18.0	−38.8	−28.5	−3185.7	−40.2
12	100.0	100.0	−8.9	−23.9	100.0	−5.8	14.8	7.8	−82.2	34.4	65.4	−17.1
13	−17.5	−25.8	8.5	−6.8	197.5	13.4	29.2	17.6	−77.2	−150.9	406.2	−26.5
22	92.6	89.9	−10.1	−24.5	100.0	−6.4	3.6	2.2	−5.2	4.9	0.3	−8.2
29	22.7	18.3	−1.4	−10.5	100.0	−0.4	7.5	21.9	−23.7	−40.6	−55.2	−20.0
30	60.6	37.8	−6.6	−15.9	100.0	−5.2	5.9	3.2	−8.1	−11.9	60.9	−1.1
46	−60.7	−79.6	3.6	1.8	1168.2	3.2	23.1	18.7	−23.1	−20.4	390.2	−26.5
59	20.9	5.5	0.0	−0.3	187.9	0.1	9.2	8.7	−3.1	−1.0	67.6	−2.3
60	−3.7	−10.2	−0.2	−0.5	53.4	−0.1	8.4	3.0	−2.5	−7.3	29.5	−0.3
73	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
83	0.1	−20.7	3.8	3.5	263.2	3.5	0.4	1.2	−1.1	−2.1	−161.8	−2.0
Average	5.7 (±67.7)	−5.5 (±64.9)	4 .2 (±14.2)	−4.7 (±12.1)	205.9 (±310.8)	5.8 (±15.0)	10.3 (±9.2)	8.6 (±8.2)	−40.2 (±62.6)	−19.5 (±45.4)	−196.9 (±955.5)	−11.9 (±13.8)
Maximum	100	100	41	16	1168	46	29.2	21.9	0	34	406	1
Minimum	−154	−129	−10	−25	0	−6	0.0	0.0	−217	−151	−3186	−40

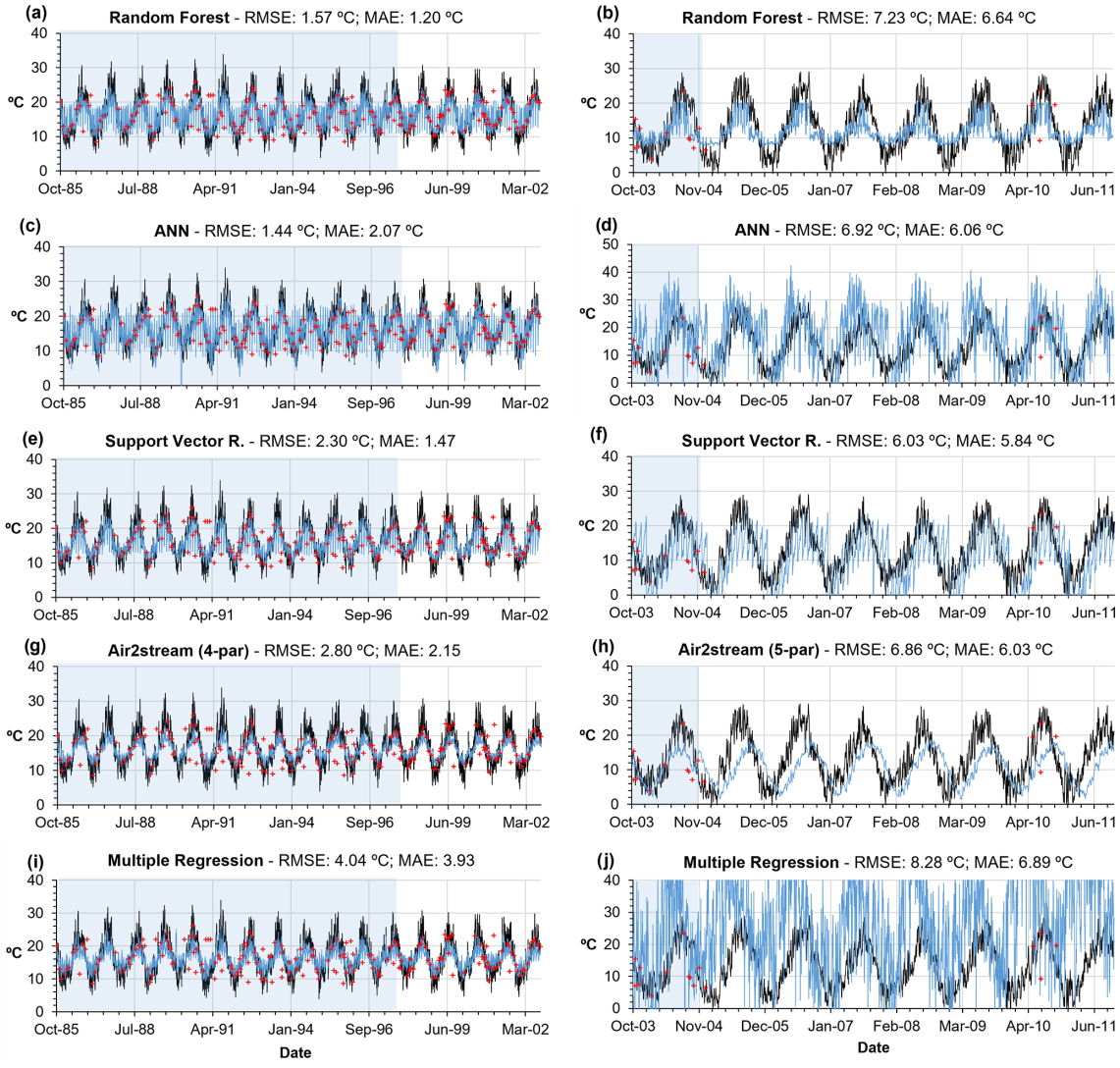

Figure 7. Root mean square error between observed (red dots) and predicted WT (blue line) values obtained during the calibration (blue shading) and testing phase (white shading) with all models for station 59 (graphs on left) and station 2 (graphs on right). Air temperature is represented by the black line.

model predictors. The results show that, on average, the inclusion of all predictor variables has a significant effect on model performance.

4.6 Effect of the watershed time of concentration on model performance

The results suggest that, tendentially, there are more training and testing datasets available for the largest watersheds (Fig. 8a and b) and that the watershed time of concentration increases with the watershed area according to a power law (Fig. 8c). Additionally, the graphic correlation of the RMSE between the observed river WT and the predicted WT (training datasets) obtained with the best-performing model run – the RF ensemble model and the best individual RF run with the watershed time of concentration – revealed the existence of a very specific linear pattern within the dataset (Fig. 9a and b). Two different data samples were extracted after the datasets' z-score normalization and the application of the Gaussian mixture model algorithm with the following parameters: n_components, 2; covariance_type, "diag"; init_params, "random"; warm_start, true (see Pedregosa et al., 2011). This small set of values, 19 (watershed area: $\mu = 106\,\mathrm{km}^2$; $\sigma = 153$) (Fig. 9a) and 19 (watershed area: $\mu = 106\,\mathrm{km}^2$; $\sigma = 153$) (Fig. 9b), corresponds to 35 % of the stations with fewer than 125 training values, a fact that enhances the non-random nature of this correlation. This correlation shows how the RMSE obtained with the RF increases with the watershed area, clearly showing the significant effect upstream conditions have on river WT. The RMSE in-

Table 7. Mean input feature importance obtained with a random forest regressor.

