Prediction of river water temperature using machine learning algorithms: a tropical river system of India

by

Maddu Rajesh, Shaik Rehana

in

International Conference, Asia Oceania Geosciences Society (AOGS)
: 1
-22

Report No: IIIT/TR/2021/-1

Centre for Spatial Informatics
International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2021

Prediction of river water temperature using machine learning algorithms: a tropical river system of India

M. Rajesh and S. Rehana

ABSTRACT

Machine learning (ML) has been increasingly adopted due to its ability to model complex and non-linearities between river water temperature (RWT) and its predictors (e.g., Air Temperature, AT). Most of these ML approaches have been applied using average AT without any detailed sensitivity analysis of other forms of AT (e.g., maximum and minimum). The present study demonstrates how new ML approaches, such as ridge regression (RR), K-nearest neighbors (KNN) regressor, random forest (RF) regressor, and support vector regression (SVR), can be coupled with Sobol' global sensitivity analysis (GSA) to predict accurate RWT estimates with the most appropriate form of AT. Furthermore, the proposed ML approaches have been combined with the Ensemble Kalman Filter (EnKF), a data assimilation (DA) technique to improve the predicted values based on the measured data. The proposed modelling framework's effectiveness is demonstrated with a tropical river system of India, Tunga-Bhadra River, as a case study. The SVR has been noted as the most robust ML model to predict RWT at a monthly time scale compared with daily and seasonal. The study demonstrates how ML methods can be coupled with a global sensitivity algorithm and DA techniques to generate accurate RWT predictions in river water quality modelling.

Key words | Ensemble Kalman Filter, K-nearest neighbors, random forest, river water temperature, Sobol' sensitivity analysis, support vector regression

M. Rajesh
S. Rehana (corresponding author)
Lab for Spatial Informatics,
International Institute of Information Technology,
Hyderabad 500032,
India
E-mail: *rehana.s@iiit.ac.in*

HIGHLIGHTS

- Machine learning models coupled with global sensitivity analysis to predict RWT.
- Ridge regression, KNN, random forest, SVR, along with Sobol' sensitivity analysis were explored.
- Maximum AT as the most sensitive variable in RWT prediction.
- The SVR as the most robust ML model to predict RWT at monthly time scale.
- Application on a tropical river system of India.

INTRODUCTION

The river water temperature (RWT) directly affects the river's physical, biological, and chemical characteristics and determines the fitness and life of all aquatic organisms. The RWT is of particular significance as (i) the discharge of excess heat from industries and municipal effluents can affect the aquatic ecosystem, (ii) temperature influences both biological and chemical reactions, and (iii) temperature fluctuations affect the density of water and hence the transport of water (Thomann & Mueller 1987). For many

doi: 10.2166/hydro.2021.121

environmental, hydrology, and ecology applications, accurate prediction and assessment of RWT have become the key problem (Zhu *et al.* 2019b, 2019c). In this context, process-based RWT models have been evolved based on heat advection-dispersion transport equations (Stefan & Sinokrot 1993) and net heat transfer processes at the surface based on thermal equilibrium concepts (Mohseni *et al.* 1999; Rehana & Mujumdar 2012). Although such process-based models give exact results, a large amount of detailed and computationally intensive data is required. Due to the simplicity of implementation, regression models have been improved using the relationship between air and water temperatures (e.g., Stefan & Preud'homme 1993; Pilgrim *et al.* 1998; Erickson Troy & Stefan Heinz 2000; Neumann David *et al.* 2003; Rehana & Mujumdar 2011). The usual illustrations are linear regression models (Morrill *et al.* 2005; Krider *et al.* 2013), non-linear regression models (Mohseni *et al.* 1998; van Vliet *et al.* 2012), stochastic regression models (Ahmadi-Nedushan *et al.* 2007; Rabi *et al.* 2015), and hybrid statistical-physical based models (Gallice *et al.* 2015; Toffolon & Piccolroaz 2015; Piccolroaz *et al.* 2016) have been developed successfully for data relating to different time scales in the past years. Artificial neural networks (ANNs) have proven to be a promising mathematical tool for predicting the non-linear relationships and their applications in RWT predictions (Chenard & Caissie 2008; Sahoo *et al.* 2009; DeWeber & Wagner 2014; Hadzima-Nyarko *et al.* 2014; Piotrowski *et al.* 2015; Rabi *et al.* 2015; Temizyurek & Dadaser-Celik 2018; Zhu *et al.* 2018, 2019d, 2019e). In recent years, Zhu *et al.* (2018, 2019a, 2019b) and Graf *et al.* (2019) developed the wavelet neural networks (WT-ANN), decision tree (DT), feedforward neural network (FFNN), Gaussian process regression (GPR), and extreme learning machine (ELM) based models to estimate RWT, and these models are very effective to a linear model and a non-linear model. However, support vector regression (SVR), which is based on structural risk minimization to avoid overfitting (Vapnik *et al.* 1996), has been adopted over ANN for RWT predictions due to the uniqueness and globalization of the solution (Rasouli *et al.* 2012; Wang *et al.* 2013; Huang *et al.* 2017; Heddam & Kisi 2018; Komasi *et al.* 2018; Rehana 2019). Random forest (RF) models have been used extensively in hydrology (Balk & Elder 2000; Tehrany *et al.* 2013; Li *et al.* 2020), and few researchers have applied for RWT modelling (Lu & Ma 2020). The K-nearest neighbors (KNN) approach has been used in many hydrology applications (Souza & Lall 2003; Beersma & Buishand 2004; Leander *et al.* 2005) and can be a proper choice for RWT predictions (Muluye 2012; Antunes *et al.* 2018; Gavahi *et al.* 2019).

In this context, the robustness of any such data-driven ML algorithms depends on the feature vector (predictors) under consideration in the prediction of RWT. Few studies have tried to model RWT by considering multiple factors, such as river flow discharge (Webb *et al.* 2003; Laanaya *et al.* 2017), solar radiation (Sahoo *et al.* 2009), riparian shade (Johnson *et al.* 2014), landform attributes, and forested land cover (DeWeber & Wagner 2014). However, the inclusion of air temperature (AT) as the sole variable in predicting RWT has gained much popularity in the research community due to the ready availability of temperature variables (e.g., Caissie 2006; Rehana & Mujumdar 2011). To this end, many studies have used average AT as the promising variable in RWT estimation using data-driven algorithms and hybrid algorithms due to the direct and linear relationships between average air and water temperatures (Piccolroaz *et al.* 2016; Rehana & Dhanya 2018; Zhu *et al.* 2018, 2019c; Graf *et al.* 2019; Rehana 2019). However, at maximum ATs, which are prevailing under seasonal temperature variations, the atmosphere's moisture-holding capacity increases, and the rate of evaporative cooling also increases, and therefore, the RWT no longer increases linearly with average AT (Mohseni *et al.* 1998; Bogan *et al.* 2003). Therefore, a thorough sensitivity analysis must be performed to identify the most influencing AT variable (average, maximum, and minimum) to predict the RWT before applying any data-driven algorithm. Given that several studies focused on average AT as the only variable to predict RWT using various ML algorithms, selecting an appropriate AT variable (average, maximum, and minimum) has not been intensively studied in the literature. To the author's best knowledge, none of the studies applied sensitivity analysis to select the best suitable and effective AT variable among maximum, minimum, and average and tested various ML models in the prediction of RWT. The present study assessed the ML model's capability with a global sensitivity analysis (GSA) to better predict RWT. The present study proposed a GSA algorithm variance based on the Sobol' method (Sobol 1990; Sobol′ 2001) to predict more influencing AT variables in the prediction of RWT. Although the Sobol' method has been used in many fields of science and engineering, it has been very limited in hydrology applications (Tang *et al.* 2006; Cloke *et al.* 2008; Pappenberger *et al.* 2008; van Werkhoven *et al.* 2009; Cibin *et al.* 2010; Yang 2011). The present study made efforts to use the Sobol' method to select highly sensitive features in RWT prediction.

One of the major limitations of ML algorithms includes the difficulty of incorporating existing physical knowledge (Boukabara *et al.* 2020). The most appropriate way forward is to combine the best of the two approaches: theory-driven and understanding-rich processes with data-driven discovery processes (Babovic 2005). Recent progress in ML inspires the idea of learning data assimilation (DA) models directly from the real observations – these are uncertain, sparsely sampled, and only indirectly sensitive to the processes of interest (Geer 2020). DA is a methodology that uses observational data and combines it with (or assimilates it into) numerical models (Babovic *et al.* 2005). The DA method can be categorized into four groups (WMO 1992; Babovic 2005): (i) updating input parameters, (ii) updating model parameters, (iii) updating state variables, and (iv) updating output variables. The fourth type updates output directly, and the possibility of forecasting these errors and superimposing them to the simulation model forecasts usually gives a good performance (Babovic *et al.* 2005).

DA has been used to enhance simulation accuracy in many engineering applications. One of the most efficient and sequential DA methods is the Kalman filter (KF) developed by Kalman (1960), and its applications in hydrology are also very impressive (Liu *et al.* 2010; Li *et al.* 2013; Wang & Babovic 2016; Wang *et al.* 2016, 2017; Mehrparvar & Asghari 2018). In RWT forecasting, only a few studies addressed the use of DA (Morrison & Foreman 2005; Yearsley 2009; Pike *et al.* 2013; Ouellet-Proulx *et al.* 2017). Besides, to the author's knowledge, a limited systematic DA method combined with ML has ever been applied in the context of RWT forecasting. Hence, this study presents an attempt to use an Ensemble Kalman Filter (EnKF) DA method to update and balance the ML model estimates by available observed historical data in RWT forecasting. This paper proposed an integrated modelling framework with ML and DA approach to improving the predicted values based on the measurement data. The proposed algorithm has been demonstrated with a river gauging station daily temperature data of the Shimoga station along the Tunga River, a tributary of the Tunga-Bhadra River, a major tributary of the Krishna River, India. In summary, the objectives of the present study are to (i) identify the most influencing AT variable by the GSA algorithm; (ii) apply various ML models (ridge regression (RR), KNN regressor, RF regressor, and SVR) with the best selected AT for RWT prediction; (iii) apply the EnKF with each ML model; and (iv) compare the performance of four advanced ML algorithms by coupling the GSA and EnKF algorithms when applied on a tropical river system of India.

