Modeling river water temperature with limiting forcing data: air2stream v1.0.0, machine learning and multiple regression

Manuel Almeida[1], Pedro Coelho[2]

[1,2] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Mare – Centro de Ciências do Mar e do Ambiente, Lisboa, 2825 - 516, Portugal.

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

Abstract.

The prediction of river water temperature (WT) is of key importance in the field of environmental science. Water temperature datasets for low order rivers are often in short supply, leaving lake/reservoir water quality modelers with the challenge of extracting as much information as possible from existing datasets, usually without the use of physically based models, due to the significant amount of data required (e.g., river morphology, degree of shading, wind velocity). Commonly, the WT observed in monitoring stations located near the downstream section of rivers are assumed to be the boundary condition of lake/reservoir water quality models. The main goal of this study is to identify a suitable WT modeling solution for these sections given the scarcity of the forcing datasets. In this study, five models are used to predict the water temperature of 83 rivers (with 98% missing data): three machine-learning (ML) algorithms (Random Forest, Artificial Neural Network and Support Vector Regression), the hybrid Air2stream model with all available parameterizations and a Multiple Regression. With the exception of Air2stream, which was forced with mean daily air temperature and discharge, all other models were forced with: mean, maximum, and minimum daily air temperature, mean daily total radiation (shortwave), mean daily discharge, month of the year and day of the year. The machine learning hyperparameters were optimized with a Tree-structured Parzen Estimators algorithm and the results of each model are presented as an ensemble of 12 individual optimized model runs. The meteorological datasets were obtained from the fifth-generation atmospheric reanalysis, ERA5. 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 WT values are limited, the application of all the models considered in this study is relevant (models ensemble mean annual – Root mean square error (RMSE): 2.75 ºC ± 1.00; Nash-Sutcliffe efficiency (NSE): 0.56 ± 0.48). Therefore, the datasets gaps can be filled with the best model of the ensemble approach. The model that performed best was Random Forest (annual mean - RMSE: 3.18 ºC ± 1.06; NSE: 0.52 ± 0.23). The results also revealed the existence of a logarithmic correlation among the RMSE between the observed and predicted river WT and the watershed time of concentration. The RMSE increases by an average of 0.1 ºC with a one-hour increase in the watershed time of concentration. (watershed area: μ= 106 km^2; σ=153). Hopefully, the high number of modeled sections considered in this model intercomparison study and the specific model forcing conditions will help to reduce the overall WT modeling uncertainty.

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) and the distribution and growth rate of aquatic organisms (e.g., primary production; fish growth and habitat) (Smith, 1972; Webb, 2003; Caissie, 2006). As such, the accurate prediction and assessment of river WT is a crucial part of many earth science applications. The thermal dynamics in rivers is quite complex as it depends 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 tributaries and groundwater) under specific meteorological and hydrological conditions (Walling and Webb, 1993; Wetzel, 2001). The complexity of river water temperature 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/reservoir water quality modeling results, namely, when lake/reservoir inflows are large. The fall and spring turnover onset, stratification strength/length and the overall heat budget can be affected. 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 water temperature with a time lag (Smith, 1981; Crisp and Howson, 1982). There are a number of examples in the existing literature of the successful implementation of linear regression models correlating air and water temperature 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 water temperature 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 not only affected 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 the linear regression (e.g., Neumann et al., 2003; Rehana, S. and Mujumdar, P. P., 2011), multiple regression (e.g., Jeppesen and Iversen, 1987; Jourdonnais et al., 1992), non-linear regression (e.g., Mohseni et al., 1998), stochastic regression models (e.g., Ahmadi-Nedushan et al., 2007; Rabi et al., 2015) and hybrid models (statistics methods combined with physical based process, e.g., Gallice et al., 2015; Toffolon and Piccolroaz, 2015). Machine-learning (ML) models, such as artificial neural networks (ANN), 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). 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 short-wave solar radiation, degree of shading and wind direction/velocity. The number and type of predictor variables considered to force river WT models in several intercomparison studies is quite different. 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 networks 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, non-linear regression and stochastic models (Zhu et al., 2018). This is not a prevailing rule, however, 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 ºC to 2.30 ºC in the case of the ML models. The worst results, as expected, correspond to the classical statistical models, namely multiple regression. From a water-quality modeling perspective, accurate time-varying boundary conditions are vital in order to calibrate lake or reservoir models. For water temperature calibration, this ideally means using continuous inflow temperatures, a condition that is very difficult to attain, as WT measurements are often scarce and rarely available, particularly for low-order streams.

Hence, the main objective of this study is to identify a suitable WT modeling solution to improve the lake/reservoir water quality models' boundary condition. It is important to mention that, for this study, an absence of significant variation between the water quality station observed WT and the WT at the downstream portion of a river was assumed, which coincides with the lake/reservoir water quality model boundary condition.

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 temperatures.

This is the primary and innovative objective of this study. The methodological approach was, therefore, defined to attempt to answer the following questions:

i) What is the best modeling solution to predict river water temperature with limiting forcing data? This modeling solution will be considered to improve the characterization of lake/reservoir WT boundary conditions (assuming that the observed WT of the downstream sections of the rivers are the boundary condition of lake/reservoir water quality models);
ii) How does the length of the calibration period and percentage of missing data affect model performance?
iii) Is it possible to relate the modeling error with 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: 3, 4, 5, 7 and 8 parameters) (Toffolon and Piccolroaz, 2015) and Multiple Regression (MR). The results of this study will hopefully prove useful from a practical perspective by helping to improve the quality of lake/reservoir model boundary conditions, as well as contributing to the overall model evaluation/development process.

Table 1: List of reviewed publications on river WT modelling and the corresponding RMSE between observed and modelled WT values

Reference	Geographic location	Number of sites	Temporal scale	Model type	RMSE (ºC)
Chenard and Caissie, 2008	Canada	1	day	ANN	0.96
DeWeber and Wagner, 2013	Eastern U.S.	96	day	ANN	1.82; 1.93
Rabi et al., 2015	Croatia	3	day	ANN	μ=1.70 σ=0.49; μ=2.06 σ=0.35; μ=2.30 σ=0.76
Zhu et al., 2019c	U.S.	3	day	ANN	0.768; 0.948; 1.242
Feigl et al., 2021	Austria; Germany and 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, U.S.	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 et al., 2020	U.S.	1	hour	RF	1.04
Feigl et al., 2021	Austria; Germany and 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., 2003	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] 5par[0.70, 1.05]; 7-par[0.65,0.78];8par[0.75,0.62] *
Zhu et al., 2019d	Europe, U.S.	8	day	Air2stream	3-par[0.64, 1.25]; 5-par[1.31, 0.76]; 8par[1.37;0.93] *
Feigl et al., 2021	Austria; Germany and Switzerland	10	day	Air2stream	8-par[0.74,1.17] *

*The model can be applied with 3, 4, 5, 7 or 8 parameters (3-par; 4-par; 5-par; 7-par and 8-par)

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 ºC in the central highlands to 25 ºC in the southeastern region. The minimum daily mean air temperature ranges from 5 ºC in the northern and central regions to 18 ºC in the south (Soares et al., 2012). The spatial and temporal heterogeneity of precipitation, which differs from a relatively wet annual maximum of over 2 500 mm/yr, in the mountainous northwest to a much drier 400 mm/yr. in the flat southeast, is defined by complex topography and coastal processes (Cardoso et al., 2013; Soares et al., 2015).