	Mean Air temperature	Maximum air temperature	Minimum air temperature	Discharge	Global radiation	Month of the year	Day of the year
Mean	0.20	0.12	0.15	0.09	0.10	0.06	0.29
Standard deviation	0.16	0.10	0.10	0.09	0.07	0.07	0.20
Maximum	0.70	0.46	0.62	0.33	0.28	0.34	0.82
Minimum	0.02	0.01	0.02	0.01	0.01	0.00	0.01

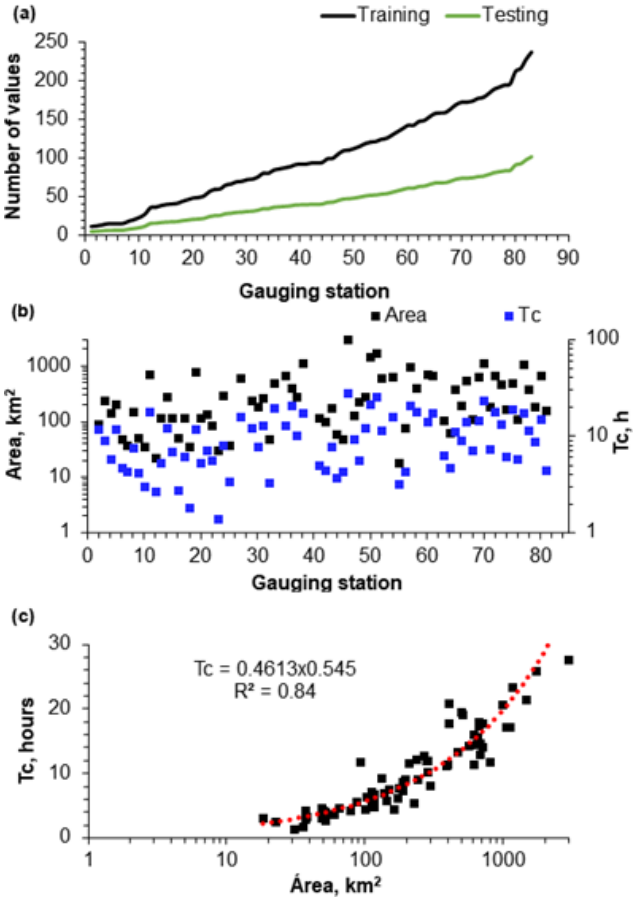

Figure 8. **(a)** Number of training and testing datasets of each station. **(b)** Watershed time of concentration and area of each station. **(c)** Watershed time of concentration versus watershed area.

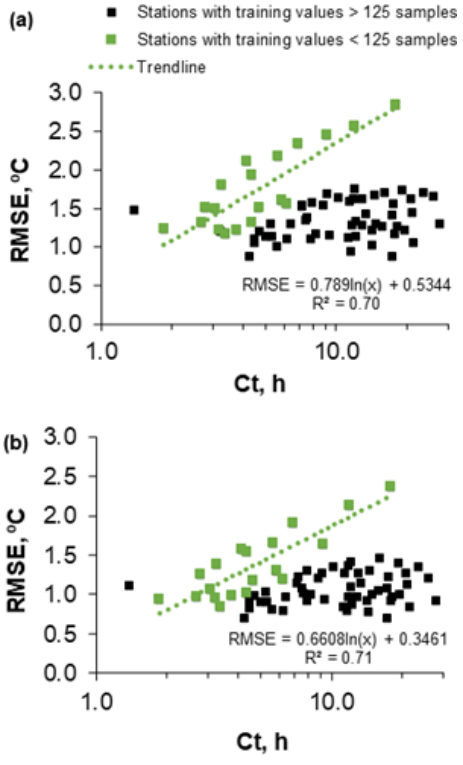

Figure 9. **(a)** RMSE between observed and simulated river WT with the random forest best model run versus the watershed time of concentration. **(b)** RMSE between observed and simulated river WT with the random forest ensemble aggregation approach versus the watershed time of concentration (Ct).

creases by an average of 0.1 °C with a 1 h increase in the watershed time of concentration, considering the RF ensemble aggregation approach (Fig. 9b).

5 Discussion

Overall, the results of the model's ensemble (mean RMSE: 2.75 °C; $\sigma = 1.00$) driven mainly by the predictions of the ML algorithms are in line with the results obtained in other studies, namely Rabi et al. (2015) (ANN – RMSE: $\mu =$ 2.06 °C) and Zhu et al. (2019a) (MR – RMSE: 2.74 °C). This is quite significant considering the scale of the missing training and testing datasets corresponding to this study ($\mu =$ 98.8 %; $\sigma = 0.68$). These results are, as expected, worse than the results obtained in some of the more recent studies in which ML algorithms were used to predict river WT (Table 1). However, the availability of training data for most of these studies was impressively good in terms of quantity and quality, which is, of course, reflected in the overall results.

The selection of the best approach to model river WT is not an easy task, as ML algorithm performance levels are very similar (Feigl et al., 2021). That said, considering all the metrics, the RF model ensemble produced the best re-

sults for the annual datasets and was the model that provided the greatest contribution in relation to overall ensemble results. As such, this was selected as the best model for modeling river WT for stations with limited forcing data. However, this is not in line with the findings of other studies. Rajesh and Rehana (2021) and Rehana (2019) concluded that the SVR model was the most robust model for predicting river WT temperature on a daily timescale. Feigl et al. (2021) concluded that the feed-forward neural networks (FNNs) and the recurrent neural networks (RNNs) performed better than the RF model. It is, however, important to highlight the significant variations in terms of the number of watersheds studied and the overall length of the training datasets used across all the different studies, which could effectively explain the different findings in relation to model performance.

One of this study's most significant conclusions is that, from a practical point of view, the application of all the models considered in this study is relevant. In fact, our results show that all models considered were best performers for some of the station datasets, including the MR, which was the best model for 14 stations. The results show that the advantages of the state-of-the-art ML models and the Air2stream model are reduced when the training datasets are very small (< 200 values) and span a long period of time. The information contained in the training datasets is not sufficient for the definition of the unknown underlying function that best relates the input variables to the output variable. Hence, the less complex approaches, such as MR, may surpass the results produced by ML algorithms.

The ML algorithms can considerably improve on the prediction results produced by the current state-of-the-art Air2stream model, regardless of the model parametrization. This finding concurs with that of Feigl et al. (2021) but is contrary to the results of the study carried out by Zhu et al. (2019d), which assessed the performance of a suite of machine learning models for daily stream WT. However, in the case of our study the performance of the Air2stream model was affected by the missing training data, namely the discharge datasets, which proved to be a significant obstacle for this model. When the dataset gap is very large, the structure of the Air2stream model with six or more parameters may become very complex when compared to the number of observed WT values, increasing the risk of overfitting (Piccolroaz, 2016). This explains the fact that, considering all the metrics, the best results were obtained with the three-parameter model, the simplest version of the Air2stream model. The three-parameter model does not consider the river discharge and depth on a daily timescale and, as such, can be successfully applied if the longitudinal gradient of temperature is small (Toffolon and Piccolroaz, 2015). The results of our study correspond to those obtained by Piccolroaz (2016) regarding the effect of missing data during the modeling of the WT of two lakes located in the USA (Lake Erie and Lake Superior) with the four- and six-parameter Air2stream model. When the length of the calibration period is 1 year and the percentage of missing data is in the range of 99 %, the RMSE between observed and predicted lake WT is $> 3.5\,^{\circ}\mathrm{C}$. It is also relevant to mention that the results of this study suggest that, besides the WT dataset gaps, the modeling results were also affected by the presence of a large number of WT outliers, by the uncertainty induced by the mean air temperature ERA5-Land reanalysis datasets and by upstream conditions, which increase with the watershed area. In terms of input dataset quality, the results of this study suggests that when the missing datasets reach 98 %, RMSE $< 3.0\,^{\circ}\mathrm{C}$ is indicative of a good modeling performance. Importantly, this error can also be further decreased by the generation of synthetic samples to some poorly represented ranges within the datasets by applying a model such as SMOGN (Branco et al., 2017).