STUDY AREA AND DATA

The river location considered for the modelling of RWT is Shimoga along the Tunga River, which confluences with the Bhadra River to form the Tunga-Bhadra River, a major tributary of the Krishna River basin, India (Figure 1). A storage dam is situated about 15 km upstream from Shimoga at Gajanur across the river Tunga. The monthly mean discharge at the Shimoga station is about 166.95 m^3/s. The observed minimum, maximum, and average air (water) temperature mean were noted as 19.66, 29.74, and 24.78 °C (27.54 °C) and standard deviation as 3.48, 3.47, and 2.77 °C (2.66 °C), respectively. A significant decrease of discharge has been noted about 3.1% at Shimoga along the Tunga River compared from 1971–1991 to 1992–2006 (Rehana & Mujumdar 2011). The Tunga River location receives the waste load from the Shimoga city municipal effluent. The daily average RWT data and average, maximum and minimum AT data from 1 January 1989 to 1 January 2004 recorded at the Shimoga station were obtained from Central Water Commission (CWC), Bangalore, Karnataka, India, and Advanced Centre for Integrated Water Resources Management (ACIWRM), Karnataka, India. The frequency of water quality data collection, i.e., water temperature, is ten times a day. The measurement of water temperature data is mean daily of ten samples (Central Water Commission 2018). To create a complete time-series

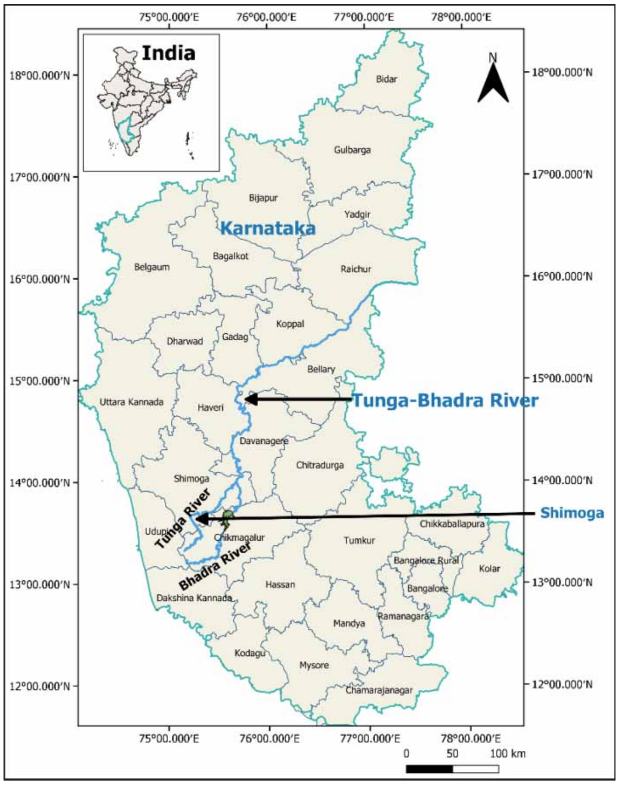

Figure 1 | Location map of the Tunga-Bhadra River and the Shimoga station, India.

dataset, the na.interp() function within the R's forecast package was used to interpolate data between missing time-series values (Hyndman *et al.* 2018). For seasonal data, na.interp uses STL (Seasonal and Trend decomposition using Loess) for this interpolation.

METHODOLOGY

The overview of the proposed modelling framework is shown in Figures 2 and 3. The first step is to apply sensitivity analysis to select the most appropriate form of AT variable to predict the RWT. Various ML approaches such as RR, KNN, RF, and SVR were applied to the study location to predict RWT at a daily time scale. Figure 2 shows the architectural flow diagram proposed for the prediction of RWT using sensitivity and ML. Figure 3 shows the ML model and the EnKF DA method's architectural flow diagram to improve the ML model's efficiency in each simulation step.

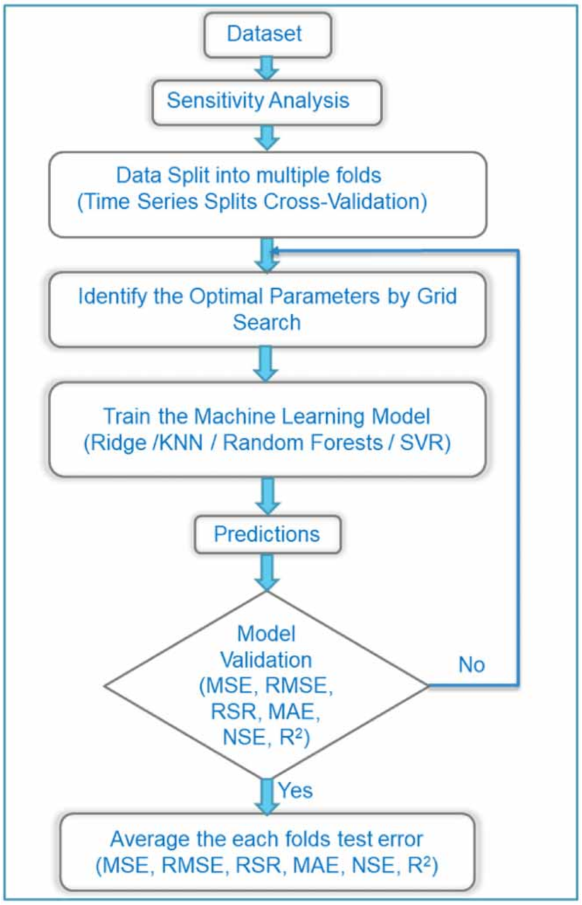
Dataset
Sensitivity Analysis
Data Split into multiple folds
(Time Series Splits Cross-Validation)
Identify the Optimal Parameters by Grid Search
Train the Machine Learning Model
(Ridge /KNN / Random Forests / SVR)
Predictions
Model Validation
(MSE, RMSE, RSR, MAE, NSE, R²)
No
Yes
Average the each folds test error
(MSE, RMSE, RSR, MAE, NSE, R²)

Figure 2 | Architectural flow diagram for ML regression models.

Sensitivity analysis

Sensitivity analysis (SA), which is often used as a powerful technique to measure the strength of relationships between model inputs and outputs, is an important assessment of any modelling, including environmental modelling (Nossent *et al.* 2011). SA is crucial in hydrologic and water quality models due to various aspects involved in modelling processes, such as spatiotemporal scales and complexity, requiring an assessment of parameters influence on the model's prediction (Yuan *et al.* 2015). In recent years, various SA environmental models are available in the literature (Saltelli *et al.* 2010; Yang 2011), based on variance decomposition. The variance-based Sobol' method is an SA method that is very common in many fields (Sobol 1990). In general, SA methods aim to measure the amount of variance that each parameter adds to the unconditional variance of the model output, these amounts are expressed as (Sobol') sensitivity indices (SIs).

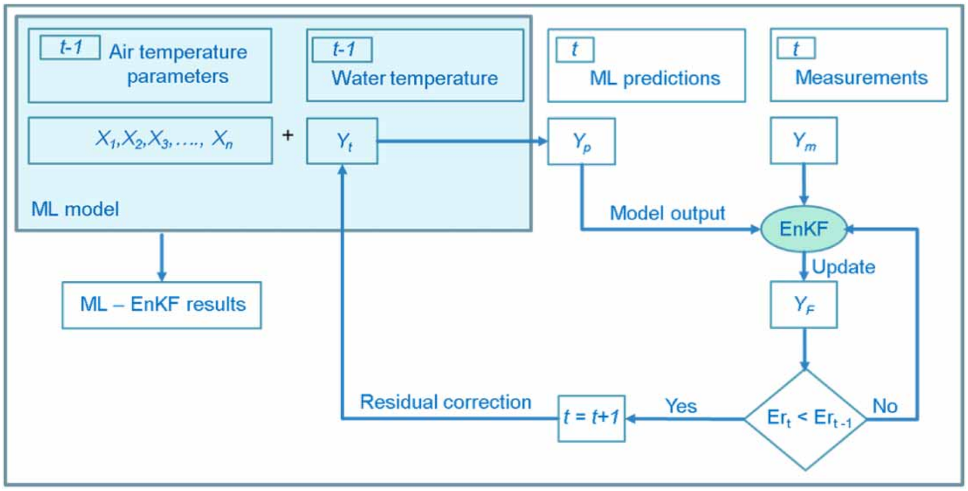

Figure 3 | Architectural flow diagram of ML model and EnKF data assimilation method.

Sobol' SA method

The method of Sobol' is an advanced, global, model-independent SA method that is based on variance decomposition. It can handle non-linear and non-monotonic functions and models. Considering a mathematical model, $Y = f(X)$, delivering the outputs of a physical system that presumably depends on M-uncertain input parameters $X = (X_1, \ldots, X_M)$. For further developments, $f_{X_i}(x_i)$ and $f_x = \Pi_{i=1}^{M} f_{X_i}(x_i)$ refer to their marginal probability density function (PDF) and the corresponding joint PDF of a given set. The sensitivity model can be defined as:

$$Y = f(X) = f(X_1, \ldots, X_M) \tag{1}$$

where Y is the objective function and $X = (X_1, \ldots, X_M)$ is the input parameter set. Sobol' proposed the decomposition of the function f into sums of increasing dimensionality:

$$f(X_1, \ldots, X_M) = f_0 + \sum_{i=1}^{M} f_i(X_i) + \sum_{i=1}^{M}\sum_{j=i+1}^{M} f_{ij}(X_i, X_j) + \cdots + f_{1,\ldots,M}(X_1, \ldots, X_M) \tag{2}$$

If the input factors are independent of each term in Equation (2) is chosen with zero average and is square-integrable, then f_0 is a constant, equal to the output expectation value, and the quantities are mutually orthogonal. The total unconditional variance can be described as:

$$V(Y) = \int_{\Omega^M} f^2(X)dX - f_0^2 \tag{3}$$

with Ω^M representing the M-dimensional unit hyperspace (i.e., the ranges of parameters are scaled between 0 and 1). The partial variances, which are the components of the total variance decomposition, are computed from each of the terms in Equation (2) as:

$$V_{i_1\ldots i_s} = \int_0^1 \cdots \int_0^1 f^2_{i_1\ldots i_s}(X_{i_1}, \ldots, X_{i_s})dX_{i_1}\ldots dX_{i_s} \tag{4}$$

where $1 \le i_1 \le \cdots \le i_s \le \mathrm{M}$ and $s = 1, \ldots, M$. Assuming that the parameters are mutually orthogonal, Equation (5) results for the variance decomposition.

$$V(Y) = \sum_{i=1}^{M} V_i + \sum_{i=1}^{M-1}\sum_{j=i+1}^{M} V_{ij} + \cdots + V_{1,\ldots,M} \tag{5}$$

In this way, the variance contributions to the total output variance of individual parameters and parameter interactions can be determined. These contributions are characterized by the ratio of partial variance to the total variance, the Sobol' sensitivity indices:

$$\text{First} - \text{order SI: } S_i = \frac{V_i}{V} \tag{6}$$

$$\text{Second} - \text{order SI: } S_{ij} = \frac{V_{ij}}{V} \tag{7}$$

$$\text{Total SI: } S_{Ti} = S_i + \sum_{j \neq i} S_{ij} + \ldots \tag{8}$$

The first order index, S_i, is a measure for the variance contribution of the individual parameter X_i to the total model variance. The partial variance V_i in Equation (6) is given by the variance of the conditional expectation $V_i = V[E(Y|X_i)]$ and is also called the 'main effect' of X_i on Y. It can be defined as the fraction of the model output variance that would disappear on average when X_i would be fixed to a value in its range (because $V(Y) = E[V(Y|X_i)] + V[E(Y|X_i)]$). The effect on the model output variance of the interaction between parameters X_i and X_j is given by S_{ij} and S_{Ti} is the result of the main effect of X_i and all its interactions with the other parameters (up to the M^{th} order).

The calculation of $S_{T\tilde{i}}$ can be based on variance $V_{\tilde{i}}$ that results from the variation of all parameters, except X_i (Homma & Saltelli 1996).

$$S_{Ti} = 1 - \frac{V_{\tilde{i}}}{V} \tag{9}$$

For additive models and assuming orthogonal input factors, S_{Ti} and S_i are equal and the sum of all S_i (and thus, all S_{Ti}) is 1. For non-additive model's interactions exist: S_{Ti} is greater than S_i and the sum of all S_i is less than 1. On the other hand, the sum of all S_{Ti} is greater than 1. By analysing the difference between S_{Ti} and S_i, the effect of interactions between parameter X_i and the other parameters can be calculated.

To compute the variances to obtain the sensitivity measures, Sobol' proposed a shortcut in the calculations, based on the assumption of mutually orthogonal summands in the decomposition. The shortcut is attained by transforming the double-loop integral of Equation (4) into an integral of the product of $f(X_{j_1}, \ldots, X_{j_{k-s'}} X_{i_1}, \ldots, X_{i_s})$ and $f(X'_{j_1}, \ldots, X'_{j_{k-s'}} X_{i_1}, \ldots, X_{i_s})$. Because environmental models are mostly complex and non-linear, it is almost impossible to calculate the variances using analytical integrals. The SIs can be calculated by performing Monte-Carlo simulations.

