Proc. 3rd International Conference on Advancements in Engineering Education (iCAEED2024)
20-23 Nov 2024, Sydney, Australia
ISBN: 978-1-7636843-3-1

Prediction of River Water Temperature using the Extreme Gradient Boosting – Tropical River System of India

Rajesh Maddu[1], Shaik Rehana[1*], Ataur Rahman[2], Taha B.M.J Ouarda[3],

[1]Hydroclimatic Research Group, Lab for Spatial Informatics, International Institute of Information Technology, Gachibowli, Hyderabad, Telangana, India
[2] School of Engineering, Design and Built Environment, Western Sydney University, Building XB, Second Avenue, Kingswood, NSW 2751, Australia
[3] National Institute of Scientific Research, Quebec, Canada
** Corresponding author E-mail: rehana.s@iiit.ac.in*

Abstract

The physical, biological, and chemical properties of a river are directly influenced by its river water temperature (RWT), which also controls the survival and fitness of all aquatic organisms. Machine Learning (ML) gained popularity because of its ability to model complex and nonlinearities between RWT and its predictors compared to process-based models that require large data. The present study demonstrates a new ML approach, Extreme Gradient Boosting (XGBoost), to predict accurate RWT estimates with the most appropriate form of AT. Further, the proposed XGBoost results are compared with the Support Vector Regressor (SVR) model. The proposed modelling framework's effectiveness is demonstrated with a tropical river system of India, Tunga-Bhadra River, as a case study. Results indicate that the XGBoost results are better than SVR for RWT prediction. The study demonstrates how ML methods can be used to generate accurate RWT predictions in river water quality modelling.

***Keywords*:** Air Temperature, Machine Learning, River Water Temperature, SVR, XGBoost

1. INTRODUCTION

Rivers and their ecosystems rely on water quality, health and proper functioning. Predicting river water quality (RWQ) variables has become essential for various environmental, hydrological, and ecological applications (Zhu & Piotrowski, 2020). Altered temperature, precipitation, and runoff patterns can impact key RWQ indicators such as River Water Temperature (RWT), Dissolved Oxygen (DO), pH, nutrient balance, and contaminant presence (Sinokrot & Stefan, 1993). For example, the rate of chemical reactions typically increases at higher RWT under elevated air temperatures (Rajesh & Rehana, 2022). The rising RWT can deplete oxygen levels, potentially disrupting aquatic habitats, harming biodiversity, and leading to the decline of sensitive species. Additionally, RWT is a key indicator in RWQ and aquatic life, influencing DO levels, algal growth, and overall aquatic production (Feigl et al., 2021). Predicting RWT is essential for safeguarding river ecosystems, public health, scientific advancement, and promoting sustainable progress (Chapra, 1998).

Predicting RWT has traditionally relied on physically based models like Delft3D, Soil and Water Assessment Tool (SWAT) and QUAL2K (Wang et al., 2022). Recently, models like Air2Stream (Shrestha & Pesklevits, 2022), Temperature Duration Curve (TDC) (Ouarda et al., 2022), and other process-based models have also been utilized (Wang et al., 2022). These models require detailed site-specific data, including solar radiation, streamflow, etc. (Feigl et al., 2021). While physical models offer accurate results, they rely heavily on such detailed data. In contrast, statistical models require fewer input variables, making them suitable for ungauged river systems, but they struggle to describe nonlinear characteristics accurately (Wang et al., 2022). To address these limitations, Machine Learning (ML) techniques offer a promising

alternative by effectively handling nonlinear relationships and requiring minimal data inputs. Data-driven algorithms, like ML models with minimal data inputs (such as AT), can effectively address data sparsity issues in simulating RWT.