The success of the models considered in this study, namely the ML algorithms, is undoubtedly linked to the hyperparameter optimization algorithm, a conclusion that is in line with the findings of Feigl et al. (2021). The feature importance analysis showed that all the predictors (mean, max. and min. daily air temperature, mean daily total radiation, discharge, MOY and DOY) are relevant to model performance, a conclusion that also concurs with the findings of Feigl et al. (2021). Nonetheless the results highlight the importance of the daily mean air temperature and DOY. The DOY was the most relevant variable. In fact, the inclusion of the DOY modified the correlation among the different variables and the performance of the models across the wet and dry season, increasing the importance of this variable to the overall modeling performance, which is in line with the findings of Zhu et al. (2019d). This suggests that the correlation associated with the other input variables and the observed river WT is, in fact, rather weak, which relates to the length and quality of the training datasets, as well as the uncertainty caused by the fact that a river's upstream environmental conditions can have a significant effect on WT predictions. However, it is also worth mentioning the lack of clarity in relation to the exact extent of the upstream area controlling the river energy balance at a given point (Moore et al., 2005), and, as such, the averaging of the predictor variables over the watershed area might not be the best solution. There are a number of limitations associated with our study that should be addressed in future studies. Firstly, regardless of the hyperparameter optimization and the inclusion of regularization and dropout layers to minimize overfitting in the ANN model, the results show that when the training datasets contain fewer than 30 values, the model will considerably overfit the datasets and considerably reduce the model's predictive capacity. This limitation might be minimized with more effective control of the number of training epochs and the regularization algorithm. It is also important to mention the fact that the hyperparameter optimization algorithm was not applied to all the station datasets; hence, the ML algorithms might be further improved. Due to the lack of physical restraints, ML models might fail when extrapolating outside

the range of their training datasets. This was not fully evaluated in this study due to the number of watersheds studied but certainly requires further investigation in the future. The modeling of 100 synthetic training datasets per station with the RF model to evaluate the SMOGN algorithm performance was very time-consuming. In fact, the average time required to model each station considered was 4.0 ± 0.45 h. Therefore, the accuracy of the RF model can probably be further increased if the number of training datasets is higher. If possible, this sensitive analysis should be combined with the evaluation of the loss of quality and consistency of the training datasets due to undersampling. The results of this study demonstrate the feasibility of finding a correlation between the prediction error for observed and predicted river WT values and the watershed time of concentration. However, the number of samples that form this correlation is small (19), and, as such, the number of watersheds studied needs to be increased to strengthen this correlation and scale it to other watersheds. The inclusion of the watershed soil type as a predictor variable would also be of relevance. It is also important to note that the results of this study are restricted to the Mediterranean region, and therefore the expansion of the study area to other latitudes to consider different climate and soil conditions would also be interesting, namely the north of Europe and Africa where data scarcity is quite relevant.

6 Conclusion

The results obtained with this study demonstrate, from a practical modeling perspective, the validity of applying all the models considered in this study – random forest, artificial neural network, support vector regression, Air2stream and multiple regression – when the number of predictor variables and observed river WT values is limited. It is also of upmost importance to optimize the ML algorithm hyperparameters. The Tree-structured Parzen Estimator algorithm has proved to be a good solution. The results of this study also show the viability of using all available predictor variables and highlights the importance of the day of the year and the mean daily air temperature. Regardless of the greater degree of modeling performance that can be attained with an ensemble of all the different models, the random forest model with the following parameters produces the best performance and may represent an effective solution for modeling river WT with limiting forcing data: n_estimators, 50; max_depth, 485; min_samples_split, 5; max_features, “auto”; bootstrap, true; random_state, 42. Importantly, our study further confirmed that the accuracy of the random forest can be significantly improved by the generation of synthetic samples to some poorly represented ranges within the training datasets by applying an oversampling–undersampling technique.

It is also relevant to mention that a logarithmic correlation exists in relation to the RMSE between the observed and predicted river WT and the watershed time of concentration. The RMSE increases by an average of 0.1 °C with a 1 h increase in the watershed time of concentration (watershed area: $\mu = 106\,\text{km}^2$; $\sigma = 153$), a conclusion that may prove useful for increasing our understanding of the effects of catchment size and landscape on runoff generation and, consequently, on river energy balance.

Appendix A

Table A1. Model parameters and optimization range.

Model	Prior distribution	Parameter	Optimization range
RF	uniform	"n_estimators"	[50, 2000]
	uniform	"max_depth"	[10, 1000]
	uniform	"min_samples_split"	[2, 10]
	–	"max_features"	[auto, sqrt]
	–	"bootstrap"	[True, False]
ANN	categorical	"n_layers"	[1, 2]
	uniform integer	"n_units_layer"	[10, 50]
	categorical	"act_func_type"	["Relu", "PRelu", "Elu", "Tanh", "Sigmoid"]
	categorical	"regularization"	[True, False]
	quantized distribution	"n_epochs"	With regularization: [500, 1000]; without regularization: [20, 300]
	uniform	"dropout"	[0, 1.0]
	loguniform	"batch_size"	[5, 20]
	uniform	"initial_value"	[0.001, 0.1]
	uniform	"reduction_freq"	[10, 200]
	uniform	"decay_rate" (regularization)	[0.0001, 0.001]
SVR	Categorical	"C"	[0.1,1, 100, 1000]
	Categorical	"kernel"	["rbf", "poly", "sigmoid", "linear"]
	Categorical	"degree"	[1, 2, 3, 4, 5, 6]
	Categorical	"gamma"	[1, 0.1, 0.01, 0.001, 0.0001]
	Categorical	"epsilon"	[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10]

Table A2. Description and optimization range (SMOGN, 2022) of modeling parameters considered for the application of SMOGN.

Parameter	Description	Parameter search space
k	Specifies the number of neighbors to consider for interpolation used in oversampling	uniform [1, 10]
samp_method	If "balance" is specified, less oversampling–undersampling is applied. If "extreme" is specified, more oversampling–undersampling is applied	Categorical [extreme, balance]
rel_thres	Specifies the threshold of rarity, takes a real number between 0 and 1	uniform [0, 1]
rel_coef	Corresponds to the box plot coefficient used to automatically determine extreme and therefore rare "minority" values in y, when rel_method = "auto"	uniform [0.01, 0.4]
rel_method	rel_method argument takes a string, either "auto" or "manual"; it specifies how relevant or rare "minority" values in y are determined – if "auto" is specified, "minority" values are automatically determined by box plot extremes	"auto"
rel_xtrm_type	The rel_xtrm_type argument takes a string, either "low" or "both" or "high"; it indicates which region of the response variable y should be considered rare or a "minority", when rel_method = "auto"	Categorical [high, both, low]

Table A3. Evaluation of model performance during the training and testing phases considering the annual datasets. Mean MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and predicted WT values.

Annual	Train					
Model/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
RF	0.86 (±0.25)	1.13 (±0.30)	0.93 (±0.03)	0.85 (±0.07)	−0.01 (±0.06)	0.96 (±0.02)
ANN	0.29 (±0.29)	0.44 (±0.40)	0.99 (±0.03)	0.98 (±0.03)	0.01 (±0.02)	0.99 (±0.03)
SVR	0.82 (±0.54)	1.19 (±0.64)	0.91 (±0.06)	0.88 (±0.09)	0.00 (±0.11)	0.92 (±0.05)
Air2stream (3 par)	2.82 (±0.86)	3.65 (±0.96)	0.33 (±0.25)	0.33 (±0.32)	0.01 (±0.01)	0.33 (±0.25)
Air2stream (4 par)	2.83 (±0.86)	3.65 (±0.97)	0.33 (±0.25)	0.34 (±0.31)	0.00 (±0.01)	0.33 (±0.25)
Air2stream (5 par)	2.72 (±0.88)	3.54 (±0.98)	0.36 (±0.25)	0.38 (±0.29)	0.00 (±0.01)	0.36 (±0.25)
Air2stream (7 par)	2.67 (±0.86)	3.50 (±0.99)	0.38 (±0.25)	0.42 (±0.28)	0.01 (±0.02)	0.38 (±0.25)
Air2stream (8 par)	2.68 (±0.87)	3.49 (±0.99)	0.39 (±0.24)	0.43 (±0.28)	0.01 (±0.04)	0.39 (±0.24)
MR	2.55 (±0.79)	3.33 (±0.95)	0.47 (±0.27)	0.49 (±0.24)	0.00 (±0.00)	0.44 (±0.22)
Ensemble	0.28 (±1.07)	0.41 (±1.36)	0.99 (±0.32)	0.98 (±0.27)	0.01 (±0.04)	0.99 (±0.30)
Annual	Test					
Model/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
RF	2.44 (±0.91)	3.18 (±1.06)	0.52 (±0.23)	0.60 (±0.20)	−0.07 (±1.11)	0.60 (±0.18)
ANN	2.50 (±0.86)	3.22 (±1.05)	0.48 (±0.28)	0.66 (±0.18)	−0.12 (±0.94)	0.55 (±0.22)
SVR	2.60 (±0.86)	3.37 (±0.96)	0.47 (±0.19)	0.53 (±0.21)	0.00 (±0.83)	0.54 (±0.18)
Air2stream (3 par)	3.17 (±1.06)	4.07 (±1.18)	0.21 (±0.32)	0.29 (±0.32)	−0.18 (±1.15)	0.34 (±0.22)
Air2stream (4 par)	3.30 (±1.15)	4.24 (±1.37)	0.11 (±0.73)	0.30 (±0.29)	−0.04 (±1.30)	0.32 (±0.23)
Air2stream (5 par)	3.53 (±1.08)	4.37 (±1.13)	0.06 (±0.59)	0.18 (±0.38)	−0.12 (±1.03)	0.30 (±0.22)
Air2stream (7 par)	3.74 (±1.15)	4.73 (±1.36)	−0.13 (±0.81)	0.19 (±0.32)	−0.50 (±1.51)	0.24 (±0.22)
Air2stream (8 par)	3.94 (±1.35)	5.06 (±1.73)	−0.56 (±2.27)	0.16 (±0.44)	−0.42 (±1.65)	0.23 (±0.22)
MR	3.34 (±1.29)	4.28 (±1.62)	0.32 (±0.34)	0.36 (±0.27)	−0.46 (±2.14)	0.34 (±0.22)
Ensemble	2.14 (±0.83)	2.75 (±1.00)	0.56 (±0.48)	0.61 (±0.25)	−0.16 (±0.73)	0.60 (±0.18)