The evaluation of the SA

Due to its advantageous properties and the drawbacks of the qualitative results of the one-factor-at-a-time (OAT) (Yang 2011) sensitive analysis approach, in this study, an attempt has been made to identify the most sensitive parameters using the Sobol' method. To analyse sensible parameters, the maximum, minimum, and average AT parameters are selected for the Sobol' sensitivity analysis of the model. One thousand independent samples of the parameter sets are generated from the Sobol sequence using the SALib module (Herman & Usher 2017) to assess the second-order sensitivity indices and total sensitivity effects. For the second-order effect, the Saltelli (Saltelli *et al.* 2008) method of the cross-sampling scheme creates a total of $N * (2D + 2)$ parameter sets, where D is the number of input parameters and N is the number of independent samples of the parameter sets. Since no prior knowledge is available on the parameters, the SA's input parameter values were sampled from a uniform distribution (Sobol 1990). The different parameter ranges were scaled between 0 and 1 with normalization. Mean from ±10% changes of AT parameters as the input values to compare the shift in mean response and changes in the entire range of simulated river temperatures. For assessment and comparison purposes, sensitivity indices can be ranked into the four classes found in Table 1 as defined by Lenhart *et al.* (2002). Normalized SIs for RWT model inputs parameters are listed in Table 4.

Table 1 | Sensitivity index categories (Lenhart *et al.* 2002)

Index	Sensitivity
$0.00 \leq \lvert \text{Index} \rvert < 0.05$	Small to negligible
$0.05 \leq \lvert \text{Index} \rvert < 0.20$	Medium
$0.20 \leq \lvert \text{Index} \rvert < 1.00$	High
$\lvert \text{Index} \rvert \geq 1.00$	Very high

Ridge regression

The method of RR was proposed by Hoerl & Kennard (1970). RR is a linear regression extension where the loss function is modified to minimize the model's complexity (Equation (11)). This adjustment is done by adding a penalty parameter equivalent to the square of the magnitude of the coefficients (2-norm or L2 norm (squared)) to avoid overfitting. Equation (10) represents the 2-norm or L2 norm.

$$||w||_2 = (w_1^2 + w_2^2 + \cdots + w_N^2)^{\frac{1}{2}} \tag{10}$$

In this study, an RR model is developed on a daily scale to predict the RWT for the Tunga-Bhadra River with minimum and maximum AT as predictor variables. RR optimizes the following:

$$\begin{aligned}\text{Objective} &= \text{RSS (Residual Sum of Squares)} \\ &\quad + \lambda * \text{(sum of the square of coefficients)}\end{aligned}$$

$$\text{Loss} = \text{Error}(y, \hat{y}\,) + \lambda \sum_{i=1}^{N} w_i^2 \tag{11}$$

KNN regressor

KNN is a simple algorithm (Cover & Hart 1967), and the input consists of the k-closest training samples in the feature space. KNN is to calculate the average of the numerical target of the KNN:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{12}$$

In this study, the KNN model is developed on a daily scale to predict the RWT with minimum and maximum AT as predictor variables. The tuning parameter choices were five neighbors to fit the model.

Support vector regression

Dibike *et al.* (2001) firstly applied the support vector machines (SVMs) approach for accurate simulation of rainfall-runoff processes in hydrology. The SVM is a kernel function learning machine, which follows the structural risk principle (Vapnik *et al.* 1996). When the training data of $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with n patterns, a function $f(x)$ will be identified with the consideration of the deviation from the actually observed target variables y_i for all the training data (Lima *et al.* 2012). The input variable, X, will be mapped into a higher-dimensional feature space using a non-linear mapping function φ.

$$f(x; w) = \langle W, \varphi(x) \rangle + b \tag{13}$$

where <, > denotes the inner product, and W and b are the regression coefficients, which can be estimated by minimizing the error between $f(x)$ and the observed values of y. The SVR uses the $\in$-insensitive error to measure the error between $f(x)$ and the observed values of y.

$$|f(x; w) - y|_{\in} = \begin{cases} 0, & \text{if}|\,f(x; w) - y\,| < \in\,, \\ |\,f(x; w) - y| - \;\in\,, & \text{otherwise,} \end{cases} \tag{14}$$

where $\in$ is the hyper-parameter.

Using the training data of (x_i, y_i), the values of w and b are estimated by minimizing the objective function:

$$F = \frac{C}{N} \sum_{i=1}^{n} |f(x_i, w) - y_i|_{\in} + \frac{1}{2} ||w||^2 \tag{15}$$

where C and $\in$ are the hyper-parameters. The minimization of the objective function, F, uses the Lagrange multiplier method, and the final regression equation with kernel function $K(X, X')$ can be in the form:

$$f(X) = \sum_i K(X, X_i) + b \tag{16}$$

Based on previous studies (Dibike *et al.* 2001; Rehana 2019), Radial Basis Function (RBF) was chosen as the kernel function to measure the performance of the model for the RWT. A detailed introduction to the SVR method may be found in Dibike *et al.* (2001).

RF regressor

RF is designed to produce output by the majority vote (for classification) and the average of the single-tree method (for regression) (Breiman 2001). Each tree creates a set of response predictor values associated with a group of independent values. After that, each independent variable data is splitting into several split points. And the sum of squared error (SSE) has been calculated for each split point between the actual values and the predicted values. This process will recursively continue until the entire data is being covered. There is no interaction between these trees while building the trees. The trees in RFs are run in parallel. The model can be written as:

$$f(x) = f_0(x) + f_1(x) + f_2(x) + \cdots \tag{17}$$

where the final model f is the sum of simple base models f_i. and each base regressor portion is the simple decision tree.

Ensemble Kalman filter

The Kalman filter (KF) (Kalman 1960) technique is one of the DA methods rooted from the Monte-Carlo and Bayesian approaches. EnKF is a variant of KF that can be used for the non-linear filtering problem. The EnKF process is a sequentially based DA method from recent DA research (Evensen 1994). The mathematics involved in EnKF is as follows: X_t^b stands for the prior state estimate ensemble $\{X_{t,1}^b, X_{t,2}^b, \ldots, X_{t,n}^b\}$ at time t; X_t^a stands for the posterior state estimate ensemble $\{X_{t,1}^a, X_{t,2}^a, \ldots, X_{t,n}^a\}$ at time t; and n is the ensemble size. The non-linear process and measurement are expressed as:

$$X_{t+1} = F(X_t) + W_t(N(0, Q)) \tag{18}$$

$$Y_t = H(X_t) + V_t(N(0, R)) \tag{19}$$

where F is a non-linear function that related state X_t at time t to state X_{t+1} at time $t+1$; H is the measurement function that converts state to observation; $W_t(N(0, Q))$ and $V_t(N(0, R))$ represent process and measurement noise, respectively; W_t and V_t are assumed to be independent white noise and white noise with normal probability distributions, and Q and R are processed noise covariance and observation noise covariance matrices, respectively, and are assumed to be constant.

The EnKF algorithm includes two steps: predicting and updating. The prior state estimate is calculated from the posterior estimation in the previous time step in the predicting step. Based on this, the state prior mean and covariance can be calculated as follows:

$$X_{t+1}^b = F(X_t^a) + W_t \tag{20}$$

$$P_{t+1}^b = E[(X_{t+1}^b - \bar{X}_{t+1}^b)(X_{t+1}^b - \bar{X}_{t+1}^b)^T] \tag{21}$$

where P_{t+1}^b represents the prior estimate of covariance, $\bar{X}_{t+1}^b$ represents the state ensemble mean, T represents matrix transposition, and E is the expectation operator. $\bar{X}_{t+1}^b$ is used as the best initial estimate as in Equation (21), and the error covariance is the directly calculated error covariance of the best estimate.

In the updating step, the field observations are treated as a random variable. In order to do this, a sample of observations is generated from a distribution with the mean equal to the field observation and the variance equal to the observation variance R. Using D to stand for the measurement sample matrix, the equations are

$$X_{t+1}^a = X_{t+1}^b + K(D - HX_{t+1}^b) \tag{22}$$

$$P_{t+1}^a = E[(X_{t+1}^a - \bar{X}_{t+1}^a)(X_{t+1}^a - \bar{X}_{t+1}^a)^T] \tag{23}$$

$$K = P_{t+1}^b H^T (HP_{t+1}^b H^T + R)^{-1} \tag{24}$$

where $(D - HX_{t+1}^b)$ is called the residual or measurement innovation. The Kalman gain K in Equation (24) defines the weight to be applied to the actual measurements. In this study, X refers to the temperature parameters, F is the ML model, and D means the water temperature measurements. Measurement error covariance R is determined by the observed dataset D and H as the observation operator.

EnKF model development

In this study, EnKF as a DA technique is implemented to improve the efficiency of ML models in each simulation step. The proposed approach is presented to enhance the performance of the integration of the ML model and

EnKF. For developing the ML model to predict or simulate RWT, EnKF is implemented to update and optimize ML model predictions. Figure 3 shows the ML and DA architectural flow diagram.

In Figure 3, Y_p is the result of ML model prediction, Y_F is the data blended by updating the ML model prediction results with the RWT observations Y_m using the EnKF technique. The steps of this model as follows:

1. The ML model is trained with the observed data at $t-1$ to form the model.
2. The subsequent observations are used to predict the RWT at t.
3. This step updates the predicted data Y_p with the available RWT measurements Y_m using the EnKF technique, and then the updated data Y_F are used as inputs to update the ML model if the error is less than the previous simulation step. The process then returns to step (1) for the next prediction until there are no new data.

MODEL EVALUATION

The accuracy of the applied ML models was evaluated using various goodness-of-fit measures, such as (Chadalawada & Babovic 2017) the coefficient of determination (R^2; Equation (25)), the mean squared error (MSE; Equation (26)), the root-mean-squared error (RMSE; Equation (26)), RMSE-observations standard deviation ratio (RSR; Equation (27); Moriasi *et al.* 2007), Nash–Sutcliffe efficiency (NSE; Equation (28); Nash & Sutcliffe 1970), the mean absolute error (MAE; Equation (29)), and Kling–Gupta efficiency (KGE; Equation (30); Kling *et al.* 2012). For assessment and comparison purposes, RSR and NSE can be ranked into the four classes found in Table 2 as defined by Moriasi *et al.* (2007).