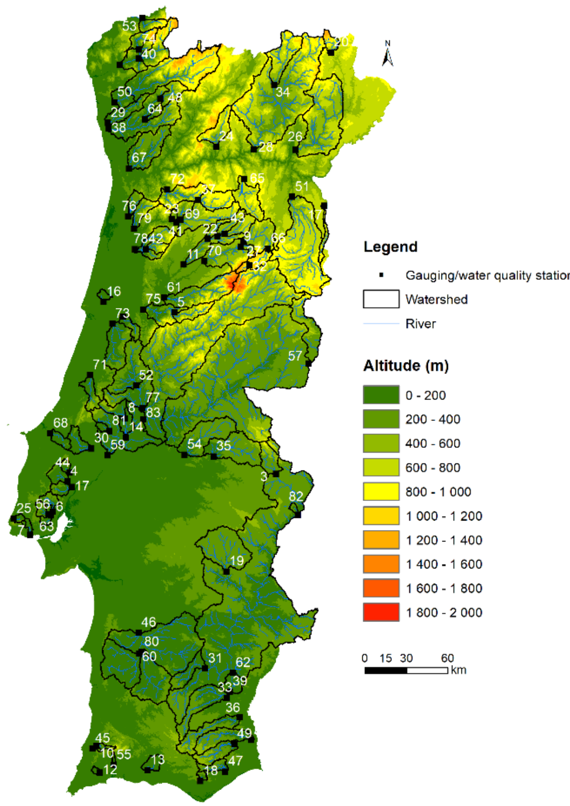


Figure 1: Location of the watersheds considered in the study (from DEM – Shuttle Radar Topography Mission (Farr et al., 2007))

The models used in this study were forced with daily mean, maximum and minimum air temperature and also global radiation values, obtained from the fifth-generation atmospheric reanalysis, ERA5, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). Hourly data on atmospheric parameters were obtained for the entire globe on a 0.25º latitude-longitude grid (Hersbach et al., 2020). The datasets are available from the Climate Data Store of the Copernicus Climate Change Service (http://cds.climate.copernicus.eu). The watershed discharge data used to force the models and the water temperature considered for the model's validation are also available from SNIRH (http://snirh.apambiente.pt). The SNIRH provides data and water temperature values for 2 471 water-quality stations, of which only 98 have gauging stations with discharge values, one of the conditions required to implement the Air2stream model. The missing discharge data was replaced with the corresponding climatological year value, hence only the gauging stations with data spanning at least a full year (365/366 values) were kept. Following this initial analysis, the number of sections 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 amount of daily river WT values are missing, ranging from 96.9% to 99.9% (μ= 98.8%; σ=0.68).

Table 2 shows the number of WT 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

Table 2: Number of WT for the annual, dry- and wet season, training and test data series

Temporal scale	Phase	Total number	Mean	Stdev	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

3 Methodology

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

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

ii)The number and type of models is 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;

iii) 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 and, as such, the forcing meteorological datasets considered in this study were obtained from the ERA5 reanalysis;

iv) There are no significant variations between the water quality station observed WT, and the WT at the downstream portion of a river, which coincides with the lake/reservoir water quality model boundary condition.

The modeling reference is the watershed main gauging station/water-quality station. Therefore, the hourly air temperature (ºC) and global radiation (shortwave) (Jm^{-2}) input datasets of the nearest ERA5 grid point were initially downloaded before the air temperature datasets were corrected according to the gauging station and the ERA5 grid point altitude. This correction was achieved by considering a linear variation of air temperature with the altitude, $\frac{dT}{dz} = -6.0, °C/km$ (Fahrer and Harris, 2004). After this correction, the mean, maximum and minimum daily air temperature values and the mean global radiation values where 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, max. and min. daily air temperature (ºC), mean daily total radiation (shortwave) (Jm^{-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)

Table 3: Model predictor variables

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

The results section starts with the evaluation of the ERA5 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 (*vide* 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 a test dataset (the remaining 30%). Due to size of the available datasets the validation phase was not considered.

Hyperparameter optimization was achieved for the ML models through the application of the Tree-structured Parzen Estimators (TPE) algorithm (*vide* Sect. 3.5). 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):

i) Considering the ordered length (L) of the 83 input datasets (training + testing), the datasets were divided into four different classes ($L \leq 50$; $50< L \leq 100$; $100< L \leq 200$; $L>200$);

ii) The ML and TPE algorithms were applied to the datasets corresponding to the minimum, mean and maximum number of values of each of the classes listed above. At this stage there are 12 model structures computed with the TPE algorithm for each ML model;

iii) The 12 models obtained for each ML were applied to the 83 datasets and the best performing model at each station was calculated based on the computed root mean square value (RMSE). Hence, the ensemble of the best results obtained across the 12 different models per station defines the overall ML results.

In other words, the ML hyperparameters are only optimized for 12 datasets and the ensemble of all the best models that were run are the ML results considered for the intercomparison of models.

Following this initial 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 of 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 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.

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 effect the river water temperature. The well-known Temez equation (Temez, J.R., 1978) (*vide* Sect. 3.7) initially defined for small-scale Mediterranean watersheds was selected for this analysis. Additionally, the Gaussian Mixture Models 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 as well as 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 results evaluation 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 provide not only a consistent interpretation of the models' results, but also to facilitate comparison with the results obtained in other studies (*vide* Sect. 3.8). The following sections describe each of the models and outline their relevant advantages/disadvantages. The ML parameters are listed in Table 3.

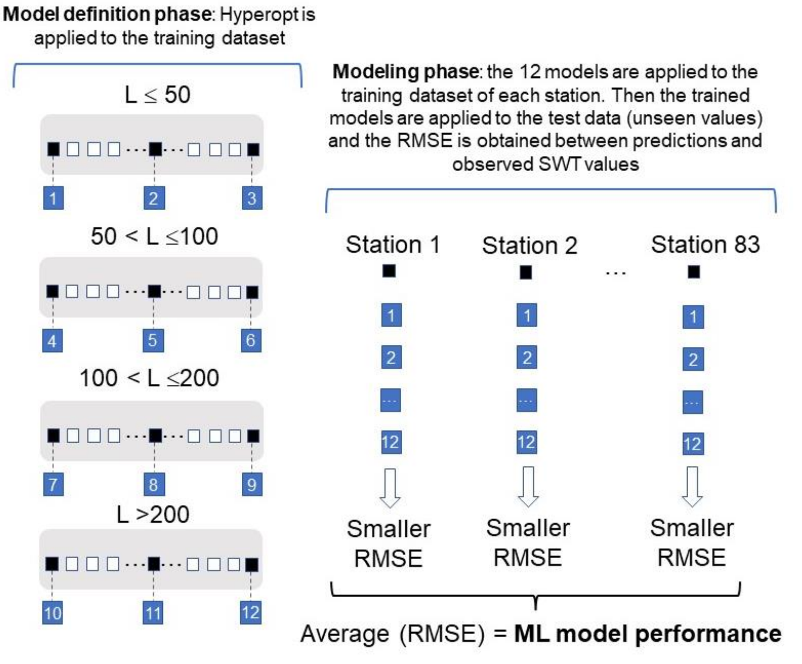


Figure 2: Schematic and simplified representation of the modeling process

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 sub-samples of the datasets and then combines the predictions. Decision trees can model complex non-linear 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).