Table A4. Evaluation of model performance during the training and testing phases considering the dry season datasets. Mean MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and predicted WT values.

Dry season	Train					
Model/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
RF	0.87 (±0.28)	1.13 (±0.34)	0.91 (±0.09)	0.83 (±0.88)	0.09 (±0.28)	0.95 (±0.02)
ANN	0.33 (±0.30)	0.47 (±0.41)	0.98 (±0.03)	0.97 (±0.03)	0.01 (±0.03)	0.98 (±0.03)
SVR	0.84 (±0.54)	1.20 (±0.68)	0.89 (±0.07)	0.86 (±0.10)	0.07 (±0.15)	0.91 (±0.06)
Air2stream (3 par)	2.93 (±0.95)	3.67 (±1.08)	0.21 (±0.25)	0.23 (±0.34)	0.30 (±0.45)	0.26 (±0.25)
Air2stream (4 par)	2.96 (±0.93)	3.69 (±1.06)	0.21 (±0.25)	0.23 (±0.32)	0.37 (±0.52)	0.27 (±0.25)
Air2stream (5 par)	2.81 (±0.95)	3.55 (±1.07)	0.25 (±0.24)	0.23 (±0.31)	0.04 (±0.19)	0.28 (±0.24)
Air2stream (7 par)	2.80 (±0.92)	3.55 (±1.05)	0.27 (±0.24)	0.29 (±0.30)	0.13 (±0.28)	0.29 (±0.23)
Air2stream (8 par)	2.82 (±0.92)	3.55 (±1.04)	0.27 (±0.24)	0.30 (±0.30)	0.19 (±0.32)	0.29 (±0.24)
MR	2.55 (±0.80)	3.22 (±0.96)	0.37 (±0.27)	0.39 (±0.24)	0.13 (±0.19)	0.41 (±0.22)
Ensemble	0.31 (±1.12)	0.44 (±1.37)	0.98 (±0.37)	0.97 (±0.33)	0.01 (±0.22)	0.98 (±0.34)
Dry season	Test					
Model/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
RF	2.37 (±1.17)	3.01 (±1.30)	0.33 (±0.62)	0.55 (±0.24)	0.29 (±1.55)	0.57 (±0.22)
ANN	2.19 (±0.93)	2.80 (±1.10)	0.31 (±0.71)	0.57 (±0.31)	0.01 (±1.03)	0.54 (±0.22)
SVR	2.39 (±0.95)	3.02 (±1.06)	0.37 (±0.34)	0.50 (±0.22)	0.29 (±1.04)	0.52 (±0.22)
Air2stream (3 par)	3.29 (±1.27)	4.12 (±1.32)	−0.13 (±0.47)	0.09 (±0.35)	0.30 (±1.73)	0.21 (±0.23)
Air2stream (4 par)	3.65 (±2.57)	4.49 (±2.51)	−0.28 (±0.87)	0.11 (±0.34)	0.79 (±3.20)	0.24 (±0.26)
Air2stream (5 par)	3.69 (±1.35)	4.48 (±1.38)	−0.41 (±1.00)	0.04 (±0.33)	0.79 (±2.15)	0.18 (±0.22)
Air2stream (7 par)	3.77 (±2.55)	4.58 (±2.50)	−0.29 (±0.75)	0.06 (±0.31)	0.57 (±3.23)	0.17 (±0.21)
Air2stream (8 par)	3.97 (±2.66)	4.84 (±2.67)	−0.59 (±1.78)	0.06 (±0.35)	0.75 (±3.36)	0.18 (±0.22)
MR	3.39 (±2.58)	4.21 (±2.71)	0.21 (±0.35)	0.22 (±0.33)	−0.44 (±3.26)	0.30 (±0.22)
Ensemble	1.98 (±0.96)	2.51 (±1.08)	0.50 (±0.55)	0.63 (±0.28)	0.12 (±1.17)	0.63 (±0.21)

Table A5. Evaluation of model performance during the training and testing phases considering the wet season datasets. Mean MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and predicted WT values.

Wet season	Train					
Model/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
RF	0.84 (±0.27)	1.11 (±0.33)	0.91 (±0.06)	0.80 (±0.09)	−0.07 (±0.15)	0.94 (±0.04)
ANN	0.25 (±0.28)	0.37 (±0.40)	0.98 (±0.04)	0.98 (±0.04)	0.01 (±0.03)	0.98 (±0.03)
SVR	0.75 (±0.53)	1.06 (±0.66)	0.91 (±0.07)	0.88 (±0.10)	−0.03 (±0.17)	0.92 (±0.06)
Air2stream (3 par)	2.72 (±0.93)	3.57 (±1.07)	0.15 (±0.26)	0.14 (±0.32)	−0.22 (±0.35)	0.20 (±0.22)
Air2stream (4 par)	2.70 (±0.92)	3.55 (±1.06)	0.15 (±0.26)	0.18 (±0.31)	−0.28 (±0.40)	0.20 (±0.22)
Air2stream (5 par)	2.64 (±0.95)	3.48 (±1.11)	0.20 (±0.25)	0.18 (±0.29)	−0.02 (±0.16)	0.23 (±0.24)
Air2stream (7 par)	2.56 (±0.96)	3.41 (±1.14)	0.24 (±0.25)	0.24 (±0.29)	−0.10 (±0.25)	0.27 (±0.25)
Air2stream (8 par)	2.55 (±0.97)	3.38 (±1.16)	0.25 (±0.26)	0.27 (±0.31)	−0.14 (±0.27)	0.28 (±0.25)
MR	2.58 (±0.89)	3.40 (±1.09)	0.30 (±0.28)	0.32 (±0.28)	−0.11 (±0.18)	0.27 (±0.23)
Ensemble	0.23 (±1.06)	0.35 (±1.37)	0.99 (±0.39)	0.98 (±0.36)	0.01 (±0.19)	0.99 (±0.37)
Wet season	Test					
Model/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
RF	2.38 (±1.07)	3.04 (±1.19)	0.13 (±1.91)	0.46 (±0.26)	−0.36 (±1.37)	0.49 (±0.23)
ANN	2.38 (±1.04)	3.03 (±1.26)	0.10 (±1.22)	0.48 (±0.36)	−0.23 (±1.06)	0.48 (±0.23)
SVR	2.52 (±0.94)	3.20 (±1.12)	0.13 (±1.10)	0.37 (±0.26)	−0.42 (±0.96)	0.40 (±0.23)
Air2stream (3 par)	3.13 (±1.47)	3.95 (±1.55)	0.02 (±0.29)	0.14 (±0.30)	−0.42 (±1.92)	0.25 (±0.22)
Air2stream (4 par)	3.15 (±1.29)	4.01 (±1.40)	−0.14 (±1.01)	0.14 (±0.29)	−0.49 (±1.77)	0.24 (±0.23)
Air2stream (5 par)	3.36 (±1.09)	4.13 (±1.18)	−0.19 (±0.64)	0.06 (±0.32)	−0.81 (±1.65)	0.21 (±0.22)
Air2stream (7 par)	3.85 (±1.23)	4.81 (±1.45)	−0.80 (±1.41)	0.01 (±0.30)	−1.27 (±2.15)	0.15 (±0.20)
Air2stream (8 par)	3.99 (±1.37)	5.10 (±1.92)	−1.27 (±3.46)	−0.04 (±0.48)	−1.25 (±2.18)	0.13 (±0.19)
MR	3.55 (±2.00)	4.42 (±2.22)	0.13 (±0.35)	0.13 (±0.36)	−0.28 (±2.61)	0.20 (±0.23)
Ensemble	2.09 (±0.86)	2.65 (±1.04)	0.31 (±0.78)	0.52 (±0.28)	−0.33 (±1.07)	0.55 (±0.18)