Table 2 | RSR and NSE performance ratings (Moriasi *et al.* 2007)

Performance rating	RSR	NSE
Very good	$0.00 \le \text{RSR} \le 0.50$	$0.75 < \text{NSE} \le 1.00$
Good	$0.50 < \text{RSR} \le 0.60$	$0.65 < \text{NSE} \le 0.75$
Satisfactory	$0.60 < \text{RSR} \le 0.70$	$0.50 < \text{NSE} \le 0.65$
Unsatisfactory	$\text{RSR} > 0.70$	$\text{NSE} \le 0.50$

$$R^2 = 1 - \frac{\sum (T_{w_{\text{pred}}} - T_{w_{obs}})^2}{\sum (T_{w_{obs}} - T_{w_{\text{mean}}})^2} \tag{25}$$

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n} (T_{w_{\text{pred}}} - T_{w_{obs}})^2}{n}} \tag{26}$$

$$\text{RSR} = \frac{\text{RMSE}}{\text{STDEV}_{\text{obs}}} = \frac{\left[\sqrt{\sum_{i=1}^{n} (T_{w_{obs}} - T_{w_{\text{pred}}})^2}\right]}{\left[\sqrt{\sum_{i=1}^{n} (T_{w_{obs}} - T_{w_{\text{mean}}})^2}\right]} \tag{27}$$

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{n} (T_{w_{obs}} - T_{w_{\text{pred}}})^2}{\sum_{i=1}^{n} (T_{w_{obs}} - T_{w_{\text{mean}}})^2}\right] \tag{28}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n} (T_{w_{\text{pred}}} - T_{w_{obs}}) \tag{29}$$

$$\text{KGE} = 1 - \sqrt{(1-r)^2 + (\gamma - 1)^2 + (\beta - 1)^2} \tag{30}$$

$$\beta = \frac{\mu_s}{\mu_0}$$

$$\gamma = \left(\frac{\sigma_s}{\mu_s} / \frac{\sigma_0}{\mu_0}\right)$$

where $T_{w_{\text{pred}}}$ is the predicted daily RWT at time step i in °C; $T_{w_{obs}}$ is the observed daily RWT at time step i in °C; $T_{w_{\text{mean}}}$ is the average daily RWT at time step i in °C; $\text{STDEV}_{\text{obs}}$ is the standard deviation of the observed daily RWT; r is the correlation coefficient between simulated and observed water temperature; β is the bias ratio (the ratio between simulated mean and observed mean), γ is the variability ratio (the ratio between simulated variance and observed variance), μ is the mean; σ is the standard deviation; and n is the number of data pairs in comparison.

Table 3 | Seasonal period Spearman's correlation coefficients between various air and water temperature variables

Season	RWT – maximum AT	RWT – minimum AT	RWT – average AT
Monsoon (Jun–Sep)	0.90	0.18	0.71
Post-monsoon (Oct–Nov)	0.77	0.26	0.59
Winter (Dec–Feb)	0.84	0.20	0.62
Summer (Mar–May)	0.77	0.55	0.76
Annual	0.84	0.31	0.70

Table 4 | Normalized sensitivity indices for RWT model input parameters

Input parameter	Sensitivity indices
Minimum air temperature	0.05
Maximum air temperature	0.95
Average air temperature	0.00

RESULTS AND DISCUSSION

The data used in this paper consist of daily water temperature and corresponding daily minimum, maximum, and mean AT for the period from 1 January 1989 to 1 January 2004. The observed minimum, maximum, and average air (water) temperature mean were noted as 19.66, 29.74, and 24.78 (27.54 °C) and standard deviation as 3.48, 3.47, and 2.77 °C (2.66 °C), respectively. To study the statistical dependency between various air and water temperature variables, Spearman's correlation coefficients have been estimated from 1 January 1989 to 1 January 2004. Spearman's correlation coefficients between RWT and maximum, minimum, and average ATs were calculated. It is observed that RWT is highly significant with the maximum, minimum, and average ATs (p-value < 0.001) (Table 3). Based on the statistical dependency measures, the maximum AT was positively correlated with daily RWT for the case study.

Furthermore, based on the SA (Table 4), it is observed that the maximum AT is highly sensitive, with a sensitivity index of 0.95 in the prediction of RWT compared with the minimum and average ATs. The SA also supports the use of maximum AT as the most important independent variable to be considered in the prediction of RWT. To show the variability of maximum AT with RWT, the daily data from 1 January 1989 to 1 January 2004 have been compared, as shown in Figure 4. Most of the earlier studies considered the average AT as the independent variable in RWT prediction. For example, Rehana & Mujumdar (2011) evaluated the average AT to predict the RWT for the Tunga-Bhadra River at the Shimoga station with the coefficient of determination (R^2) value as 0.53 with discharge as another independent variable. As the present study's main objective is to select an appropriate AT among average, maximum, and minimum to model RWT, the study has not used river discharge in the RWT prediction.

Furthermore, the improved performance in the prediction of RWT with consideration of maximum AT and the

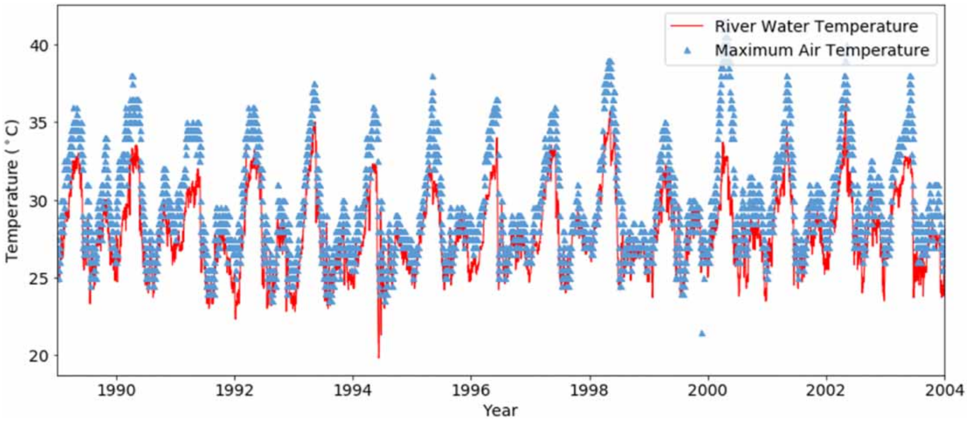

Figure 4 | Time series of daily maximum air temperatures, water temperatures (1989–2004) of the Tunga-Bhadra River at the Shimoga station, India.

average AT was compared with the linear regression model. The resulting R^2 value in RWT prediction was obtained as 0.58 and 0.83 with the average and maximum ATs, respectively. Such improved performance of the RWT prediction model was convincing with an earlier study by Rehana & Mujumdar (2011), which used average AT as the predictor variable in RWT modelling.

To understand the variability of air and water temperature changes for long-term periods, the study estimated the linear trends of both variables (Figure 7(a) and 7(b)). As can be observed, the long-term maximum AT and the RWT are varied during the period from 1989 to 2004 (Figure 5). The monthly seasonal dynamics of RWT and maximum AT based on 15 years averages at the Shimoga station (1989–2004) are presented in Figure 6. It is shown that RWT and maximum AT give a strong seasonal pattern with larger values in summer and lower values in winter. As shown in Figure 7, the long-term AT and the water temperature increased during the period 1989–2004 at the Shimoga station. AT has been increased about 0.077 °C year^{-1}, while RWT increased about 0.062 °C year^{-1}. Such increasing trends of RWT have been investigated in many parts of the world. For example, the observed RWT has shown a growing trend of about 0.029–0.046 °C year^{-1} over China (Chen *et al.* 2016), over the USA of about 0.009–0.077 °C year^{-1} (Isaak *et al.* 2012; van Vliet *et al.* 2013; Rice & Jastram 2015) and Europe as 0.006–0.18 °C year^{-1} (Albek & Albek 2009; Orr *et al.* 2015; Hardenbicker *et al.* 2017). AT increased by 1.0 °C over the 15-year interval from the plot, while the water temperature increased by 0.8 °C. Such increasing air and water temperature trends agreed with the case study's earlier research findings (Rehana & Mujumdar 2011). Furthermore, there is strong evidence of climate change's impact on the river water quality due to the increase of RWTs and decrease of stream flows for the river of interest (e.g., Rehana & Mujumdar 2012; Rehana & Dhanya 2018).

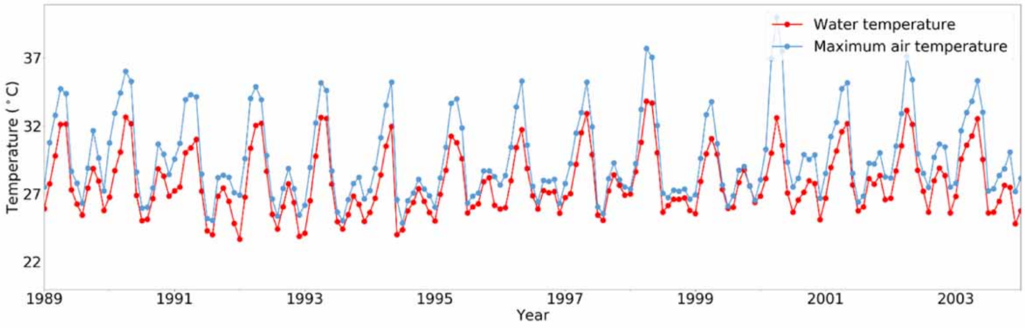

Figure 5 | Time series of monthly mean maximum air temperature and water temperature for the period 1989–2004.

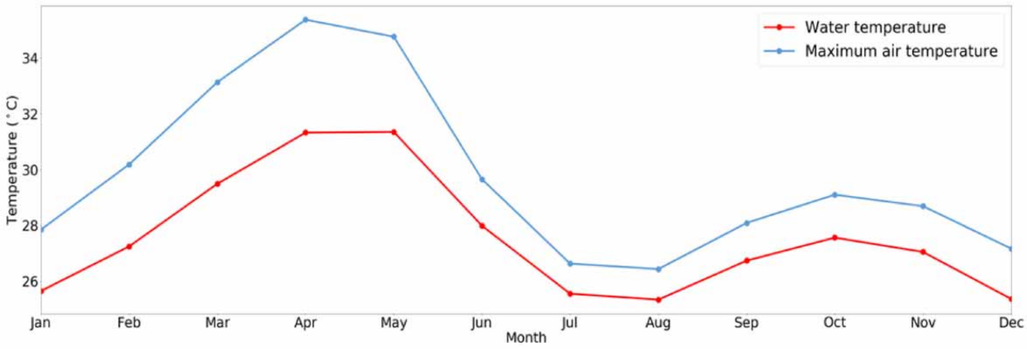

Figure 6 | Monthly mean maximum air temperature and water temperature based on 15 years average at the Shimoga station (1989–2004).

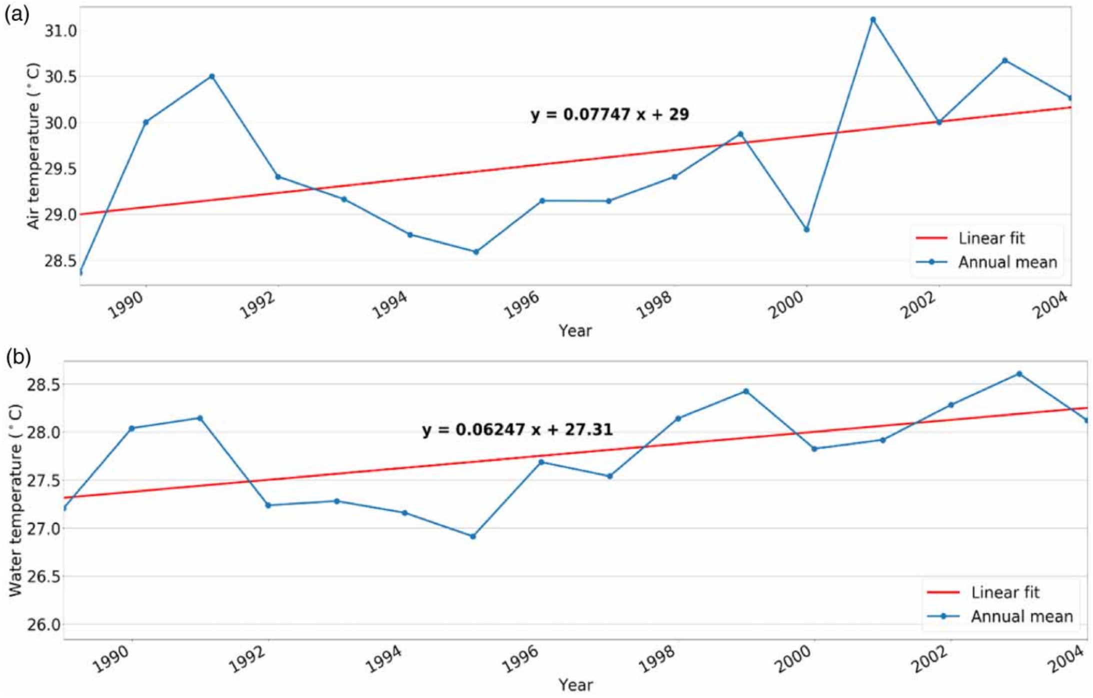

Figure 7 | Time series of annual average (a) maximum air temperatures and (b) water temperatures for 1989–2004.