3.2 Artificial Neural Network

The ANN prototyping and building was achieved with the NeuPy python library (Shevchuk, Y., 2015). This library uses Tensorflow (an open-source platform for machine learning) as a computational backend for deep learning models (Abadi et al., 2015). 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 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 water temperature (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 was 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 et al., 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 less 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, and the atmosphere (Toffolon and Piccolroaz, 2015). The river WT variation is described by the following equation:

$$\rho C_p V \frac{dT_W}{dt} = AH + \rho C_p \left(\sum_i Q_i T_{W,i} - Q T_w \right), \tag{1}$$

Where T_w is the water temperature of a river section with a volume V and surface area A, 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_{Wi} is the i-th water temperature 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. Eq. (2) is the simplified form of Eq.(1) (*vide* Toffolon and Piccolroaz, 2015). This equation, with 8 parameters, forms the basis of the Air2stream model:

$$\frac{dT_W}{dt} = \frac{1}{\theta^{a_4}}\left(a_1 + a_2\, T_a - a_3 T_w + \theta\left(a_5 + a_6 cos\left(2\pi\left(\frac{t}{t_y} - a_7\right)\right) - a_8 T_w\right)\right), \qquad (2)$$

Where Ta is the air temperature, θ is the dimensionless discharge ($\theta = Q/\bar{Q}$) (3), $\bar{Q}$ is the mean discharge. The parameter, a_4 is related with the exponent of the rating curve. The model is fitted to the entire input dataset (air temperature, water temperature 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 3, 4, 5, 7 and 8 parameter versions. Please refer to Toffolon and Piccolroaz (2015) for a full description of each one of the models' parameterizations.

3.5 Hyperparameter optimization

Hyperparameter optimization was achieved using the Tree-structured Parzen Estimators (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 set up 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 as it fixes some of the disadvantages associated with the latter. One notable drawback, however, is that this model 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 hyperot 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 as the algorithm score. The algorithm uses 20% of best observations to estimate the next set of parameters (gama = 0.2). Table 3 shows the models parameters and the corresponding optimization range. The algorithm was applied to the training data set. Table A1 shows the model parameters and the optimization range.

3.6 Multiple regression

This model was implemented using the machine learning python package, scikit-learn (Pedregosa et al., 2011). In this model the predicted value ($\hat{y}$) is expected to be a linear combination of the input features:

$$\hat{y}(w, x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p, \tag{4}$$

Where, x_i are the model input features, w_0 is the intercept and w_i the model coefficients. The model fits a linear model with coefficients w_i to minimize the residual sum of squares between the observed and predicted values.

3.7 Time of concentration

The time of concentration was estimated using the Temez equation (Temez, J.R., 1978), which was defined for small natural watersheds located in Spain.

$$T_C = 0.3 \left(\frac{L}{J^{1/4}}\right)^{0.76}, \tag{5}$$

T_C – time of concentration, h

L – length of the main water line, km

J – mean steepness (ratio between the mean fall and the L length of the water line), m/m

3.8 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 (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, $\bar{m}$ and $\bar{o}$ their means, σ_m is the standard deviation of the modeled values, σ_o the standard deviation of the observed values and r is the Pearson coefficient:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|m_i - o_i|, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2}, \tag{7}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N}(o_i - m_i)^2}{\sum_{i=1}^{N}(o_i - \bar{o}_i)^2}\right], \quad (8)$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_m}{\sigma_o} - 1\right)^2 + \left(\frac{\bar{m}}{\bar{o}} - 1\right)^2}, \quad (9)$$

$$Bias = \bar{m} - \bar{o}, \quad (10)$$

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{N}(m_i - \bar{o})^2}{\sum_{i=1}^{N}(o_i - \bar{o})^2} \times 100, \quad (11)$$

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

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2, \quad (12)$$

4 Results

4.1 Air temperature - ERA5 *versus* ground observed datasets

In this analysis the observed air temperature datasets of a total of eleven meteorological stations were considered. These are all available air temperature datasets observed within a 5 km radius of the stations considered in this study. Results show that the mean RMSE obtained between the two datasets considering all stations varied from 1.26ºC to 2.05ºC (μ=1.54ºC; σ=0.24ºC) and that, according to the mean bias values, the ERA tends to overestimate the observed air temperature datasets at 91% of the stations. Overall, a mean value of 1.54ºC (σ=0.24ºC) and a mean NSE value of 0.90 (σ=0.07) is indicative of a good performance. However, as shown in Figure 3, there are some sporadic significant discrepancies between the two datasets (Fig. 3). Additionally, results show that the stations with a RMSE higher than 2ºC are scattered all over the country. In this context it is relevant to mention that McNicholl et al. (2021) also found large biases between the ERA5 daily air temperature datasets and land data for a temperate region (Dublin) and for a tropical region (Singapore). Generally, these results suggest that the consideration of the ERA5 air temperature datasets for WT modeling can, sporadically, induce some important discrepancies between the two datasets. The error can significantly increase if the model’s training/testing dataset is small.

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

Station	N*	MAE	RMSE	NSE	KGE	bias	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

*number of dataset values

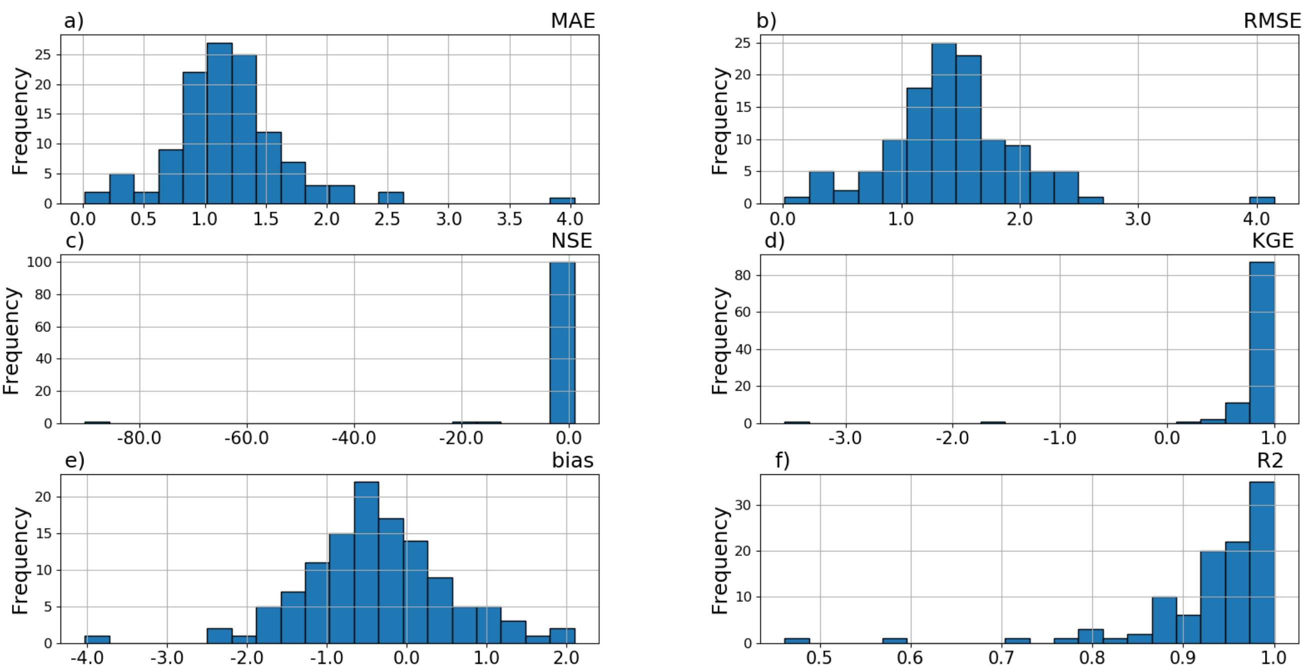


Figure 3: Metrics histograms of daily air temperature - ERA5 *versus* ground observed datasets

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 ºC (σ = 1.06), considering all station results, varying from 1.57 ºC to 7.23 ºC and a mean NSE value of 0.52 (σ=0.23) to be the best performing model, closely followed by the ANN model ensemble, with a mean RMSE of 3.22 ºC (σ = 1.05), varying from 1.79 ºC to 3.22 ºC (Table A2) and by the SVR model ensemble, with a mean RMSE of 3.37 ºC (σ = 0.96), varying from 1.34 ºC to 6.21 ºC. The SVR model had the lowest RMSE of all the simulations run: 1.34 ºC 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 ºC (σ = 0.96) varying from 1.34 ºC to 6.21 ºC.