In recent years, Artificial Neural Networks (ANN) (Qiu et al., 2020), Random Forest (RF) models (Rajesh & Rehana, 2021), K-nearest neighbors (KNN) approach (Gavahi et al., 2019), Support vector regression (SVR) (Rehana, 2019) have garnered significant attention for RWT prediction. To this end, numerous studies have utilized ML models for predicting RWT. However, selecting the appropriate ML model is important, as the results are heavily influenced by the specific model during modelling.

2. STUDY AREA AND DATA

The RWT modelling was conducted for the Tunga River at Shimoga, which merges with the Bhadra River to form the Tunga-Bhadra River, a major tributary of the Krishna River basin in India (Figure 1). Observed mean air (water) temperatures was 24.78°C (27.54°C), with standard deviations of 2.77°C (2.66°C), respectively. Data from January 1, 1989, to January 1, 2004, was obtained from the Central Water Commission (CWC) and Advanced Centre for Integrated Water Resources Management (ACIWRM). The Tunga-Bhadra River is significantly impacted by climate change, experiencing increased RWTs and discharges, and is one of India's most polluted rivers due to municipal and industrial effluents (Rehana & Mujumdar, 2012) (CPCB, 2019).

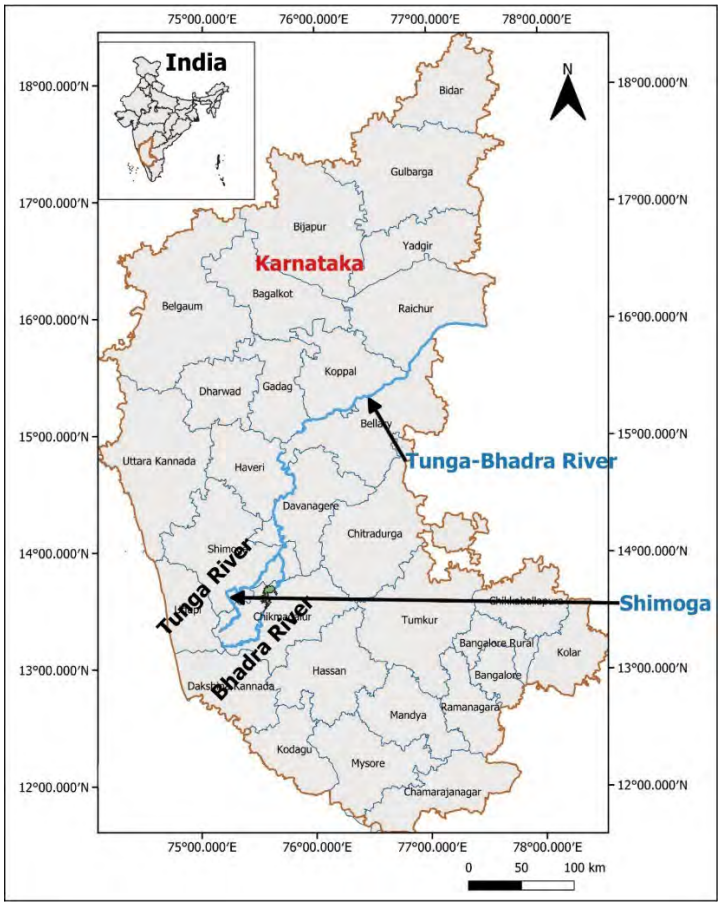

Figure 1. Location map of Tunga-Bhadra River and Shimoga station, India.

3. METHODOLOGY

The proposed model includes data pre-processing, identifying the model parameters, and modelling with performance measures at daily and monthly timescale (Figure 2). Particularly, this work explores XGBoost technique for prediction of RWT. Five statistical measures, including the Nash-Sutcliffe efficiency (NSE), Kling–Gupta efficiency (KGE), RMSE-observations standard deviation ratio (RSR), the root mean squared error (RMSE) are considered to measure the performances of ML model. In Rajesh & Rehana (2021), these metrics are explained in detail.