ML model performance

The next step in the prediction of RWT is to use appropriate ML, which can work accurately in terms of calibration and validation with a comparison of acceptable performance measures, as shown in Figure 2. To utilize the data better, assessing the effectiveness of the model and avoid overfitting, the cross-validation (CV) technique was applied. When dealing with time-series data, traditional CV (like k-fold) cannot be used since the adjacent data points are often highly dependent, so standard CV will fail. To overcome these issues, the time-series splits CV technique was used in the present study (Pedregosa *et al.* 2011; Scavuzzo *et al.* 2018). This CV was performed chronologically, started with a small subset of data for training purposes, estimated the last data points, and then checked the accuracy for the calculated data points. The same estimated data points are then included as part of the next training dataset, and subsequent data points were estimated. This CV procedure provides an almost unbiased estimate of the true error (Varma & Simon 2006). The error on each split is averaged in order to compute a robust estimate of model error, as shown in Figure 2. While fitting a model on a dataset, all the possible combinations of parameter values are evaluated using the GridSearchCV python library module (Pedregosa *et al.* 2011), and the best combination is taken to make the model performant.

The results of the ML approaches (Ridge, KNN, RF, and SVR) for the prediction of RWT were evaluated using several goodness-of-fit statistics (MSE, MAE, RMSE, RSR, NSE, and R^2), and graphical tools (seasonal plots and box plots). The experiment results showed a good trade-off between training and validation performance, confirming the stable generalization capacity of ML approaches. The developed models were able to predict RWT using AT as input successfully. Figure 8 shows the box plot for observed and predicted RWT using Ridge, KNN, RF, and SVR models, and it is observed that the minimum RWT is 21 °C and max RWT is 31 °C for the observed data while the lower and quartile range between 24 and 28 °C with median RWT of 26 °C. According to Figure 8, all the four models performed almost comparable predictions with a difference of 1 °C based on the median, and there is a clear resemblance between the observed RWT and the

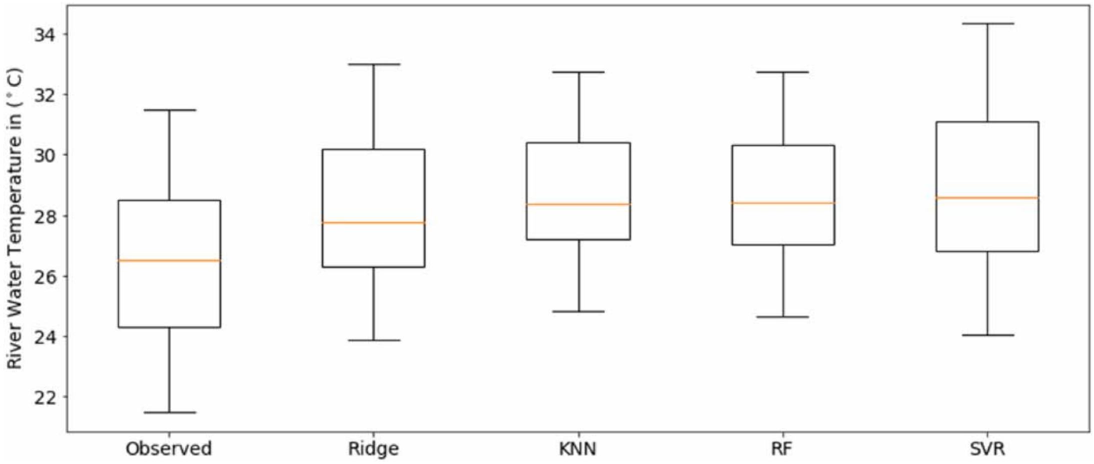

Figure 8 | Box plots of observed and calculated RWT (°C) in the validation phase with the four ML models.

predicted value, in addition the lower and the upper quartile ranges predicted using these models were marginally varied compared with the observed data.

The performance of the Ridge, KNN, RF, and SVR models for daily data at the Shimoga station is provided in Table 5 and Figure 9. Results showed that the seasonal variations of predicted RWT are almost synchronous and comparable with the observed values (Figure 9), but the Ridge model performed poorly with overestimated values in high water temperature period and performance statistics (R^2, MSE, RMSE RSR, NSE, and MAE) can be found in Table 5. From Table 5, the SVR ($R^2 = 0.84$, KGE = 0.86, MSE = 0.99, RMSE = 0.99, RSR = 0.40, NSE = 0.84, and MAE = 0.77) model has performed slightly better than KNN ($R^2 = 0.82$, KGE = 0.87, MSE = 1.11, RMSE = 1.05, RSR = 0.42, NSE = 0.82, and MAE = 0.84), RF ($R^2 = 0.83$,

Table 5 | Performances of different models in the prediction of RWT for the period of 1989–2004

Data	Model	R^2	KGE	MSE	RMSE	RSR	NSE	MAE
Daily	Ridge	0.76	0.87	1.44	1.01	0.31	0.76	0.90
	KNN	0.82	0.87	1.11	1.05	0.42	0.82	0.84
	RF	0.83	0.87	1.05	1.03	0.41	0.83	0.81
	SVR	0.84	0.86	0.99	0.99	0.40	0.84	0.77
Monthly	Ridge	0.79	0.87	1.02	1.00	0.35	0.79	0.74
	KNN	0.85	0.85	0.87	0.93	0.38	0.84	0.74
	RF	0.87	0.94	0.71	0.84	0.39	0.87	0.67
	SVR	0.88	0.88	0.61	0.78	0.39	0.88	0.57
Season (Jan–Apr)	Ridge	0.64	0.72	1.93	1.38	0.30	0.64	1.06
	KNN	0.76	0.90	1.42	1.19	0.35	0.76	0.97
	RF	0.80	0.89	1.15	1.07	0.36	0.80	0.86
	SVR	0.82	0.92	1.00	1.00	0.36	0.82	0.80
Season (May–Aug)	Ridge	0.84	0.88	1.42	1.19	0.27	0.84	0.88
	KNN	0.86	0.89	1.30	1.14	0.28	0.85	0.86
	RF	0.87	0.86	1.17	1.08	0.28	0.87	0.82
	SVR	0.87	0.95	1.18	1.08	0.28	0.86	0.76
Season (Sep–Dec)	Ridge	0.52	0.86	0.71	0.84	0.56	0.52	0.68
	KNN	0.50	0.70	0.77	0.88	0.53	0.49	0.69
	RF	0.53	0.72	0.73	0.85	0.53	0.52	0.68
	SVR	0.61	0.74	0.61	0.78	0.58	0.60	0.60

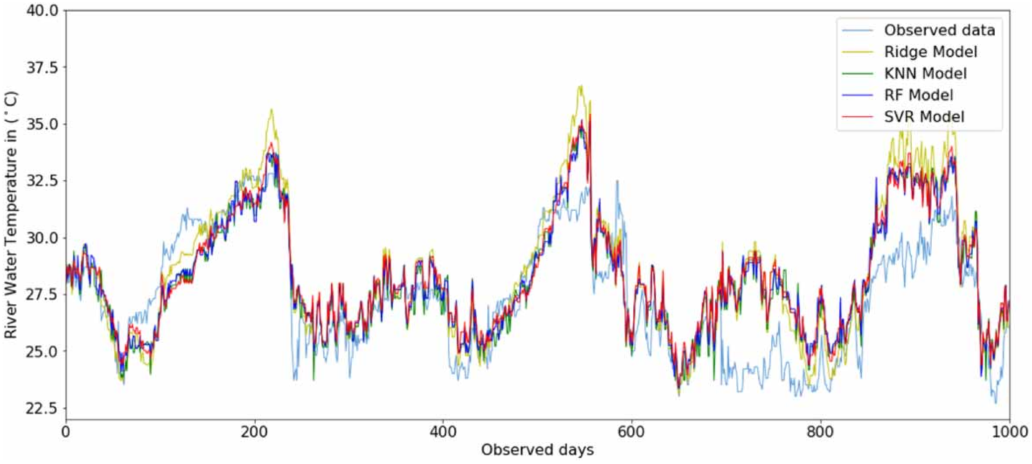

Figure 9 | Comparison between the daily predicted values and observed values of RWT (°C) in the validation phase, with the four ML models.

KGE = 0.87, MSE = 1.05, RMSE = 1.03, RSR = 0.41, NSE = 0.83, and MAE = 0.81), and Ridge (R^2 = 0.76, KGE = 0.87, MSE = 1.44, RMSE = 1.01, RSR = 0.31, NSE = 0.76, and MAE = 0.90) for daily time scale. The accuracy for the ML approaches showed excellent performance in terms of NSE (NSE >0.75) and RSR (RSR <0.50) (Moriasi *et al.* 2007; Table 2) with lower values of MSE and RMSE. The relationship between daily RWT and maximum AT at the Shimoga station has a relatively strong correlated value for all four models (R^2 values). The RMSE values for the Shimoga station range from 0.99 to 1.05 for all the four ML models (Table 5) for daily data, which are reasonable compared with Jackson *et al.* (2018) (1.57) and Sohrabi *et al.* (2017) (1.25), and far better than that of Temizyurek & Dadaser-Celik (2018) (2.10–2.64). Based on RSR and NSE performance ratings (Moriasi *et al.* 2007; Table 2), the best performing model was noted as the SVR (NSE = 0.84; KGE = 0.86; R^2 = 0.84; RSR <0.50) for RWT prediction based on the performance measures (Table 5) for daily time scale. The superiority of SVR in the prediction of RWT as revealed in the present study was found to agree with the study of Rehana (2019) for the same case study. However, it can be noted that the study by Rehana (2019) used the average AT as the independent variable without testing for the most influencing AT variables in the prediction of RWT, as demonstrated in the present study. Furthermore, it can also be noted that the model performance has improved using the SVR with maximum AT (NSE: 0.84 and RMSE: 0.99) as an independent variable compared with the average AT (NSE: 0.61 and RMSE: 1.69) (Rehana 2019) for the same case study at daily time scale.

A summary of the Ridge, KNN, RF, and SVR model performances for monthly data is illustrated in Table 5 and Figure 10. ML results showed that the seasonal variations of predicted RWT are almost synchronous and comparable with the observed values (Figure 10), but the Ridge model performed poorly with overestimated values in high water temperature period and performance statistics are given in Table 5. Compared with the four ML models, the SVR (R^2 = 0.88, KGE = 0.88, MSE = 0.61, RMSE = 0.78, RSR = 0.39, NSE = 0.88, and MAE = 0.57) model performed slightly better than KNN (R^2 = 0.85, KGE = 0.85, MSE = 0.87, RMSE = 0.93, RSR = 0.38, NSE = 0.84, and MAE = 0.74), RF (R^2 = 0.87, KGE = 0.94, MSE = 0.71, RMSE = 0.84, RSR = 0.39, NSE = 0.87, and MAE = 0.67), and Ridge (R^2 = 0.79, KGE = 0.87, MSE = 1.02, RMSE = 1.00, RSR = 0.35, NSE = 0.79, and MAE = 0.74) for monthly time scale. It can be noticed that performance coefficients of monthly time scale were improved in terms of higher R^2, NSE, and lower RMSE and MAE values when compared with daily time scale (Table 5). The ML model accuracy has been increased with monthly data for RWT predictions compared with daily data, with SVR (RSR = 0.39; NSE = 0.88), RF (RSR = 0.39; NSE = 0.87), KNN (RSR = 0.38; NSE = 0.84), and Ridge (RSR = 0.35; NSE = 0.79) showed very good performance based on RSR and NSE performance ratings (Moriasi *et al.* 2007; Table 2).