The Air2stream model with 3-parameters is the best of the hybrid model parameterizations, with a mean RMSE of 4.06 ºC (σ = 1.17), followed by the MR, with an annual mean RMSE of 4.28 ºC (σ = 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 (μ=0.52; σ=0.23) can be considered satisfactory (0.50<NSE<0.65). According to the same classification, the ANN and the SVR with a mean annual NSE value of 0.48 (σ=0.28) and 0.47 (σ=0.19) produce an unsatisfactory modeling performance (NSE ≤ 0.50). The same classification was obtained with the all the parameterizations of the Air2stream model and the MR, but with a very reduced NSE value. The mean annual RMSE considering the ensemble of all model results for the testing phase is 2.75 ºC (σ = 1.00), varying from 1.34 ºC to 6.03 ºC and, according to the mean NSE value (μ=0.56; σ=0.48), the model's ensemble can be considered satisfactory. The individual model's contribution 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.

Fig. 4 and 5 show the RMSE obtained with each model during the training and testing phases, respectively, and the interannual variability 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 that:

i) During the training phase, all models exhibited a very low mean RMSE and interannual variability, except the Air2stream (3par) and the MR;

ii) The RF underfitted the training datasets with less than 30 values and, consequently, the predicted WT values exhibited a high RMSE and interannual variability during the testing phase (σ=1.28) (Fig. 4 and 5);

iii) During the training phase, the ANN exhibited the lowest mean annual RMSE (μ=0.44ºC; σ=0.40) (Table 4). This model clearly overfitted the training datasets, with less than 30 values, which increased the RMSE obtained for Stations 1 to 11 (Fig. 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 (σ=1.28);

iv) 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 (σ=1.25), including for the stations 1 to 10;

v) The Air2stream (3-parameters) 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 (σ=4.10) for the datasets with less than 30 values (Stations 1 to 10), which was reflected in the results obtained during the testing phase.

Fig. 6 was included to provide greater insight into the underfitting and overfitting associated with the ML models. The training datasets with less than 30 values are clearly underfitted by the RF model (Fig. 6a) and overfitted by the ANN and SVR (Fig. 6c and 6e). 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 also interesting to look at the results obtained from the models with regard to levels of performance. Fig. 7 shows the temporal evolution of the WT values obtained during the training and testing datasets for Station 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 with a monitoring error and, by doing so, the predicted temporal evolution of the river WT oscillates widely (Fig. 7a, 7c, and 7e). 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 7i). 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, 7d and 7f). In this example, the Air2stream (5-parameters) 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).

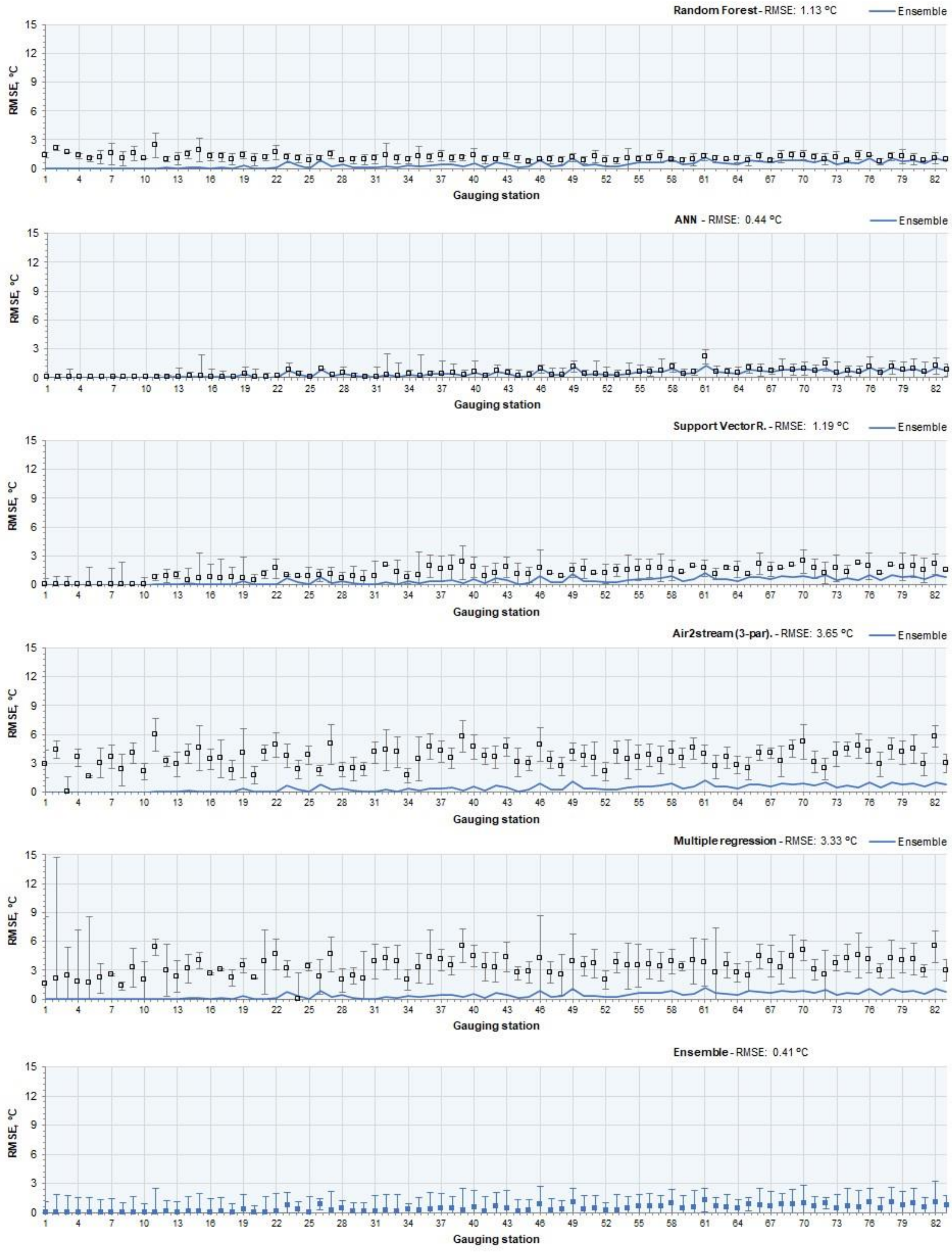


Figure 4: Root-mean-square error between observed and predicted WT values obtained during the training phase with all models (with standard deviation of interannual RMSE), considering the model results and the ensemble of all models results. Stations are ordered by the number of training dataset values,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