Table A6. Evaluation of RF model performance during the training (raw datasets) and testing (raw datasets) phases considering the annual datasets. Mean MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and predicted WT values.

Annual	Train (raw datasets)						Test (raw datasets)					
Station/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
1	1.18	1.40	0.96	0.80	0.15	0.99	3.53	5.03	−0.01	0.33	1.56	0.17
	(±1.77)	(±2.13)	(±0.40)	(±0.37)	(±0.15)	(±0.29)	(±0.56)	(±0.48)	(±0.22)	(±0.20)	(±1.02)	(±0.06)
7	1.07	1.55	0.84	0.69	−0.08	0.91	2.26	3.19	0.46	0.50	0.01	0.46
	(±0.65)	(±0.61)	(±0.20)	(±0.19)	(±0.08)	(±0.11)	(±0.36)	(±0.28)	(±0.11)	(±0.17)	(±0.16)	(±0.09)
12	0.78	0.97	0.92	0.81	0.03	0.95	2.98	3.53	0.15	0.35	−1.56	0.32
	(±0.22)	(±0.30)	(±0.07)	(±0.06)	(±0.03)	(±0.07)	(±0.06)	(±0.10)	(±0.05)	(±0.08)	(±0.05)	(±0.05)
13	0.83	1.04	0.86	0.71	0.04	0.93	2.51	3.09	0.29	0.20	0.28	0.50
	(±0.20)	(±0.26)	(±0.09)	(±0.08)	(±0.03)	(±0.08)	(±0.09)	(±0.09)	(±0.05)	(±0.04)	(±0.15)	(±0.09)
22	1.22	1.66	0.91	0.79	−0.19	0.94	2.57	3.30	0.45	0.57	−1.67	0.60
	(±0.69)	(±0.49)	(±0.08)	(±0.07)	(±0.03)	(±0.07)	(±0.09)	(±0.11)	(±0.04)	(±0.05)	(±0.06)	(±0.05)
29	0.61	0.88	0.95	0.88	−0.02	0.96	1.47	2.32	0.62	0.60	−0.16	0.65
	(±0.16)	(±0.19)	(±0.03)	(±0.03)	(±0.01)	(±0.03)	(±0.06)	(±0.06)	(±0.02)	(±0.03)	(±0.05)	(±0.02)
30	0.85	0.92	0.92	0.85	0.01	0.93	1.56	2.31	0.44	0.67	−1.09	0.56
	(±0.16)	(±0.20)	(±0.05)	(±0.04)	(±0.01)	(±0.05)	(±0.06)	(±0.06)	(±0.03)	(±0.03)	(±0.05)	(±0.03)
46	0.69	0.94	0.96	0.91	0.01	0.97	2.79	3.63	0.60	0.61	0.32	0.61
	(±0.25)	(±0.34)	(±0.04)	(±0.06)	(±0.02)	(±0.03)	(±0.09)	(±0.14)	(±0.03)	(±0.07)	(±0.54)	(±0.02)
59	0.65	0.83	0.96	0.90	0.01	0.97	1.20	1.57	0.84	0.90	0.33	0.85
	(±0.18)	(±0.25)	(±0.03)	(±0.04)	(±0.02)	(±0.03)	(±0.06)	(±0.11)	(±0.02)	(±0.05)	(±0.06)	(±0.02)
60	0.71	0.92	0.96	0.90	−0.03	0.97	2.00	2.48	0.70	0.78	0.85	0.74
	(±0.23)	(±0.30)	(±0.04)	(±0.06)	(±0.02)	(±0.03)	(±0.05)	(±0.07)	(±0.02)	(±0.06)	(±0.13)	(±0.02)
73	0.84	1.16	0.92	0.81	0.04	0.95	1.92	2.47	0.58	0.63	0.28	0.58
	(±0.24)	(±0.27)	(±0.05)	(±0.05)	(±0.01)	(±0.04)	(±0.14)	(±0.12)	(±0.04)	(±0.07)	(±0.05)	(±0.04)
83	0.65	0.88	0.95	0.87	0.00	0.96	1.61	2.07	0.69	0.72	−0.13	0.69
	(±0.15)	(±0.20)	(±0.03)	(±0.03)	(±0.02)	(±0.03)	(±0.08)	(±0.12)	(±0.04)	(±0.05)	(±0.04)	(±0.04)
Average	0.84	1.09	0.93	0.83	0.00	0.95	2.20	2.92	0.48	0.57	−0.08	0.56
	(±0.21)	(±0.28)	(±0.04)	(±0.07)	(±0.08)	(±0.02)	(±0.70)	(±0.92)	(±0.24)	(±0.20)	(±0.95)	(±0.18)

Table A7. Evaluation of RF model performance during the training (modified datasets) and testing (raw datasets) phases considering the annual datasets. Mean MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and predicted WT values.