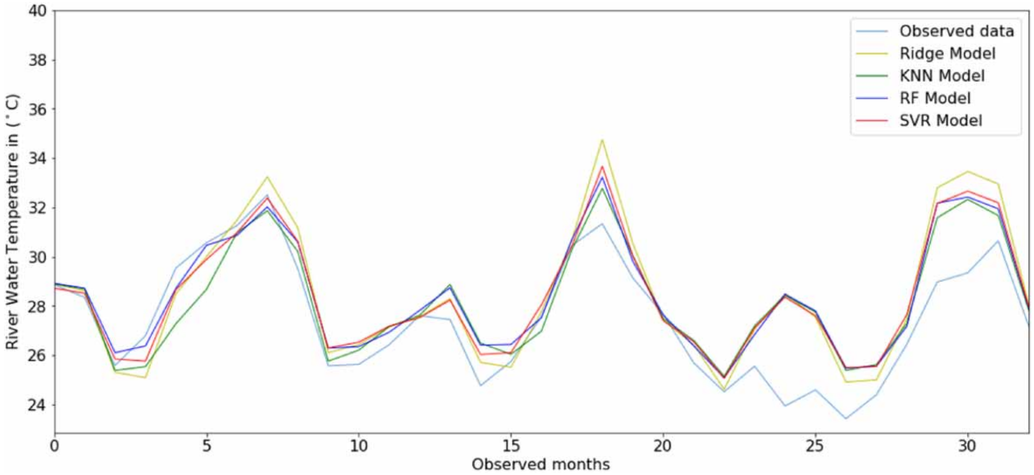

Figure 10 | Comparison between the monthly predicted values and observed values of RWT (°C) in the validation phase, with the four ML models.

The performance of the Ridge, KNN, RF, and SVR models for seasonal data (Jan–Apr, May–Aug, and Sep–Dec) (Laizé *et al.* 2017; Zhu *et al.* 2019c) is shown in Figure 11. Results showed that the seasonal variations of predicted RWT are almost in agreement with the observed values (Figure 11), but the Ridge model performed poorly with overestimated values in high water temperature periods and performance statistics are given in Table 5. From Table 5, the SVR model performed slightly better than KNN, RF, and Ridge in all three seasons (Jan–Apr, May–Aug, and Sep–Dec). It can be noticed that NSE and RSR values were poor for the season (Sep–Dec) when compared with the other two seasons, daily time scale and monthly time scale values. Table 5 shows that the four models constructed in this paper may learn the RWT variation rules from the historical data and reproduce the seasonal dynamics of RWT. This case study demonstrates that integrating the scientific knowledge into ML tools promises to improve many important environmental variables predictions.

ML-EnKF model performance

In the next step in the prediction of RWT, the EnKF DA technique is implemented to improve the efficiency of ML models in each simulation step. Table 6 shows the results of the ML-EnKF model at different simulation steps with the assimilated data. Table 6 shows that the blended data show the improved results from simulation-1 (1 January 2001 to 1 January 2002) to simulation-2 (1 January 2002 to 1 January 2003). These results demonstrate that the blended data are best. It can be concluded that the ML-EnKF model can do a better job with assimilated data in RWT prediction. It dramatically enhances the direct ML models. If the simulation steps continue, the ML-EnKF model is improved and the simulation results are significantly improved, according to Table 6. As the first section states, the ML-EnKF model is designed to improve the ML model performance by a combination of both ML models and a DA approach to enhance the predicted values based on the measurement data.

CONCLUSIONS

ML techniques represent a potentially disruptive force for many scientific disciplines. The purpose of this study was to assess the performance of a suite of ML models for RWT prediction for the Tunga-Bhadra River, India, with the aid of the minimum and maximum AT at daily, monthly, and seasonal time scales. In this study, an attempt has been made to identify the most sensitive AT variable (average, maximum, and minimum) using the Sobol' sensitivity analysis method, which can serve as an input variable in the prediction of RWT. The results indicated that the maximum AT was the most important variable in the prediction of

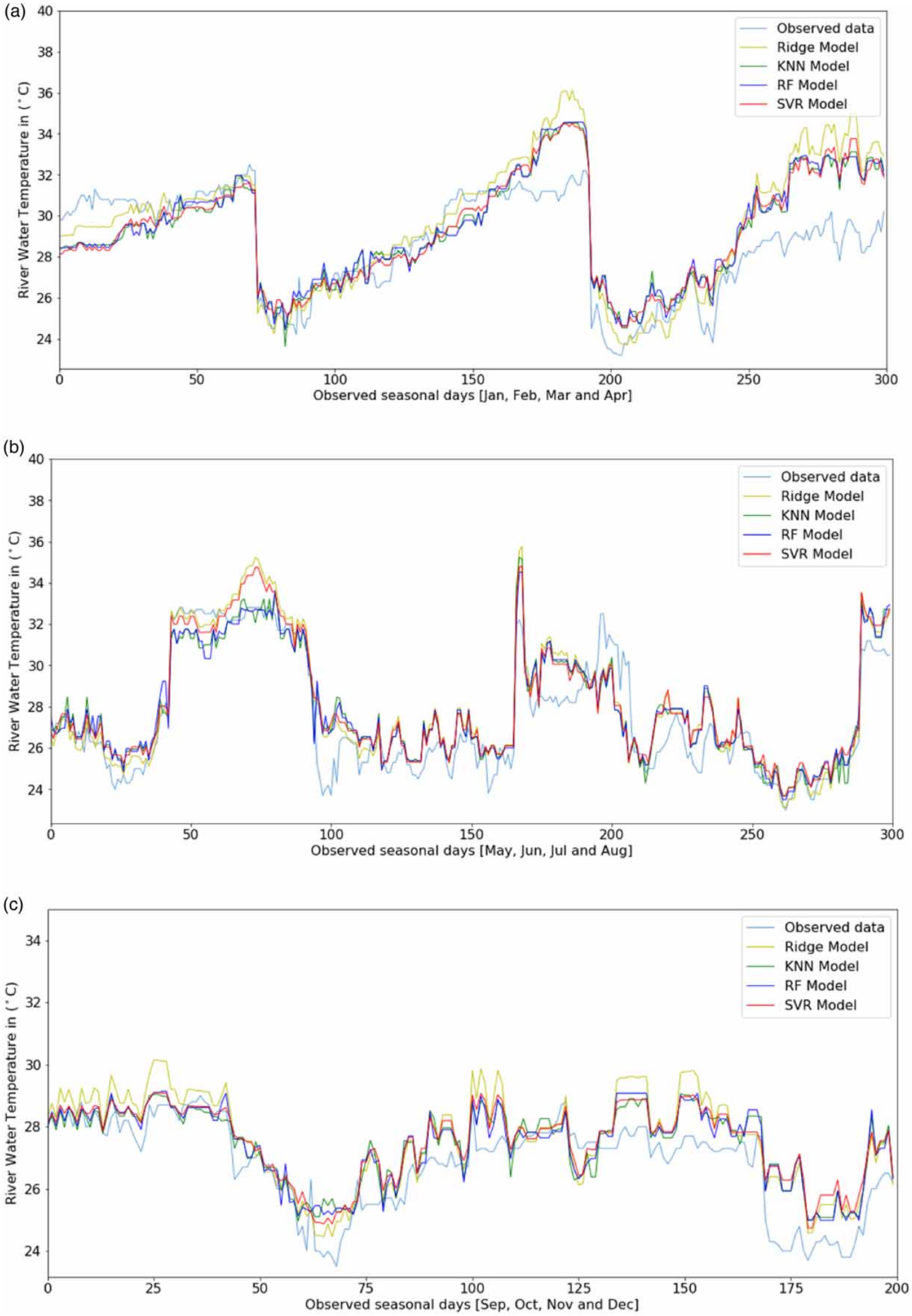

Figure 11 | Comparison between the (a) Jan–Apr months, (b) May–Aug months, and (c) Sep–Dec months seasonal predicted values and observed values of RWT (°C) in the validation phase, with the four ML models.

Table 6 | Performances of different models with assimilated data in the prediction of RWT

Data	Model	R^2	KGE	MSE	RMSE	RSR	NSE	MAE
Simulation-1 (1 Jan 2001 to 1 Jan 2002)	Ridge	0.829	0.807	0.829	0.910	0.413	0.829	0.759
	KNN	0.855	0.925	0.699	0.836	0.379	0.855	0.667
	RF	0.860	0.934	0.676	0.822	0.373	0.860	0.656
	SVR	0.886	0.915	0.555	0.745	0.338	0.885	0.593
Simulation-2 (1 Jan 2002 to 1 Jan 2003)	Ridge	0.867	0.843	0.841	0.917	0.363	0.867	0.710
	KNN	0.855	0.883	0.921	0.959	0.379	0.856	0.764
	RF	0.865	0.880	0.898	0.947	0.375	0.859	0.741
	SVR	0.911	0.921	0.564	0.741	0.303	0.908	0.573

RWT for the river location of interest. In general, it can be concluded that the Sobol' sensitivity analysis can be successfully applied for input variable fixing and prioritization of any RWT model. Therefore, the Sobol' sensitivity analysis method can be considered as a robust and powerful sensitivity analysis method for RWT prediction modelling.

Furthermore, each model's configurable variable is optimized, and the performances of various ML models are analysed to test the applicability of the data-driven models in the RWT being investigated. The study revealed that ML model performance coefficients are improved in monthly data compared with the daily time scale. The seasonal time scale RWT prediction models also performed poorly compared with daily and monthly time scale data. Overall, the monthly time scale RWT prediction ML models have performed better than daily and seasonal for interest study location. The SVR has been noted as the most robust ML model to predict RWT. Furthermore, the EnKF DA algorithm with ML approaches improves the predicted values based on the measurement data. The ML-EnKF model update of the prediction data with the observed data using the DA method shows a better result. Generally, the assimilation method is just considered to bring model predictions close to the observations rather than improve the model structure. Here, as the updated data are used to train the ML model for the next prediction, it does enhance the model and makes the model more practical in hydrologic applications. If the simulation steps continue, the ML-EnKF model is improved and the simulation results are significantly improved.

This case study demonstrated how a data-driven modelling framework could be scaled up and used for the prediction of RWT. The DA methods can also combine with ML models to improve the predicted values based on the measurement data. Overall, the data-driven modelling framework presented in the study indicated that all ML models were proven to be effective in RWT prediction. This case study demonstrates that integrating scientific knowledge into ML tools for improving predictions of many important environmental variables and the applicability of data-driven models in the field of the water sector. Simultaneously, ML models architecture and the law of parameter setting demonstrated in the present study can be valuable for the river water quality management problems.