3.1. Extreme Gradient Boosting Regressor (XGBoost)

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting that optimizes the performance of the model and has gained immense popularity in recent years. XGBoost is an ensemble learning method that leverages second-order Taylor expansion for loss function approximation to optimize the tree structure and reduce overfitting (Chen & Guestrin, 2016). The XGBoost algorithm iteratively improves the model by computing gradients of the loss function with respect to current predictions, fitting a base learner to these gradients, and updating the model with an optimal step length while applying regularization. It incorporates various regularization techniques, i.e., L1 (regularization adds a penalty equal to the absolute value of the magnitude of coefficients), and L2 (regularization adds a penalty equal to the square of the magnitude of coefficients) to improve performance. The model parameters, inputs, and algorithm are described as follows.

Parameters: G = Model, z_i = Input set, t_i = Output set of $G(z_i)$, J = Loss function, f_m = Base learner function at stage m, δ_m = Step length, s_{im} = Pseudo residual, k = number of samples, μ= Regularization parameter, ϑ= Minimum loss reduction to make a further partition, ρ = L1 regularization term on weights, σ = L2 regularization term on weights.

Inputs: Inputs to the XGBoost are the number of iterations (P), differentiable loss function $J(t, G(z))$, training dataset for $\{(z_i, t_i)\}$ for $i=1$ to k, and hyperparameters $\mu, \vartheta, \rho, \sigma$.

Algorithm (Chen & Guestrin, 2016):

1. Initialize the model with a constant prediction:
   $$G_0(z) = arg\ min_t \sum_{i=1}^{k} J(t_i, \mu) \quad (1)$$
2. For $m = 1$ to P, repeat:
   a. Compute the gradients (pseudo-residuals (s_{im})):
      $$s_{im} = -\left[\frac{\partial J(t_i, G(z_i))}{\partial G(z_i)}\right] \quad \text{for } i = 1,\ldots,k \quad (2)$$
      where $G(z) = G_{m-1}(z)$
   b. Fit a base learner $f_m(z)$ is pseudo residuals $\{(z_i, s_{im})\}$ for i=1 to k:
      $$f_m(z) \leftarrow \text{train } (z_i, s_{im}) \quad (3)$$
   c. Compute the optimal multiplier δ_m:
      $$\delta_m = arg\ min_\sigma \sum_{i=1}^{k} J(t_i, G_{m-1}(z_i) + \delta f_m(z_i)) \quad (4)$$
   d. Update the model:
      $$G_m(z) = G_{m-1}(z) + \delta_m f_m(z) \quad (5)$$
   e. Apply regularization:
      - L1 Regularization (Lasso):
        $$Penalty = \rho \sum_j |w_j| \quad (6)$$
      - L2 Regularization (Ridge):
        $$Penalty = \sigma \sum_j w_j^2 \quad (7)$$
   f. Pruning (if the loss reduction is less than ϑ, stop splitting):
      If Loss Reduction < ϑ the stop splitting
3. Final model output: $G_p(z)$

3.2. Support Vector Regressor (SVR)

The Support Vector Machine (SVM) is a kernel function learning machine that adheres to the structural risk minimization principle (Vapnik et al., 1996). In addition to classification tasks, Support Vector Regression (SVR) is an extension of SVM for regression problems. SVR is particularly effective for capturing complex, non-linear relationships in hydrological data. The SVR model is a strong choice for smaller datasets and is less sensitive to outliers in the data. The SVR is generally less computationally expensive than the ML models. Detailed information of the SVR algorithm for the prediction of RWT can be found in Rajesh & Rehana (2021).