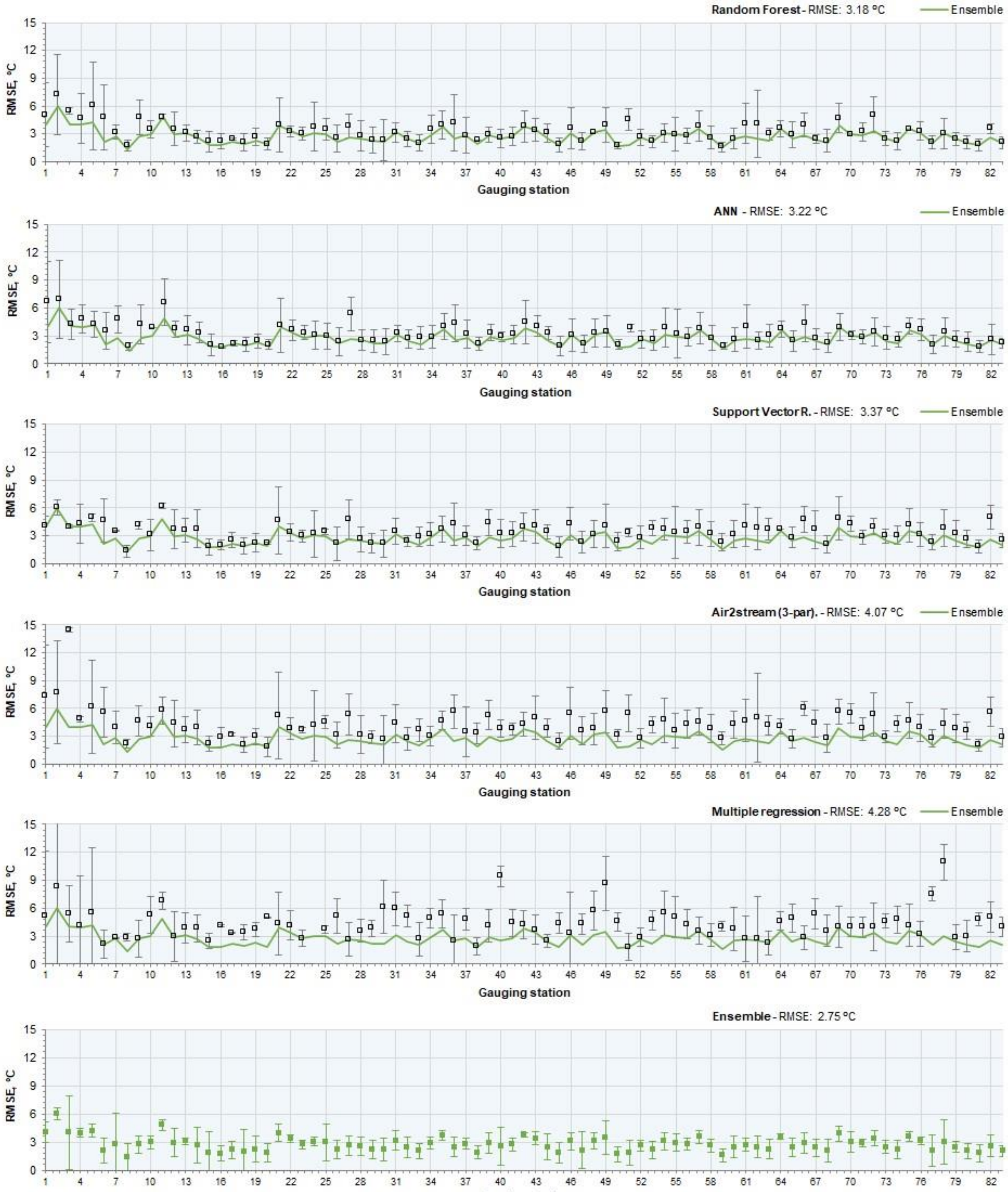
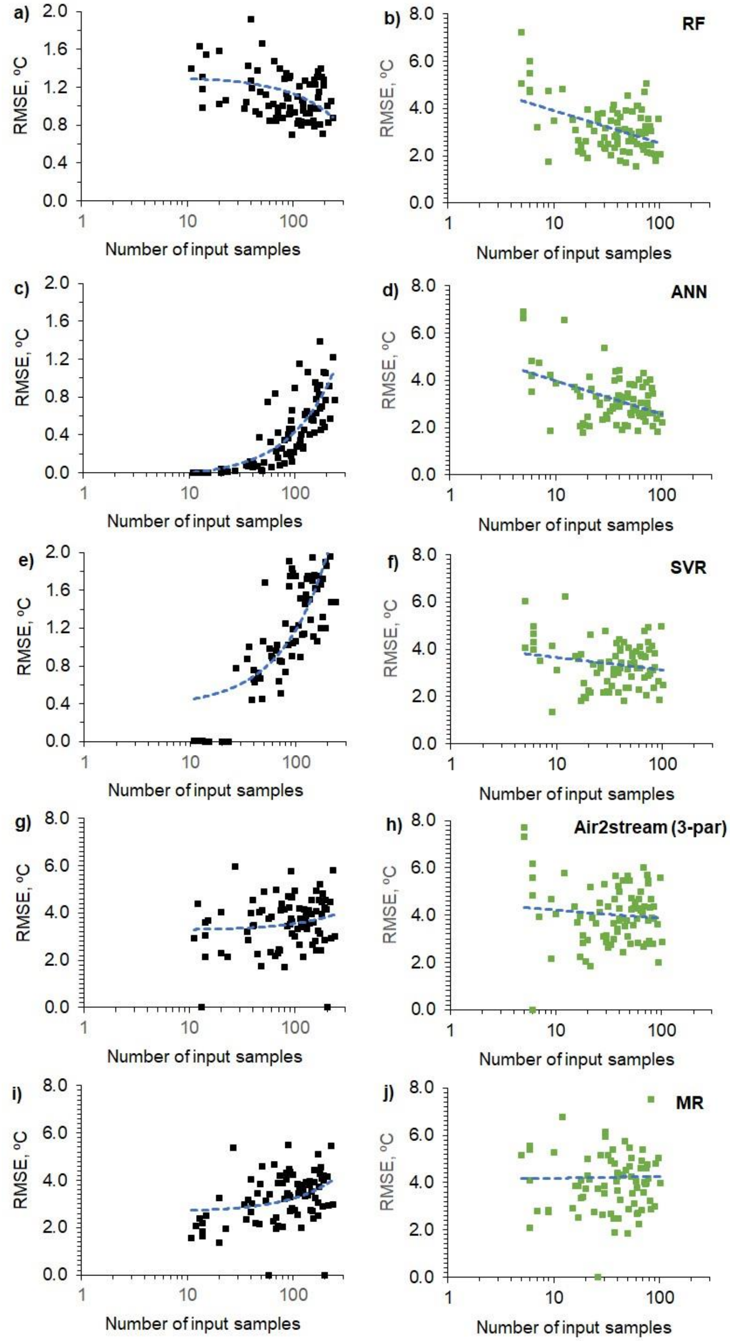


Figure 5: Root-mean-square error between observed and predicted WT values obtained during the testing phase with all models (with standard deviation of interannual RMSE), considering the model results and the ensemble of all models results. Stations are ordered by the number of testing dataset values,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