Despite the robustness of the modelling frameworks as presented in the study, it has some caveats. One of the major limitations of the study is consideration of the data for the period from 1989 to 2004, which is the only long period of data available along the river stretch with minimal missing and erroneous data. The proposed modelling framework of RWT prediction can always be implemented with newly updated data as demonstrated in the present study, which can be extended to other stations based on data availability. RWT prediction models should consider the spatial dependency of air and water temperature variables when the modelling framework is proposed to be implemented with multiple stations of a river stretch. Furthermore, the study demonstrated the modelling framework to consider the most sensitive variables in predicting RWT using various AT variables, such as average, maximum, and minimum. However, there are several variables, which have a direct impact on RWT, such as streamflow (Isaak *et al.* 2010; Toffolon & Piccolroaz 2015; Sohrabi *et al.* 2017) and river geometry (Gu & Li 2002), which need to be considered in the sensitivity analysis and consequently in the ML algorithms. Further research into the robust and hybrid approaches to RWT modelling is required, as an accurate simulation of RWT plays an important role in water resources management.

ACKNOWLEDGEMENT

We thank the Advanced Centre for Integrated Water Resources Management (ACIWRM), Bengaluru, Karnataka, India, for providing the water quality data. The research work is supported by the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India, through Core Research Grant Project No. CRG/2020/002028 to Dr. S. Rehana.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Ahmadi-Nedushan, B., St-Hilaire, A., Ouarda, T. B. M. J., Bilodeau, L., Robichaud, É., Thiémonge, N. & Bobée, B. 2007 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.

Albek, M. & Albek, E. 2009 Stream temperature trends in Turkey. *CLEAN – Soil, Air, Water* **37**, 142–149. https://doi.org/10.1002/clen.200700159.

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A. & Oliveira, M. S. 2018 Short-term water demand forecasting using machine learning techniques. *Journal of Hydroinformatics* **20**, 1343–1366. https://doi.org/10.2166/hydro.2018.163.

Babovic, V. 2005. Data mining in hydrology. *Hydrological Processes* **19**, 1511–1515. https://doi.org/10.1002/hyp.5862.

Babovic, V., Sannasiraj, S. A. & Chan, E. S. 2005 Error correction of a predictive ocean wave model using local model approximation. *Journal of Marine Systems* **53**, 1–17. https://doi.org/10.1016/j.jmarsys.2004.05.028.

Balk, B. & Elder, K. 2000 Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. *Water Resources Research* **36**, 13–26. https://doi.org/10.1029/1999WR900251.

Beersma, J. & Buishand, T. 2004 Joint probability of precipitation and discharge deficits in the Netherlands. *Water Resources Research* **40**. https://doi.org/10.1029/2004WR003265.

Bogan, T., Mohseni, O. & Stefan, H. G. 2003 Stream temperature-equilibrium temperature relationship. *Water Resources Research* **39**. https://doi.org/10.1029/2003WR002034.

Boukabara, S.-A., Krasnopolsky, V., Penny, S. G., Stewart, J. Q., McGovern, A., Hall, D., Hoeve, J. E. T., Hickey, J., Huang, H.-L. A., Williams, J. K., Ide, K., Tissot, P., Haupt, S. E., Casey, K. S., Oza, N., Geer, A. J., Maddy, E. S. & Hoffman, R. N. 2020 Outlook for exploiting artificial intelligence in the earth and environmental sciences. *Bulletin of the American Meteorological Society* **1**, 1–53. https://doi.org/10.1175/BAMS-D-20-0031.1.

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

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

Central Water Commission 2018 *Hydro-Meteorological Data Dissemination Policy*. http://www.cwc.gov.in/sites/default/files/hddp2018_0.pdf.

Chadalawada, J. & Babovic, V. 2017 Review and comparison of performance indices for automatic model induction. *Journal of Hydroinformatics* **21**, 13–31. https://doi.org/10.2166/hydro.2017.078.

Chen, D., Hu, M., Guo, Y. & Dahlgren, R. A. 2016 Changes in river water temperature between 1980 and 2012 in Yongan watershed, eastern China: magnitude, drivers and models. *Journal of Hydrology* **533**, 191–199. https://doi.org/10.1016/j.jhydrol.2015.12.005.

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

Cibin, R., Sudheer, K. P. & Chaubey, I. 2010 Sensitivity and identifiability of stream flow generation parameters of the SWAT model. *Hydrological Processes* **24**, 1133–1148. https://doi.org/10.1002/hyp.7568.

Cloke, H. L., Pappenberger, F. & Renaud, J.-P. 2008 Multi-method global sensitivity analysis (MMGSA) for modelling floodplain hydrological processes. *Hydrological Processes* **22**, 1660–1674. https://doi.org/10.1002/hyp.6734.

Cover, T. & Hart, P. 1967 Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13**, 21–27. https://doi.org/10.1109/TIT.1967.1053964.

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

Dibike, Y., Velickov, S., Solomatine, D. & Abbott, M. 2001 Model induction with support vector machines: introduction and applications. *Journal of Computing in Civil Engineering* **15**. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208).

Erickson Troy, R. & Stefan Heinz, G. 2000 Linear air/water temperature correlations for streams during open water periods. *Journal of Hydrologic Engineering* **5**, 317–321. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:3(317).

Evensen, G. 1994 Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99**, 10143–10162. https://doi.org/10.1029/94JC00572.

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

Gavahi, K., Mousavi, S. J. & Ponnambalam, K. 2019 Adaptive forecast-based real-time optimal reservoir operations: application to Lake Urmia. *Journal of Hydroinformatics* **21**, 908–924. https://doi.org/10.2166/hydro.2019.005.

Geer, A. 2020 Learning earth system models from observations: machine learning or data assimilation? https://doi.org/10.21957/7fyj2811r.

Graf, R., Zhu, S. & Sivakumar, B. 2019 Forecasting river water temperature time series using a wavelet-neural network hybrid modelling approach. *Journal of Hydrology* **578**, 124115. https://doi.org/10.1016/j.jhydrol.2019.124115.

Gu, R. R. & Li, Y. 2002 River temperature sensitivity to hydraulic and meteorological parameters. *Journal of Environmental Management* **66**, 43–56. https://doi.org/10.1006/jema.2002.0565.

Hadzima-Nyarko, M., Rabi, A. & Šperac, M. 2014 Implementation of artificial neural networks in modeling the water-air temperature relationship of the river Drava. *Water Resources Management* **28**, 1379–1394. https://doi.org/10.1007/s11269-014-0557-7.

Hardenbicker, P., Viergutz, C., Becker, A., Kirchesch, V., Nilson, E. & Fischer, H. 2017 Water temperature increases in the river Rhine in response to climate change. *Regional Environmental Change* **17**, 299–308. https://doi.org/10.1007/s10113-016-1006-3.

Heddam, S. & Kisi, O. 2018 Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology* **559**, 499–509. https://doi.org/10.1016/j.jhydrol.2018.02.061.

Herman, J. & Usher, W. 2017 SALib: an open-source Python library for Sensitivity Analysis. *Journal of Open Source Software* **2**, 97. https://doi.org/10.21105/joss.00097.

Hoerl, A. E. & Kennard, R. W. 1970 Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Homma, T. & Saltelli, A. 1996 Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety* **52**, 1–17. https://doi.org/10.1016/0951-8320(96)00002-6.

Huang, F., Huang, J., Jiang, S.-H. & Zhou, C. 2017 Prediction of groundwater levels using evidence of chaos and support vector machine. *Journal of Hydroinformatics* **19**, 586–606. https://doi.org/10.2166/hydro.2017.102.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. & Yasmeen, F. 2018. forecast: Forecasting functions for time series and linear models. R package version 8.0, http://github.com/robjhyndman/forecast.

Isaak, D. J., Luce, C. H., Rieman, B. E., Nagel, D. E., Peterson, E. E., Horan, D. L., Parkes, S. & Chandler, G. L. 2010 Effects of climate change and wildfire on stream temperatures and salmonid thermal habitat in a mountain river network. *Ecological Applications* **20**, 1350–1371. https://doi.org/10.1890/09-0822.1.

Isaak, D. J., Wollrab, S., Horan, D. & Chandler, G. 2012 Climate change effects on stream and river temperatures across the northwest U.S. from 1980–2009 and implications for salmonid fishes. *Climatic Change* **113**, 499–524. https://doi.org/10.1007/s10584-011-0326-z.

Jackson, F. L., Fryer, R. J., Hannah, D. M., Millar, C. P. & Malcolm, I. A. 2018 A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change. *Science of The Total Environment* **612**, 1543–1558. https://doi.org/10.1016/j.scitotenv.2017.09.010.

Johnson, M. F., Wilby, R. L. & Toone, J. A. 2014 Inferring air–water temperature relationships from river and catchment properties. *Hydrological Processes* **28**, 2912–2928. https://doi.org/10.1002/hyp.9842.

Kalman, R. E. 1960 A new approach to linear filtering and prediction problem. *Transactions of the AMSE – Journal of Basic Engineering* **82** (D), 35–45.

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

Komasi, M., Sharghi, S. & Safavi, H. R. 2018 Wavelet and cuckoo search-support vector machine conjugation for drought forecasting using Standardized Precipitation Index (case study: Urmia Lake, Iran). *Journal of Hydroinformatics* **20**, 975–988. https://doi.org/10.2166/hydro.2018.115.

Krider, L. A., Magner, J. A., Perry, J., Vondracek, B. & Ferrington, L. C. 2013 Air-water temperature relationships in the trout streams of southeastern Minnesota's carbonate-sandstone landscape. *JAWRA Journal of the American Water Resources Association* **49**, 896–907. https://doi.org/10.1111/jawr.12046.

Laanaya, F., St-Hilaire, A. & Gloaguen, E. 2017 Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models. *Hydrological Sciences Journal* **62**, 1078–1093. https://doi.org/10.1080/02626667.2016.1246799.

Laizé, C. L. R., Bruna Meredith, C., Dunbar, M. J. & Hannah, D. M. 2017 Climate and basin drivers of seasonal river water temperature dynamics. *Hydrology and Earth System Sciences* **21**, 3231–3247. https://doi.org/10.5194/hess-21-3231-2017.

Leander, R., Buishand, A., Aalders, P. & Wit, M. D. 2005 Estimation of extreme floods of the River Meuse using a stochastic weather generator and a rainfall–runoff model / Estimation des crues extrêmes de la Meuse à l'aide d'un générateur stochastique de variables météorologiques et d'un modèle pluie–débit. *Hydrological Sciences Journal* **50**, 1103. https://doi.org/10.1623/hysj.2005.50.6.1089.

Lenhart, T., Eckhardt, K., Fohrer, N. & Frede, H.-G. 2002 Comparison of two different approaches of sensitivity analysis. *Physics and Chemistry of the Earth, Parts A/B/C* **27**, 645–654. https://doi.org/10.1016/S1474-7065(02)00049-9.

Li, X.-L., Lü, H., Horton, R., An, T. & Yu, Z. 2013 Real-time flood forecast using the coupling support vector machine and data assimilation method. *Journal of Hydroinformatics* **16**, 973–988. https://doi.org/10.2166/hydro.2013.075.

Li, M., Zhang, Y., Wallace, J. & Campbell, E. 2020 Estimating annual runoff in response to forest change: a statistical method based on random forest. *Journal of Hydrology* **589**, 125168. https://doi.org/10.1016/j.jhydrol.2020.125168.

Lima, A. R., Cannon, A. J. & Hsieh, W. W. 2012 Downscaling temperature and precipitation using support vector regression with evolutionary strategy. In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. https://doi.org/10.1109/IJCNN.2012.6252383.

Liu, D., Yu, Z. & H, L. 2010 Data assimilation using support vector machines and ensemble Kalman filter for multi-layer soil moisture prediction. *Water Science and Engineering* **3**, 361–377. https://doi.org/10.3882/j.issn.1674-2370.2010.04.001.

Lu, H. & Ma, X. 2020 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.