Annual	Train (modified datasets)						Test (raw datasets)					
Station/metric	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2	MAE (°C)	RMSE (°C)	NSE	KGE	Bias (°C)	R^2
1	2.99	3.20	0.77	0.77	0.00	0.77	3.44	5.00	−0.04	0.37	1.26	0.17
	(±1.24)	(±1.47)	(±0.31)	(±0.30)	(±0.15)	(±0.15)	(±0.34)	(±0.31)	(±0.14)	(±0.22)	(±1.28)	(±0.07)
7	0.99	2.35	0.49	0.58	0.00	0.49	1.86	2.62	0.64	0.65	0.16	0.65
	(±0.45)	(±0.66)	(±0.09)	(±0.12)	(±0.07)	(±0.08)	(±0.43)	(±0.38)	(±0.14)	(±0.12)	(±0.62)	(±0.11)
12	0.00	0.00	1.00	1.00	0.00	1.00	2.54	3.25	0.28	0.23	−0.54	0.37
	(±0.38)	(±0.57)	(±0.08)	(±0.11)	(±0.05)	(±0.07)	(±0.62)	(±0.78)	(±0.50)	(±0.21)	(±0.98)	(±0.12)
13	0.98	1.31	0.79	0.75	−0.04	0.80	1.78	2.55	0.52	0.51	−0.86	0.63
	(±0.33)	(±0.45)	(±0.13)	(±0.15)	(±0.05)	(±0.11)	(±0.50)	(±0.54)	(±0.29)	(±0.34)	(±1.08)	(±0.19)
22	0.09	0.17	1.00	0.99	0.00	1.00	2.47	3.23	0.47	0.54	−1.66	0.65
	(±0.69)	(±0.89)	(±0.11)	(±0.15)	(±0.07)	(±0.09)	(±0.78)	(±0.90)	(±0.43)	(±0.15)	(±1.57)	(±0.15)
29	0.47	0.72	0.96	0.97	0.00	0.96	1.36	1.81	0.77	0.85	−0.25	0.77
	(±0.36)	(±0.54)	(±0.06)	(±0.08)	(±0.05)	(±0.05)	(±0.41)	(±0.52)	(±0.21)	(±0.13)	(±1.00)	(±0.13)
30	0.33	0.57	0.98	0.98	0.00	0.98	1.47	2.24	0.47	0.75	−0.43	0.57
	(±0.29)	(±0.47)	(±0.07)	(±0.09)	(±0.03)	(±0.06)	(±0.47)	(±0.50)	(±0.32)	(±0.11)	(±0.71)	(±0.09)
46	1.11	1.68	0.93	0.90	−0.09	0.94	2.15	2.95	0.73	0.73	−0.92	0.77
	(±0.40)	(±0.58)	(±0.05)	(±0.06)	(±0.05)	(±0.04)	(±0.66)	(±0.85)	(±0.23)	(±0.13)	(±1.50)	(±0.11)
59	0.51	0.78	0.96	0.90	−0.01	0.97	1.09	1.43	0.87	0.91	0.11	0.87
	(±0.33)	(±0.44)	(±0.04)	(±0.07)	(±0.05)	(±0.04)	(±0.40)	(±0.52)	(±0.16)	(±0.12)	(±0.67)	(±0.10)
60	0.73	1.02	0.96	0.91	−0.02	0.97	1.83	2.41	0.72	0.83	0.60	0.74
	(±0.35)	(±0.47)	(±0.04)	(±0.05)	(±0.03)	(±0.04)	(±0.36)	(±0.53)	(±0.17)	(±0.11)	(±0.99)	(±0.12)
73	0.84	1.16	0.92	0.81	0.04	0.95	1.92	2.47	0.58	0.69	0.28	0.58
	(±0.39)	(±0.52)	(±0.06)	(±0.08)	(±0.05)	(±0.04)	(±0.39)	(±0.47)	(±0.21)	(±0.13)	(±1.17)	(±0.11)
83	0.65	1.06	0.91	0.84	0.01	0.93	1.61	2.05	0.69	0.73	−0.34	0.70
	(±0.28)	(±0.37)	(±0.04)	(±0.06)	(±0.03)	(±0.03)	(±0.41)	(±0.49)	(±0.21)	(±0.14)	(±0.72)	(±0.10)
Average	0.81	1.17	0.89	0.87	−0.01	0.90	1.96	2.67	0.56	0.65	−0.22	0.62
	(±0.77)	(±0.90)	(±0.15)	(±0.12)	(±0.03)	(±0.15)	(±0.63)	(±0.91)	(±0.25)	(±0.20)	(±0.77)	(±0.19)

Table A8. SMOGN parameters for the best RF predictions.

Station	k	samp_method	rel_thres	rel_coef	rel_xtrm_type
1	4.0	extreme	0.53	0.02	high
7	3.0	extreme	0.46	0.36	both
12	2.0	extreme	0.29	0.17	both
13	7.0	extreme	0.81	0.02	high
22	3.0	balance	0.46	0.01	both
29	7.0	extreme	0.63	0.14	high
30	6.0	extreme	0.98	0.40	both
46	2.0	balance	0.62	0.29	both
59	5.0	extreme	0.40	0.10	both
60	5.0	extreme	0.77	0.17	both
73	5.0	extreme	0.53	0.38	both
83	5.0	extreme	0.04	0.28	both

Table A9. Evaluation of the random forest performance during the training and testing phases considering the annual datasets and the sequential increase in the models' predictors. Mean MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and predicted WT values. (1) Mean air temperature; (2) mean air temperature + discharge; (3) mean air temperature + discharge + radiation; (4) mean air temperature + discharge + radiation + maximum air temperature; (5) mean air temperature + discharge + radiation + maximum air temperature + minimum air temperature; (6) mean air temperature + discharge + radiation + maximum air temperature + minimum air temperature + MOY; (7) mean air temperature + discharge + radiation + maximum air temperature + minimum air temperature + MOY + DOY.

Annual	Train						
Metric/predictor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
MAE (°C)	1.84 (±0.51)	1.61 (±0.45)	1.50 (±0.41)	1.48 (±0.40)	1.44 (±0.40)	1.41 (±0.40)	1.07 (±0.30)
RMSE (°C)	2.35 (±0.57)	2.09 (±0.51)	1.98 (±0.47)	1.94 (±0.47)	1.90 (±0.46)	1.86 (±0.46)	1.43 (±0.38)
NSE	0.72 (±0.10)	0.78 (±0.09)	0.80 (±0.07)	0.81 (±0.07)	0.82 (±0.07)	0.82 (±0.07)	0.89 (±0.06)
KGE	0.67 (±0.12)	0.70 (±0.11)	0.71 (±0.11)	0.71 (±0.10)	0.72 (±0.10)	0.72 (±0.10)	0.82 (±0.09)
Bias (°C)	0.00 (±0.11)	0.01 (±0.09)	0.01 (±0.08)	0.01 (±0.09)	0.01 (±0.11)	0.00 (±0.08)	0.00 (±0.08)
R^2	0.76 (±0.08)	0.82 (±0.0.7)	0.85 (±0.05)	0.86 (±0.04)	0.87 (±0.04)	0.87 (±0.05)	0.92 (±0.04)
Annual	Test						
Metric/predictor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
MAE (°C)	3.55 (±0.97)	3.43 (±1.01)	3.37 (±1.10)	3.35 (±1.08)	3.35 (±1.10)	3.29 (±1.10)	2.51 (±0.95)
RMSE (°C)	4.54 (±1.18)	4.40 (±1.17)	4.30 (±1.19)	4.29 (±1.19)	4.30 (±1.23)	4.23 (±1.21)	3.29 (±1.12)
NSE	0.03 (±0.35)	0.08 (±0.34)	0.12 (±0.35)	0.13 (±0.34)	0.13 (±0.34)	0.16 (±0.33)	0.48 (±0.26)
KGE	0.32 (±0.25)	0.32 (±0.26)	0.31 (±0.28)	0.32 (±0.27)	0.31 (±0.28)	0.33 (±0.28)	0.60 (±0.20)
Bias (°C)	−0.15 (±1.33)	−0.26 (±1.37)	−0.25 (±1.18)	−0.22 (±1.21)	−0.22 (±1.26)	−0.21 (±1.21)	−0.10 (±1.25)
R^2	0.23 (±0.20)	0.26 (±0.21)	0.28 (±0.22)	0.28 (±0.23)	0.28 (±0.23)	0.29 (±0.23)	0.58 (±0.18)

Code and data availability. The Python code used to generate all results for this publication and the Fortran code of the Air2stream model can be found in Almeida and Coelho (2023; https://doi.org/10.5281/zenodo.7870379). Additionally, this repository includes the input data considered in this study (83 datasets). It is also possible to download the code and data from https://github.com/mcvta/WaterPythonTemp (last access: 17 July 2023).

Author contributions. MA conceived the study, performed the simulations and wrote the paper. PC contributed to the study design and to the results analysis. All authors contributed to the discussion and paper revision. All authors read and approved the final paper.

Competing interests. The contact author has declared that neither of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. We thank the anonymous reviewers for their careful reading of our paper and their many insightful comments and suggestions.

Financial support. The authors received funding from the Fundação para a Ciência e a Tecnologia (FCT, Portugal) through the strategic projects UIDB/04292/2020 and UIDP/04292/2020 granted to MARE and the project LA/P/0069/2020 granted to the Associate Laboratory ARNET.

Review statement. This paper was edited by Andrew Wickert and reviewed by three anonymous referees.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., 620 Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous distributed systems, in: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016, 265–283, https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf (last access: 17 July 2023), 2016.

Agrawal, A. and Petersen, M. R.: Detecting Arsenic Contamination Using Satellite Imagery and Machine Learning, Toxics, 9, 333, https://doi.org/10.3390/toxics9120333, 2021.

Ahmadi-Nedushan, B., St-Hilaire, A., Ouarda, T. B. M. J., Bilodeau, L., Robichaud, É., Thiémonge, N., and Bobée, B.: Predicting river water temperatures using stochastic models: case study of the Moisie River (Québec, Canada), Hydrol. Process., 21, 21–34, https://doi.org/10.1002/hyp.6353, 2007.