455 **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)**

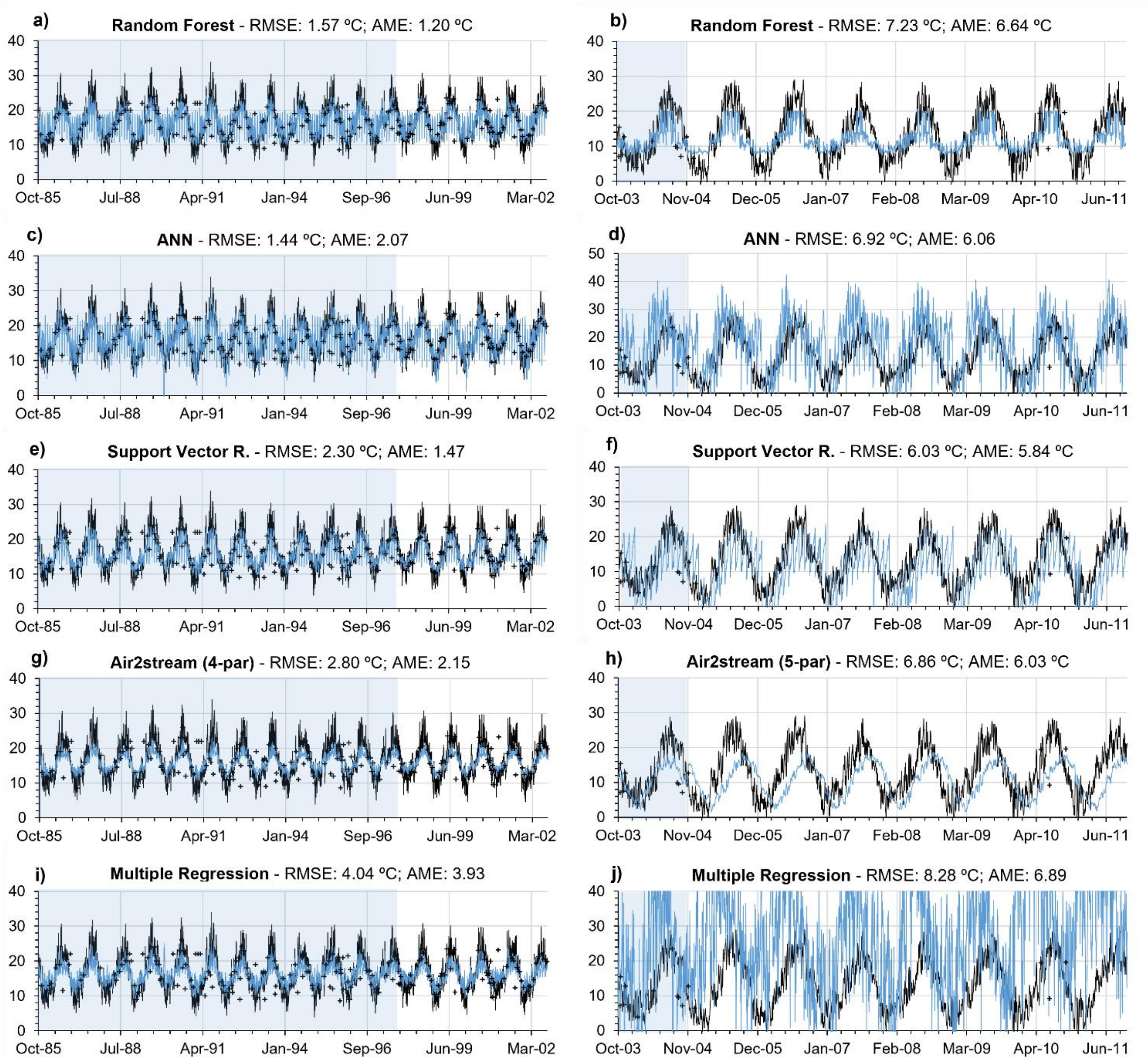


Figure 7: Root-mean-square error between observed (black dots) and predicted WT (Blue line) values obtained during the calibration (blue shadow area) and testing phase (white shadow area) with all models for Station 59 (graphs on left) and Station 2 (graphs on right). Air temperature (black line)

4.3 Model intercomparison – seasonal datasets

The results obtained for the dry and wet season testing datasets, considering all metrics, suggest that models performance is better for the dry season, with the exception of the results obtained with the Air2stream model using 3 and 4 and 5 parameters (Tables A3 and A4). The model using the 3 and 4 parameters does not consider the effect of river discharge and the 5-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 7 and 8 Air2stream model parameterizations. 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% (μ=30.0; σ=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 (Random Forest) 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 (Table A5). 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 ensembles for both seasons, considering the mean MAE and RMSE values. 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 (±1.91); SVR: 0.13 (±1.10); ANN: 0.10 (±1.22)), nonetheless the mean KGE values favor the ANN ensemble over the other ML results (KGE: RF: 0.46 (±0.26); SVR: 0.37 (±0.26); ANN: 0.48 (±0.36)). The Air2stream model with 3-parameters is the best of the hybrid model parameterizations followed by the MR (Tables A3 and A4).

4.4 Feature importance

Table 5 shows the mean feature importance obtained with the best performing model (Random Forest Regressor, Pedregosa et al., 2011) considering the mean annual RMSE, a RF with the following parameters: n_estimators = 50; max_depth = 485; min_samples_split = 5; max_features = 'auto'; bootstrap = True; random_state = 42; considering all stations datasets. 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 A6 shows the evaluation of the RF model performance during the training and testing phases considering the annual datasets and the sequential increase of the model predictors. The results show that, on average, the inclusion of all predictor variables have a significant effect on model performance.

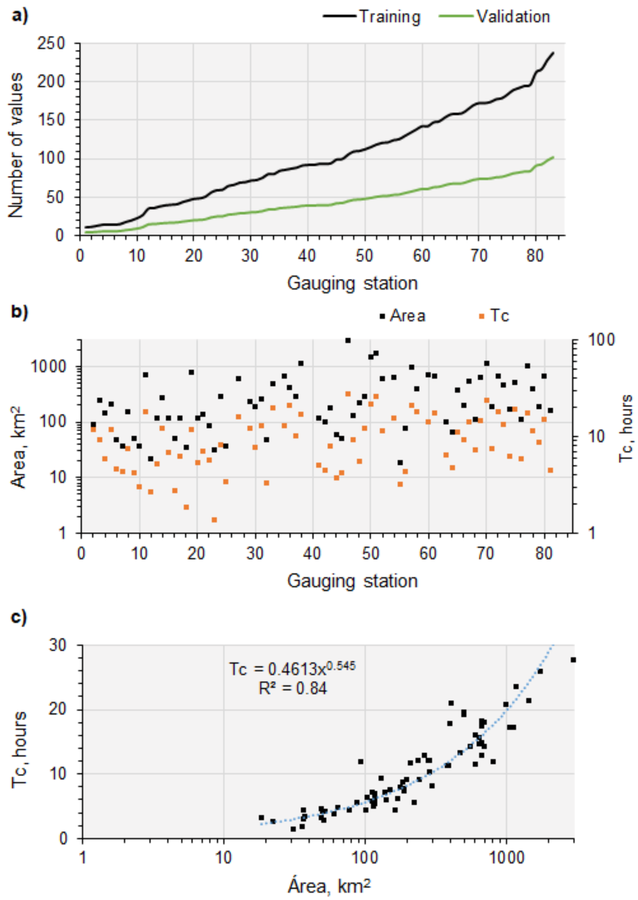
Table 5: Mean input feature importance obtained with a Random Forest regressor

	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
Stdev	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

4.5 Effect of the watershed time of concentration on model performance

Not surprisingly, the results suggest that, tendentially, there are more training and testing datasets available for the largest watersheds (Fig. 8a and 8b) 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 9c). After the data sets z-score normalization and the application of the Gaussian Mixture Models algorithm with the following parameters: n_components=2, covariance_type='diag',init_params='random',warm_start=True (*vide* Pedregosa et al., 2011), two different data samples were extracted. This small set of values, 19 (watershed area: μ= 106 km^2; σ=153) (Fig. 9b) and 19 (watershed area: μ= 106 km^2; σ=153) (Fig. 9d) 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 increases by an average of 0.1 ºC with a one-hour increase in the watershed time of concentration, considering the RF ensemble aggregation approach (Fig. 9d).