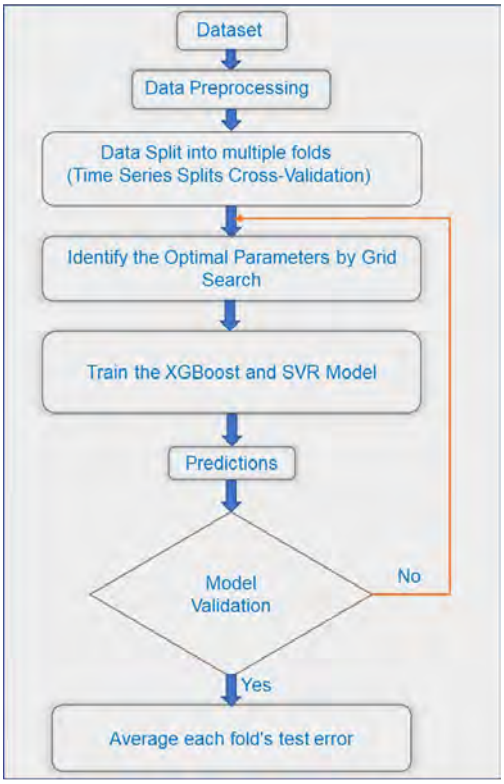

Figure 2. Flow diagram for XGBoost and SVR regression model.

4. RESULTS AND DISCUSSION

The dataset used in this study contains daily minimum, maximum, and mean ATs, and the RWTs for the period from 1st January 1989 to 1st January 2004. The variability of ATs and RWT changes are visualized in Figure 3. To accurately predict RWT, the next step is to employ an appropriate ML model that ensures precise calibration and validation. In this study we have proposed the XGBoost model to predict the RWT and further compared the results with SVR model. First, we used the time-series splits cross-validation technique which trains on an initial small subset of data, estimates the subsequent data points, and adding them to the training set for the next iteration to provide an almost unbiased error estimate. We then used the GridSearchCV (Pedregosa *et al.* 2011) to evaluate all parameter combinations and selected the best one to optimize performance. Finally, we evaluated the ML models against acceptable performance measures, as shown in Figure 2.

The performance of the XGBoost, and SVR models for daily, and monthly data at Shimoga station are provided in Table 1. From Table 1, XGBoost (R^2 = 0.91, MSE = 0.62, RMSE = 0.78, MAE = 0.55, and

RSR = 0.30) model has performed slightly better than SVR (R^2 = 0.89, MSE = 0.69, RMSE = 0.83, MAE = 0.60, and RSR = 0.32) for daily time scale. From Figure 4, ML results showed that the seasonal patterns of predicted RWT are almost synchronous and comparable with the observed values. Compared to the two ML models, XGBoost (R^2 = 0.96, MSE = 0.21, RMSE = 0.46, MAE = 0.19, and RSR = 0.19) model has performed slightly better than SVR (R^2 = 0.93, MSE = 0.38, RMSE = 0.61, MAE = 0.44, and RSR = 0.26) for monthly time scale. The present study confirms the superiority of ML models in predicting RWT, aligning with earlier research findings based on Rehana (2019) and Rajesh & Rehana (2021) for the same case study. It can be noted that performance coefficients improved at the monthly time scale, showing higher R^2 and NSE, and lower RMSE and MAE values compared to the daily time scale. It can be noted that the improved ML model accuracy with monthly data, showing higher R^2 and NSE, and lower RMSE and MAE values compared to daily data due to taking the daily values into monthly totals (i.e., averaging all the daily values, the errors will be distributed and get better results). Overall, this case study demonstrates the potential of integrating scientific knowledge with ML tools to enhance RWT predictions.