Almeida, M. C. and Coelho, P. S.: mcvta/WaterPythonTemp: Release 0.2.0, Zenodo [code and data set], https://doi.org/10.5281/zenodo.7870379, 2023.

Araújo, C. S. P., Silva, I. A. C., Ippolito, M., and Almeida, C. D.: Evaluation of air temperature estimated by ERA5-Land reanalysis using surface data in Pernambuco, Brazil. Environ. Monit. Assess., 194, 381, https://doi.org/10.1007/s10661-022-10047-2, 2022.

Awad, M. and Khanna, R. (Eds.): Support vector regression, in: Efficient learning machines, Springer, 67–80, https://doi.org/10.1007/978-1-4302-5990-9, 2015.

Bergstra, J., Yamins, D., and Cox, D. D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, in: TProc. of the 30th International Conference on Machine Learning (ICML 2013), 11523, https://doi.org/10.5555/3042817.3042832 (last access: 17 July 2023), 2013.

Branco, P., Ribeiro, R. P., and Torgo, L.: UBL: an R package for utility-based learning. arXiv preprint, arXiv:1604.08079, https://doi.org/10.48550/arXiv.1604.08079, 2016.

Branco, P., Ribeiro, R. P., Torgo, L., Krawczyk, B., and Moniz, N.: Smogn: a pre-processing approach for imbalanced regression, Proc. Mach. Learn. Res., 74, 36–50, 2017.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Caissie, D.: The thermal regime of rivers: a review, Freshwater Biol., 51, 1389–1406, https://doi.org/10.1111/j.1365-2427.2006.01597.x, 2006.

Cardoso, R. M., Soares, P. M. M., Miranda, P. M. A., and Belo-Pereira, M.: WRF High resolution simulation of Iberian mean and extreme precipitation climate, Int. J. Climatol., 33, 2591–2608, https://doi.org/10.1002/joc.3616, 2013.

Chenard, J.-F. and Caissie, D.: Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada, Hydrol. Process., 22, 3361–3372, https://doi.org/10.1002/hyp.6928, 2008.

Crisp, D. T. and Howson, G.: Effect of air temperature upon mean water temperature in streams in the north Pennines and English Lake District, Freshwater Biol., 12, 359–367, https://doi.org/10.1111/j.1365-2427.1982.tb00629.x, 1982.

DeWeber, J. T. and Wagner, T.: A regional neural network ensemble for predicting mean daily river water temperature, J. Hydrol., 517, 187–200, https://doi.org/10.1016/j.jhydrol.2014.05.035, 2014.

Du, X., Shrestha, N. K., Ficklin, D. L., and Wang, J.: Incorporation of the equilibrium temperature approach in a Soil and Water Assessment Tool hydroclimatological stream temperature model, Hydrol. Earth Syst. Sci., 22, 2343–2357, https://doi.org/10.5194/hess-22-2343-2018, 2018.

Ducharne, A.: Importance of stream temperature to climate change impact on water quality, Hydrol. Earth Syst. Sci., 12, 797–810, https://doi.org/10.5194/hess-12-797-2008, 2008.

Fahrer, C. and Harris, D., 2004. LAMPPOST A Mnemonic device for teaching climate variables, J. Geogr., 103, 86–90, https://doi.org/10.1080/00221340408978579, 2004.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The Shuttle Radar Topography Mission, Rev. Geophys., 45, RG2004, https://doi.org/10.1029/2005RG000183, 2007.

Feigl, M., Lebiedzinski, K., Herrnegger, M., and Schulz, K.: Machine-learning methods for stream water temperature prediction, Hydrol. Earth Syst. Sci., 25, 2951–2977, https://doi.org/10.5194/hess-25-2951-2021, 2021.

Gallice, A., Schaefli, B., Lehning, M., Parlange, M. B., and Huwald, H.: Stream temperature prediction in ungauged basins: review of recent approaches and description of a new physics-derived statistical model, Hydrol. Earth Syst. Sci., 19, 3727–3753, https://doi.org/10.5194/hess-19-3727-2015, 2015.

Jeppesen, E. and Iversen, T. M.: Two Simple Models for Estimating Daily Mean Water Temperatures and Diel Variations in a Danish Low Gradient Stream, Oikos, 49, 149–155, https://doi.org/10.2307/3566020, 1987.

Jourdonnais, J. H., Walsh, R. P., Pickett, F., and Goodman D.: Structure and calibration strategy for a water temperature model of the lower Madison River, Montana, Rivers, 3, 153–169, 1992.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Louppe, G.: Understanding Random Forests: From Theory to Practice, arXiv 2014, arXiv:1407.7502, 28 July 2014.

Lu, H. and Ma, X.: Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere, 249, 126169, https://doi.org/10.1016/j.chemosphere.2020.126169, 2020.

Macedo, M. N., Coe, M. T., DeFries, R., Uriarte, M., Brando, P. M., Neill, C., and Walker, W. S.: Land-use-driven stream warming in southeastern Amazonia, Philos. T. R. Soc. B, 368, 1619, https://doi.org/10.1098/rstb.2012.0153, 2013.

Mackey, A. P. and Berrie, A. D.: The prediction of water temperatures in chalk streams from air temperatures, Hydrobiologia, 210, 183–189, https://doi.org/10.1007/BF00034676, 1991.

Mohseni, O., Stefan, H. G., and Erickson, T. R.: A nonlinear regression model for weekly stream temperatures, Water Resour. Res., 10, 2685–2692, https://doi.org/10.1029/98WR01877, 1998.

Mohseni, O., Erickson T. R., and Stefan, H. G.: Upper bounds for stream temperatures in the contiguous United States, Journal of Environmental Engineering, Am. Soc. Civil Eng., 128, 4–11, https://doi.org/10.1029/98WR01877, 2002.

Moore, R. D., Spittlehouse, D. L., and Story, A.: Riparian microclimate and stream temperature response to forest harvesting: a review, J. Am. Water Resour. As., 41, 813–834, https://doi.org/10.1111/j.1752-1688.2005.tb03772.x, 2005.

Moore, R. D., Nelitz, M., and Parkinson, E.: Empirical modelling of maximum weekly average stream temperature in British Columbia, Canada, to support assessment of fish habitat suitability, Can. Water Resour. J., 38, 135–147, https://doi.org/10.1080/07011784.2013.794992, 2013.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, T. ASABE, 50, 885–900, https://doi.org/10.13031/2013.23153, 2007.

Muñoz-Sabater, J.: ERA5-Land hourly data from 1981 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.e2161bac, 2019.

Muñoz-Sabater, J.: ERA5-Land hourly data from 1950 to 1980, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.e2161bac, 2021.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models: Part 1. A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Neumann, D., W., Balaji, R., and Zagona E., A.: 2011, Regression model for daily maximum stream temperature, J. Environ. Eng., 129, 667–674, https://doi.org/10.1061/(ASCE)0733-9372(2003)129:7(667), 2003.

Pedregosa, F., Varoquaux, G., Gramfort, A, Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, https://doi.org/10.48550/arXiv.1201.0490, 2011.

Piccolroaz, S.: Prediction of lake surface temperature using air 2stream model: guidelines, challenges, and future perspectives, Adv. Oceanogr. Limnol., 7, 36–50, https://doi.org/10.4081/aiol.2016.5791, 2016.

Piotrowski, A. P., Napiorkowski, M. J., Napiorkowski, J. J., and Osuch, M.: Comparing various artificial neural types for water temperature prediction in rivers, J. Hydrol., 529, 302–315, https://doi.org/10.1016/j.jhydrol.2015.07.044, 2015.

Platt, J.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft, Open File Rep., MSR-TR-98-14, 98–14, https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/ (last access: 19 July 2023), 1998.

Qian, N.: On the momentum term in gradient descent learning algorithms, Neural Networks, 12, 145–151, https://doi.org/10.1016/S0893-6080(98)00116-6, 1999.

Rabi, A., Hadzima-Nyarko, M., and Šperac, M.: Modelling river temperature from air temperature: case of the River Drava (Croatia), Hydrol. Sci. J., 60, 1490–1507, https://doi.org/10.1080/02626667.2014.914215, 2015.