530 **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**

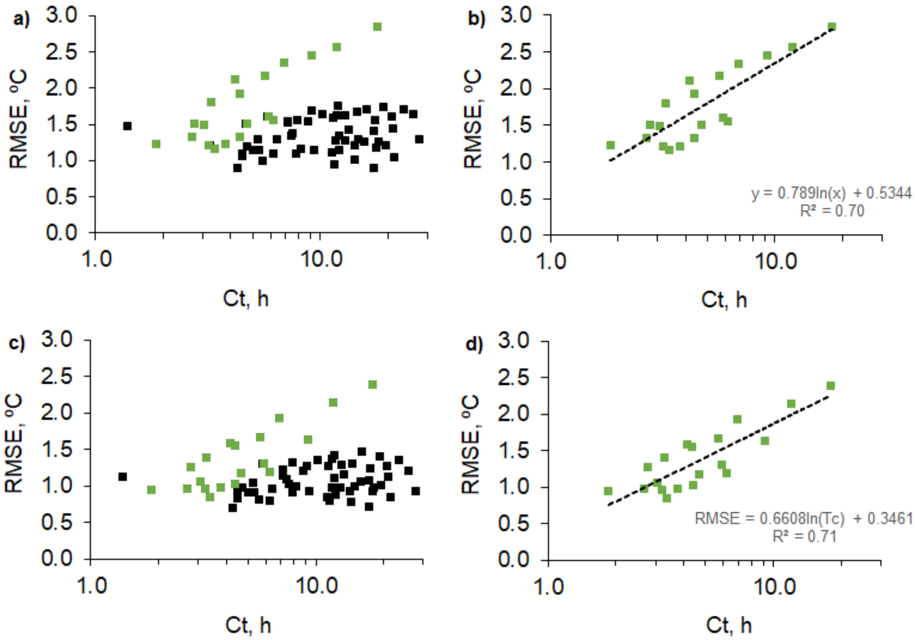


Figure 9: a) RMSE between observed and simulated river WT with the Random Forest best model run versus the watershed time of concentration. b) Data extraction from a). c) RMSE between observed and simulated river WT with the Random Forest ensemble aggregation approach versus the watershed time of concentration. d) Data extraction from c)

5 Discussion

Overall, the results of the model's ensemble (mean RMSE: 2.75 ºC; σ = 1.00) driven mainly by the ML algorithms predictions are in line with the results obtained in other studies, namely Rabi et al., (2015) (ANN - RMSE: μ=2.06 ºC) and Zhu et al., (2019a) (MR – RMSE: 2.74 ºC), and do not differ greatly from the results obtained in other studies (*vide* Table 1). This is quite significant considering the scale of the missing training and testing datasets corresponding to this study (μ= 98.8%; σ=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 (*vide* Table1). 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 (e.g., Feigl et al., 2021). That said, the RF model ensemble produced the best results considering 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 limiting 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 FNNs and the recurrent neural networks (RNNs) performed better than the Random Forest 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 effectively could 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 spans 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 water temperature. 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 6 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 the best results were obtained with the 3-parameter model, the simplest version of the Air2stream models, which is a model that 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 4- and 6-parameter Air2stream model. When the length of the calibration period is of one year and the percentage of missing data is in the range of 99%, the RMSE between observed and predicted lake WT is >3.5 ºC. 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 reanalysis datasets and by upstream conditions, which increases with the watershed area. The results of this study considering the quality of the input datasets suggests that when the missing datasets reach 98%, a RMSE <3.0ºC is indicative of a good modeling performance. Also relevant is the fact that this error can 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, but can also be associated with the uncertainty caused by the fact that river WT is not only affected by local environmental conditions, but also by upstream conditions. 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, the fact that, 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 results of this study demonstrate the feasibility of finding a correlation between the prediction error between 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.

5 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 algorithms hyperparameters. The Tree-structured Parzen Estimators 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: n_estimators = 50; max_depth = 485; min_samples_split = 5; max_features = 'auto'; bootstrap = True; random_state = 42) produce the best performance and may represent an effective solution for model river WT with limiting forcing data.

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 ºC with a one-hour increase in the watershed time of concentration (watershed area: μ= 106 km^2; σ=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: 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 (ºC)**	**RMSE (ºC)**	**NSE**	**KGE**	**bias (ºC)**	**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	**TESTING**					
Model/metric	**MAE (ºC)**	**RMSE (ºC)**	**NSE**	**KGE**	**bias (ºC)**	**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 A3: 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 (ºC)**	**RMSE (ºC)**	**NSE**	**KGE**	**bias (ºC)**	**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 (ºC)**	**RMSE (ºC)**	**NSE**	**KGE**	**bias (ºC)**	**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 A4: 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 (ºC)**	**RMSE (ºC)**	**NSE**	**KGE**	**bias (ºC)**	**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 (ºC)**	**RMSE (ºC)**	**NSE**	**KGE**	**bias (ºC)**	**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 A5: Evaluation of the Random Forest performance during the training and testing phases considering the dry and wet seasons and the sequential increase of the model predictors. Mean MAE, RMSE, NSE, KGE, bias and R2 (with standard deviation) between observed and predicted WT values.

Dry season	**TRAIN**						
Metric/predictor	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**	**7)**
MAE **(ºC)**	1.88 (±0.52)	1.65 (±0.47)	1.53 (±0.45)	1.52 (±0.44)	1.47 (±0.44)	1.44 (±0.43)	1.09 (±0.39)
RMSE **(ºC)**	2.35 (±0.61)	2.10 (±0.56)	1.97 (±0.53)	1.95 (±0.52)	1.90 (±0.52)	1.88 (±0.52)	1.44 (±0.48)
NSE	0.65 (±0.22)	0.71 (±0.25)	0.75 (±0.26)	0.75 (±0.26)	0.76 (±0.26)	0.77 (±0.26)	0.85 (±0.28)
KGE	0.63 (±0.14)	0.66 (±0.14)	0.66 (±0.13)	0.67 (±0.13)	0.67 (±0.13)	0.67 (±0.14)	0.79 (±0.12)
bias **(ºC)**	0.23 (±0.46)	0.21 (±0.41)	0.17 (±0.39)	0.17 (±0.40)	0.16 (±0.41)	0.14 (±0.38)	0.13 (±0.45)
R^2	0.73 (±0.09)	0.80 (±0.07)	0.84 (±0.05)	0.85 (±0.05)	0.86 (±0.05)	0.85 (±0.10)	0.91 (±0.04)
Dry season	**TEST**						
Metric/predictor	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**	**7)**
MAE **(ºC)**	3.66 (±1.22)	3.48 (±1.24)	3.40 (±1.28)	3.40 (±1.26)	3.38 (±1.30)	3.35 (±1.27)	2.49 (±1.22)
RMSE **(ºC)**	4.58 (±1.41)	4.40 (±1.42)	4.25 (±1.37)	4.24 (±1.36)	4.24 (±1.45)	4.19 (±1.39)	3.17 (±1.36)
NSE	-0.44 (±0.62)	-0.34 (±0.66)	-0.24 (±0.58)	-0.23 (±0.61)	-0.23 (±0.58)	-0.20 (±0.61)	0.25 (±0.81)
KGE	0.16 (±0.31)	0.17 (±0.33)	0.16 (±0.32)	0.17 (±0.32)	0.16 (±0.34)	0.17 (±0.34)	0.54 (±0.25)
bias **(ºC)**	0.45 (±2.00)	0.41 (±2.01)	0.33 (±1.78)	0.36 (±1.84)	0.33 (±1.83)	0.28 (±1.81)	0.24 (±1.75)
R^2	0.16 (±0.22)	0.20 (±0.24)	0.20 (±0.24)	0.21 (±0.24)	0.21 (±0.25)	0.22 (±0.24)	0.55 (±0.22)
Wet season	**TRAIN**						
Metric/predictor	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**	**7)**
MAE **(ºC)**	1.80 (±0.59)	1.59 (±0.54)	1.49 (±0.51)	1.46 (±0.50)	1.42 (±0.50)	1.38 (±0.49)	1.06 (±0.36)
RMSE **(ºC)**	2.31 (±0.67)	2.07 (±0.60)	1.97 (±0.57)	1.91 (±0.57)	1.88 (±0.57)	1.83 (±0.55)	1.42 (±0.43)
NSE	0.64 (±0.13)	0.71 (±0.11)	0.73 (±0.10)	0.75 (±0.11)	0.76 (±0.10)	0.77 (±0.10)	0.85 (±0.10)
KGE	0.60 (±0.14)	0.63 (±0.13)	0.63 (±0.12)	0.64 (±0.21)	0.65 (±0.12)	0.66 (±0.11)	0.76 (±0.11)
bias **(ºC)**	-0.16 (±0.27)	-0.13 (±0.23)	-0.11 (±0.21)	-0.10 (±0.21)	-0.09(±0.21)	-0.09 (±0.19)	-0.08 (±0.19)
R^2	0.70 (±0.12)	0.78 (±0.10)	0.82 (±0.08)	0.83 (±0.0.8)	0.84 (±0.07)	0.85 (±0.07)	0.89 (±0.07)
Wet season	**TEST**						
Metric/predictor	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**	**7)**
MAE **(ºC)**	3.47 (±1.42)	3.44 (±1.52)	3.39 (±1.63)	3.37 (±1.59)	3.36 (±1.56)	3.27 (±1.55)	2.56 (±1.40)
RMSE **(ºC)**	4.39 (±1.52)	4.31 (±1.57)	4.25 (±1.65)	4.25 (±1.63)	4.23 (±1.61)	4.16 (±1.63)	3.27 (±1.46)
NSE	-0.66 (±3.06)	-0.52 (±2.36)	-0.64 (±.3.86)	-0.61 (±3.63)	-0.51 (±2.88)	-0.49 (±2.81)	0.04 (±2.16)
KGE	0.14 (±0.32)	0.13 (±0.32)	0.09 (±0.35)	0.11 (±0.30)	0.09 (±0.33)	0.13 (±0.34)	0.45 (±0.26)
bias **(ºC)**	-0.56 (±1.95)	-0.77 (±1.92)	-0.63 (±2.00)	-0.62 (±2.00)	-0.59 (±1.96)	-0.49 (±1.95)	-0.29 (±1.80)
R^2	0.18 (±0.23)	0.19 (±0.23)	0.20 (±0.23)	0.19 (±0.23)	0.19 (±0.22)	0.20 (±0.22)	0.45 (±0.24)