Mehrparvar, M. & Asghari, K. 2018 Modular optimized data assimilation and support vector machine for hydrologic modeling. *Journal of Hydroinformatics* **20**, 728–738. https://doi.org/10.2166/hydro.2018.009.

Mohseni, O., Stefan, H. G. & Erickson, T. R. 1998 A nonlinear regression model for weekly stream temperatures. *Water Resources Research* **34**, 2685–2692. https://doi.org/10.1029/98WR01877.

Mohseni, O., Erickson, T. R. & Stefan, H. G. 1999 Sensitivity of stream temperatures in the United States to air temperatures projected under a global warming scenario. *Water Resources Research* **35**, 3723–3733. https://doi.org/10.1029/1999WR900193.

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R. D. & Veith, T. 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* **50**. https://doi.org/10.13031/2013.23153.

Morrill, J. C., Bales, R. C. & Conklin, M. H. 2005 Estimating stream temperature from air temperature: implications for future water quality. *Journal of Environmental Engineering* **131**, 139–146. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:1(139).

Morrison, J. & Foreman, M. G. G. 2005 Forecasting Fraser river flows and temperatures during upstream salmon migration. *Journal of Environmental Engineering and Science* **4**, 101–111. https://doi.org/10.1139/s04-046.

Muluye, G. Y. 2012 Comparison of statistical methods for downscaling daily precipitation. *Journal of Hydroinformatics* **14**, 1006–1023. https://doi.org/10.2166/hydro.2012.197.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10**, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Neumann David, W., Balaji, R. & Zagona Edith, A. 2003 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).

Nossent, J., Elsen, P. & Bauwens, W. 2011 Sobol' sensitivity analysis of a complex environmental model. *Environmental Modelling & Software* **26**, 1515–1525. https://doi.org/10.1016/j.envsoft.2011.08.010.

Orr, H. G., Simpson, G. L., des Clers, S., Watts, G., Hughes, M., Hannaford, J., Dunbar, M. J., Laizé, C. L. R., Wilby, R. L., Battarbee, R. W. & Evans, R. 2015 Detecting changing river temperatures in England and Wales. *Hydrological Processes* **29**, 752–766. https://doi.org/10.1002/hyp.10181.

Ouellet-Proulx, S., Chimi Chiadjeu, O., Boucher, M.-A. & St-Hilaire, A. 2017 Assimilation of water temperature and discharge data for ensemble water temperature forecasting. *Journal of Hydrology* **554**, 342–359. https://doi.org/10.1016/j.jhydrol.2017.09.027.

Pappenberger, F., Beven, K. J., Ratto, M. & Matgen, P. 2008 Multi-method global sensitivity analysis of flood inundation models. *Advances in Water Resources* **31**, 1–14. https://doi.org/10.1016/j.advwatres.2007.04.009.

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. 2011 Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Piccolroaz, S., Calamita, E., Majone, B., Gallice, A., Siviglia, A. & Toffolon, M. 2016 Prediction of river water temperature: a comparison between a new family of hybrid models and statistical approaches. *Hydrological Processes* **30**, 3901–3917. https://doi.org/10.1002/hyp.10913.

Pike, A., Danner, E., Boughton, D., Melton, F., Nemani, R., Rajagopalan, B. & Lindley, S. 2013 Forecasting river temperatures in real time using a stochastic dynamics approach. *Water Resources Research* **49**, 5168–5182. https://doi.org/10.1002/wrcr.20389.

Pilgrim, J. M., Fang, X. & Stefan, H. G. 1998 Stream temperature correlations with Air temperatures in Minnesota: implications for climate warming1. *JAWRA Journal of the American Water Resources Association* **34**, 1109–1121. https://doi.org/10.1111/j.1752-1688.1998.tb04158.x.

Piotrowski, A. P., Napiorkowski, M. J., Napiorkowski, J. J. & Osuch, M. 2015 Comparing various artificial neural network types for water temperature prediction in rivers. *Journal of*

Hydrology **529**, 302–315. https://doi.org/10.1016/j.jhydrol.2015.07.044.

Rabi, A., Hadzima-Nyarko, M. & Šperac, M. 2015 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.

Rasouli, K., Hsieh, W. W. & Cannon, A. J. 2012 Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology* **414–415**, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039.

Rehana, S. & Mujumdar, P. P. 2011 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.

Rehana, S. & Mujumdar, P. 2012 Climate change induced risk in water quality control problems. *Journal of Hydrology* **444–445**, 63–77. https://doi.org/10.1016/j.jhydrol.2012.03.042.

Rehana, S. & Dhanya, C. T. 2018 Modeling of extreme risk in river water quality under climate change. *Journal of Water and Climate Change* **9**, 512–524. https://doi.org/10.2166/wcc.2018.024.

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

Rice, K. C. & Jastram, J. D. 2015 Rising air and stream-water temperatures in Chesapeake Bay region, USA. *Climatic Change* **128**, 127–138. https://doi.org/10.1007/s10584-014-1295-9.

Sahoo, G. B., Schladow, S. G. & Reuter, J. E. 2009 Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. *Journal of Hydrology* **378**, 325–342. https://doi.org/10.1016/j.jhydrol.2009.09.037.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. 2008 *Global Sensitivity Analysis. The Primer*. https://doi.org/10.1002/9780470725184.ch6.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. & Tarantola, S. 2010 Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications* **181**, 259–270. https://doi.org/10.1016/j.cpc.2009.09.018.

Scavuzzo, J. M., Trucco, F., Espinosa, M., Tauro, C. B., Abril, M., Scavuzzo, C. M. & Frery, A. C. 2018 Modeling Dengue vector population using remotely sensed data and machine learning. *Acta Tropica* **185**, 167–175. https://doi.org/10.1016/j.actatropica.2018.05.003.

Sobol, I. 1990 On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie* **1**, 112–118.

Sobol', I. M. 2001 Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation* **55**, 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6.

Sohrabi, M. M., Benjankar, R., Tonina, D., Wenger, S. J. & Isaak, D. J. 2017 Estimation of daily stream water temperatures with a Bayesian regression approach. *Hydrological Processes* **31**, 1719–1733. https://doi.org/10.1002/hyp.11139.

Souza, F. A. & Lall, U. 2003 Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: applications of a multivariate, semiparametric algorithm. *Water Resources Research* **39**. https://doi.org/10.1029/2002WR001373.

Stefan, H. G. & Preud'homme, E. B. 1993 Stream temperature estimation from air temperature. *JAWRA Journal of the American Water Resources Association* **29**, 27–45. https://doi.org/10.1111/j.1752-1688.1993.tb01502.x.

Stefan, H. G. & Sinokrot, B. A. 1993 Projected global climate change impact on water temperatures in five north central U. S. streams. *Climatic Change* **24**, 353–381. https://doi.org/10.1007/BF01091855.

Tang, Y., Reed, P., Wagener, T. & Van Werkhoven, K. 2006 Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrology and Earth System Sciences Discussions* **3**. https://doi.org/10.5194/hessd-3-3333-2006.

Tehrany, M. S., Pradhan, B. & Jebur, M. N. 2013 Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. *Journal of Hydrology* **504**, 69–79. https://doi.org/10.1016/j.jhydrol.2013.09.034.

Temizyurek, M. & Dadaser-Celik, F. 2018 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.

Thomann, R. V. & Mueller, J. A. 1987 *Principles of Surface Water Quality Modeling and Control*. Harper-Collins, New York, 644 p.

Toffolon, M. & Piccolroaz, S. 2015 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.

van Vliet, M., Yearsley, J., Franssen, W., Ludwig, F., Haddeland, I., Lettenmaier, D. & Kabat, P. 2012 Coupled daily streamflow and water temperature modeling in large river basins. *Hydrology and Earth System Sciences* **16**, 4303–4321. https://doi.org/10.5194/hess-16-4303-2012.

van Vliet, M. T. H., Franssen, W. H. P., Yearsley, J. R., Ludwig, F., Haddeland, I., Lettenmaier, D. P. & Kabat, P. 2013 Global river discharge and water temperature under climate change. *Global Environmental Change* **23**, 450–464. https://doi.org/10.1016/j.gloenvcha.2012.11.002.

van Werkhoven, K., Wagener, T., Reed, P. & Tang, Y. 2009 Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. *Advances in Water Resources* **32**, 1154–1169. https://doi.org/10.1016/j.advwatres.2009.03.002.

Vapnik, V., Golowich, S. E. & Smola, A. 1996 Support vector method for function approximation, regression estimation and signal processing. In: *Proceedings of the 9th*

International Conference on Neural Information Processing Systems. NIPS'96, pp. 281–287.

Varma, S. & Simon, R. 2006 Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91. https://doi.org/10.1186/1471-2105-7-91.

Wang, W., Xu, D., Chau, K. & Chen, S. 2013 Improved annual rainfall-runoff forecasting using PSO–SVM model based on EEMD. *Journal of Hydroinformatics* **15**, 1377–1390. https://doi.org/10.2166/hydro.2013.134.

Wang, X. & Babovic, V. 2016 Application of hybrid Kalman filter for improving water level forecast. *Journal of Hydroinformatics* **18**, 773–790. https://doi.org/10.2166/hydro.2016.085.

Wang, X., Zhang, J. & Babovic, V. 2016 Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. *Ecological Indicators* **66**, 428–439. https://doi.org/10.1016/j.ecolind.2016.02.016.

Wang, X., Babovic, V. & Li, X. 2017 Application of spatial-temporal error correction in updating hydrodynamic model. *Journal of Hydro-Environment Research* **16**, 45–57. https://doi.org/10.1016/j.jher.2017.07.001.

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

WMO 1992 *Simulated Real-Time Intercomparison of Hydrological Models*. 38-WMO No.779. WMO Operational Hydrology Report (OHR).

Yang, J. 2011 Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. *Environmental Modelling & Software* **26**, 444–457. https://doi.org/10.1016/j.envsoft.2010.10.007.

Yearsley, J. R. 2009 A semi-Lagrangian water temperature model for advection-dominated river systems. *Water Resources Research* **45**. https://doi.org/10.1029/2008WR007629.

Yuan, Y., Khare, Y., Wang, X., Parajuli, P. B., Kisekka, I. & Finsterle, S. 2015 Hydrologic and water quality models: sensitivity. *Transactions of the ASABE* **58**, 1721–1744. https://doi.org/10.13031/trans.58.10611.

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

Zhu, S., Nyarko, E. K., Hadzima-Nyarko, M., Heddam, S. & Wu, S. 2019a 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.

Zhu, S., Heddam, S., Wu, S., Dai, J. & Jia, B. 2019b 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.

Zhu, S., Bonacci, O., Oskoruš, D., Hadzima-Nyarko, M. & Wu, S. 2019c Long term variations of river temperature and the influence of air temperature and river discharge: case study of Kupa River watershed in Croatia. *Journal of Hydrology and Hydromechanics* **67**. https://doi.org/10.2478/johh-2019-0019.

Zhu, S., Heddam, S., Nyarko, E. K., Hadzima-Nyarko, M., Piccolroaz, S. & Wu, S. 2019d Modeling daily water temperature for rivers: comparison between adaptive neuro-fuzzy inference systems and artificial neural networks models. *Environmental Science and Pollution Research* **26**, 402–420. https://doi.org/10.1007/s11356-018-3650-2.

Zhu, S., Hadzima-Nyarko, M., Gao, A., Wang, F., Wu, J. & Wu, S. 2019e Two hybrid data-driven models for modeling water-air temperature relationship in rivers. *Environmental Science and Pollution Research* **26**, 12622–12630. https://doi.org/10.1007/s11356-019-04716-y.

First received 23 July 2020; accepted in revised form 24 March 2021. Available online 7 April 2021