Rajesh, M. and Rehana, S.: Prediction of river water temperature using machine learning algorithms: a tropical river system of India, J. Hydroinform., 23, 605–626, https://doi.org/10.2166/hydro.2021.121, 2021.

Rehana, S.: River water temperature modelling under climate change using support vector regression, in: Hydrology in a Changing World: Challenges in Modeling, edited by: Singh, S. K. and Dhanya, C. T., World Springer, 171–183, https://doi.org/10.1007/978-3-030-02197-9_8, 2019.

Rehana, S. and Mujumdar, P. P.: River water quality response under hypothetical climate change scenarios in Tunga-Bhadra river, India, Hydrol. Process., 25, 3373–3386, https://doi.org/10.1002/hyp.8057, 2011.

Segura, C., Caldwell, P., Sun, G., McNulty, S., and Zhang, Y.: A model to predict stream water temperature across the conterminous USA, Hydrol. Process., 29, 2178–2195, https://doi.org/10.1002/hyp.10357, 2014.

Shevchuk, Y.: Python library, https://neupy.com/pages/home.html, last access: 7 July 2022.

Shi, Y. and Eberhart, R. C.: A modified particle swarm optimizer, in: International Conference on Evolutionary Computation Proceedings, IEEE World Congress on Computational Intelligence (Cat. No.98TH8360), 69–73, https://doi.org/10.1109/ICEC.1998.699146, 1998.

Sinokrot, B. A. and Stefan, H. G.: Stream temperature dynamics: measurements and modeling, Water Resour. Res., 29, 2299–2312, https://doi.org/10.1029/93WR00540, 1993.

Smith, K.: River water temperatures - an environmental review, Scottish Geographical Magazine, 88, 211–220, https://doi.org/10.1080/00369227208736229, 1972.

Smith K.: The prediction of river water temperatures, Hydrol. Sci. Bull., 26, 19–32, https://doi.org/10.1080/02626668109490859, 1981.

Smith, K. and Lavis, M. E.: Environmental influences on the temperature of a small upland stream, Oikos, 26, 228–236, 1975.

SMOGN: Synthetic Minority Oversampling Technique for Regression with Gaussian Noise, GitHub repository [code], https://github.com/nickkunz/smogn (last access: 17 July 2023), 2022.

Smola, A. J. and Schölkopf, B. A.: tutorial on support vector regression, Stat. Comput., 14, 199–222, https://doi.org/10.1023/B:STCO.0000035301.49549.88, 2004.

Soares, P. M. M., Cardoso, R. M., Medeiros, J., Miranda, P. M. A., Belo-Pereira, M., and Espirito-Santo, F.: WRF High resolution dynamical downscaling of ERA-Interim for Portugal, Clim. Dynam., 39, 2497–2522, https://doi.org/10.1007/s00382-012-1315-2#Bib1, 2012.

Soares, P. M. M., Cardoso, R. M., Ferreira, J. J., and Miranda, P. M. A.: Climate change and the Portuguese precipitation: ENSEMBLES regional climate models results, Clim. Dynam., 45, 1771–1787, https://doi.org/10.1007/s00382-014-2432-x, 2015.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, 2014.

Temez, J. R.: Calculo hidrometeorologico de caudales maximos em pequeñas cuencas naturales, Madrid: Ministério de Obras 790 Publicas y Urbanismo (MOPU), Direccion General de Carreteras, 12, 1978.

Temizyurek, M. and Dadaser-Celik, F.: Modelling the effects of meteorological parameters on water temperature using artificial neural networks, Water Sci. Technol., 77, 1724–1733, https://doi.org/10.2166/wst.2018.058, 2018.

Toffolon, M. and Piccolroaz, S.: A hybrid model for river water temperature as a function of air temperature and discharge, Environ. Res. Lett., 10, 114011, https://doi.org/10.1088/1748-9326/10/11/114011, 2015.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P.: Smote for regression, in: Progress in Artificial Intelligence, Springer, 378–389, https://doi.org/10.1007/978-3-642-40669-0_33, 2013.

Torgo, L., Branco, P., Ribeiro, R. P., and Pfahringer, B.: Resampling strategies for regression, Expert Systems, 32, 465–476, https://doi.org/10.1111/exsy.12081, 2015.

Vanella D., Longo-Minnolo, G., Belfiore, O.R. Ramírez-Cuesta, J. M., Pappalardo, S., Consoli, S., D'Urso, G.,Chirico, G.B.,Coppola, A., Comegna, A. Toscano, T., Quarta, R., Provenzano, G., Ippolito, M., Castagna, A., and Gandolfi, C.: Comparing the use of ERA5 reanalysis dataset and ground-based agrometeorological data under different climates and topography in Italy, J. Hydrol., 42, 101182, https://doi.org/10.1016/j.ejrh.2022.101182, 2022.

Walling, D. E. and Webb, B. W.: Water quality. I. Physical characteristics, in: The rivers Handbook. I. Hydrological and ecological principles, edited by: Calow, P. and Petts, G. E., Blackwell Scientific Publs., Oxford, 48–72, ISBN 978-1-444-31386-4, 1993.

Wang, B., Spessa, A., Feng, P., Hou, X., Yue, C., Luo, J.-J., Ciais, P., Waters, C., Cowie, A., Nolan, R. H., Nikonovas, T., Jin, H., Walshaw, H., Wei, J., Guo, X., Liu, D. L., and Yu, Q.: Extreme Fire Weather Is The Major Driver Of Severe Bushfires In Southeast Australia, Sci. Bull., 67, 655–664, https://doi.org/10.1016/j.scib.2021.10.001, 2021.

Webb, B. W. and Nobilis, F.: A long-term perspective on the nature of the air-water temperature relationship: a case study, Hydrol. Process., 11, 137–147, https://doi.org/10.1002/(SICI)1099-1085(199702)11:2<137::AID-HYP405>3.0.CO;2-2, 1997.

Webb, B. W., Clark, P. D., and Walling, D. E.: Water-air temperature relationships in a Devon river system and the role of flow, Hydrol. Process., 17, 3069–3084, https://doi.org/10.1002/hyp.1280, 2003.

Wetzel, R. G. (Ed.): Limnology. Lake and River Ecosystems, Third Edition, Academic Press, ISBN 9780127447605, 2001.

Younus, M., Hondzo, M., and Engel, B. A.: Stream Temperature Dynamics in Upland Agricultural Watershed, J. Environ. Eng., 126, 518–526, https://doi.org/10.1061/(ASCE)0733-9372(2000)126:6(518), 2000.

Zhao, P. and He, Z.: A First Evaluation of ERA5-Land Reanalysis Temperature Product Over the Chinese Qilian Mountains, Front. Earth Sci., 10, 907730, https://doi.org/10.3389/feart.2022.907730, 2022.

Zhu, S., Nyarko, E. K., and Hadzima-Nyarko, M.: Modelling daily water temperature from air temperature for the Missouri River, PeerJ, 6, e4894, https://doi.org/10.7717/peerj.4894, 2018.

Zhu, S., Hadzima-Nyarko, M., Gao, A., Wang, F., Wu, J., and Wu, S.: Two hybrid data-driven models for modeling water-air temperature relationship in rivers, Environ. Sci. Pollut. R., 26, 12622–12630, https://doi.org/10.1007/s11356-019-04716-y, 2019a.

Zhu, S., Heddam, S., Nyarko, E. K., Hadzima-Nyarko, M., Piccolroaz, S., and Wu, S.: Modeling daily water temperature for rivers: comparison between adaptive neuro-fuzzy inference systems and artificial neural networks models, Environ. Sci. Pollut. Res., 26, 402–420, https://doi.org/10.1007/s11356-018-3650-2, 2019b.

Zhu, S., Heddam, S., Wu, S., Dai, J., and Jia, B.: Extreme learning machine-based prediction of daily water temperature for rivers, Environ. Earth Sci., 78, 202, https://doi.org/10.1007/s12665-019-8202-7, 2019c.

Zhu, S., Nyarko, E. K., Hadzima-Nyarko, M., Heddam, S., and Wu, S.: Assessing the performance of a suite of machine learning models for daily river water temperature prediction, PeerJ, 7, e7065, https://doi.org/10.7717/peerj.7065, 2019d.