Table A6: Evaluation of the Random Forest performance during the training and testing phases considering the annual datasets and the sequential increase of 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 **(ºC)**	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 **(ºC)**	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 **(ºC)**	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)
R2	0.76 (±0.08)	0.82 (±0.0.7)	0.85 (±0.05)	0.86 (±0.0.4)	0.87 (±0.04)	0.87 (±0.05)	0.92 (±0.04)
Annual	**TEST**						
Metric/predictor	**1)**	**2)**	**3)**	**4)**	**5)**	**6)**	**7)**
MAE **(ºC)**	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 **(ºC)**	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 **(ºC)**	-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)
R2	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 (2022). Additionally, this repository includes the input data considered in this study (83 datasets). It is also possible to download the code/data from https://github.com/mcvta/WaterPythonTemp.

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

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study had the support of national funds through Fundação para a Ciência e Tecnologia (FCT), under the project LA/P/0069/2020 granted to the Associate Laboratory ARNET, and the strategic project UDIB/04292/2020 granted to MARE.

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, URL: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf.

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), Hydrological Processes, 21, 21–34, https://doi.org/10.1002/hyp.6353, 2007.

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

Almeida, M. C., and Coelho, P. S.: mcvta/WaterPythonTemp: HESS Submission (v0.1.0). Zenodo. https://doi.org/10.5281/zenodo.6620581, 2022.

Branco, P., Ribeiro, R. P., Torgo, L., Krawczyk, B., Moniz, N.: Smogn: a pre-processing approach for imbalanced regression, Proceedings of Machine Learning Research 74, 36–50. 2017.

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, 2013.

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

Caissie, D.: The thermal regime of rivers: a review, Freshwater Biol., 51, 1389–406, https://doi.org/10.1111/j.13652427.2006.01597.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. Journal Climat., 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, Hydrological Processes, 22 (17), 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 Biology, 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.

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. Review of Geophysics, 45, RG2004, https://doi.org/10.1029/2005RG000183, 2007.

Fahrer, C., and Harris, D., 2004. LAMPPOST A Mnemonic device for teaching climate variables. Journal of Geography 103:86-90, https://doi.org/10.1080/00221340408978579, 2004.

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.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Munõz-Sabater J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc ,146(730): 1999-2049l, https://doi.org/10.1002/qj.3803, 2020.

Jeppesen, E. and Torben Moth Iversen. T.: Two Simple Models for Estimating Daily Mean Water Temperatures and Diel Variations in a Danish Low Gradient Stream, Oikos, 49(2), 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.

McNicholl, B., Lee, Y.H., Campbell, A.G., Dev, S.: Evaluating the Reliability of Air Temperature from ERA5 Reanalysis Data. IEEE, Geosci. Remote Sens. Lett, 19, 1–5, https://doi.org/10.1109/lgrs.2021.3137643, 2021.

Mohseni, O., Erickson T. R., and Stefan, H. G.: Upper bounds for stream temperatures in the contiguous United States, Journal of Environmental Engineering, American Society of Civil Engineers, 128(1), 4–11, https://doi.org/10.1061/(ASCE)07339372(2002)128:1(4), 2002.

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

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, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 38(2), 135-147, https://doi.org/10.1080/07011784.2013.794992, 2013.

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.

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, Tranactions of the ASABE, 50(3), 885-900. https://doi.org/10.13031/2013.23153, 2007.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models: Part 1. A discussion of principles, J. Hydrol., 10(3), 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, Journal of Environmental Engineering, 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, JMLR 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, Advances in Oceanography and Limnology, 7(1), 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. Open File Rep., 98–14, 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), Hydrological Sciences Journal, 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, Journal of Hydroinformatics, 23 (3), 605–626, https://doi.org/10.2166/hydro.2021.121, 2021.

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

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.

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 Resources Research 29:2299_2312, https://doi.org/10.1029/93WR00540, 1993.

Smith, K.: River water temperatures - an environmental review. Scottish Geographical Magazine, 88(3), 211–220, https://doi.org/10.1080/00369227208736229, 1972.

Smith K.: The prediction of river water temperatures. Hydrological Sciences Bulletin, 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. https://www.jstor.org/stable/3543713, 1975.

Smola, A. J., and Schölkopf, B. A.: tutorial on support vector regression, Statistics and Computing 14, 199–222, 780 https://doi.org/10.1023/B:STCO.0000035301.49549.88, 2004.

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 Dyn, 45, 1771–1787, https://doi.org/10.1007/s00382-014-2432-x , 2015.

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 Dyn, 39, 2497–2522,

https://doi.org/10.1007/s00382-012-1315-2#Bib1, 2012.

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, https://doi.org/10.5555/2627435.2670313, 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 Science and Technology 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, Environmental Research Letters, 10, 114011, https://doi.org/10.1088/1748-9326/10/11/114011, 2015.

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

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, 3069–84, https://doi.org/10.1002/hyp.1280, 2003.

Webb, B. W., and Nobilis, F.: A long-term perspective on the nature of the air-water temperature relationship: a case study, Hydrological Processes, 11, 137–147, https://doi.org/10.1002/(SICI)1099-1085(199702)11:2<137::AID-HYP405>3.0.CO;22, 1997.

Wetzel, R.G.: Limnology. Lake and River Ecosystems, Third Edition, Academic Press. https://doi.org/10.1016/C2009-002112-6, 2001.

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

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/s11356-019-04716y, 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. 2019 Extreme learning machine-based prediction of daily water temperature for rivers, Environmental Earth Sciences, 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.