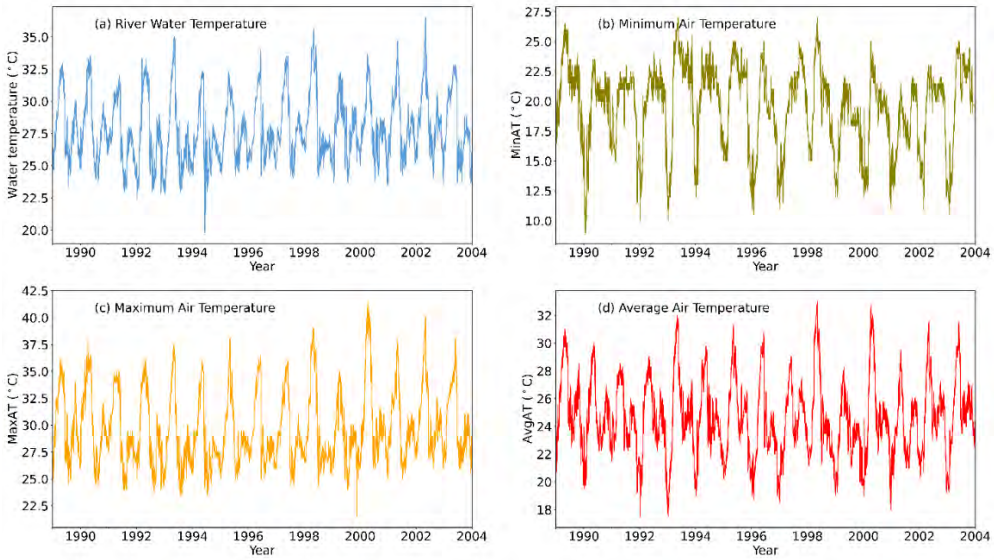

Figure 3. Time series of daily river water temperature, minimum air temperature, maximum air temperature, and average air temperature for the period 1989-2004.

Table 1. Performance of ML model in the prediction of RWT.

Data	Model	R^2	MSE	RMSE	MAE	RSR
Daily	SVR	0.89	0.69	0.83	0.60	0.32
	XGBoost	0.91	0.62	0.78	0.55	0.30
Monthly	SVR	0.93	0.38	0.61	0.44	0.26
	XGBoost	0.96	0.21	0.46	0.19	0.19

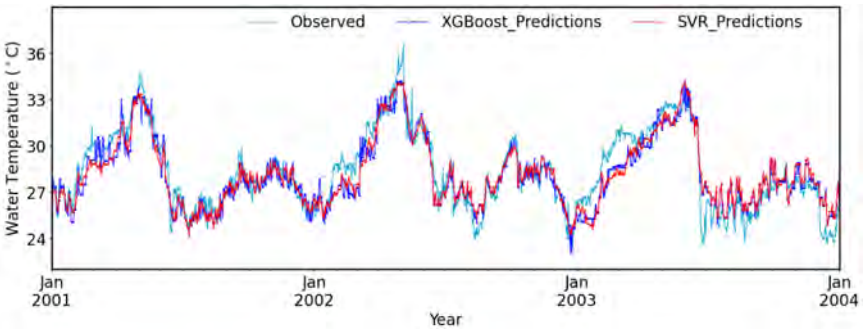

Figure 4: Comparison of time series results of observed, XGBoost and SVR predictions for Tunga-Bhadra River, India.

5. CONCLUSIONS

This study aims to predict the RWT for the Tunga-Bhadra River basin using minimum, maximum, and average ATs with a ML model. The XGBoost model demonstrated superior performance on daily data and monthly data. The data-driven approach successfully predicted RWT, but the study has limitations, including reliance on data from 1989 to 2004, which is the only extensive period with sufficient data. The proposed RWT modelling framework can be updated with new data and extended to additional stations. Further research is needed on robust and hybrid approaches incorporating streamflow, solar radiation, and other factors.

ACKNOWLEDGEMENT

The authors thank the funding agencies: Ministry of Science & Technology, Department of Science and Technology (DST), Government of India and India-Canada Centre for Innovative Multidisciplinary Partnership to Accelerate Community Transformation (IC-IMPACTS) Canada for their funding support (Grant No: DST/IC/IC-IMPACTS/2022/P-9); Scheme for Promotion of Academic and Research Collaboration (SPARC) Program funded by Ministry of Human Resource Development, Government of India with Project ID-P2488.

REFERENCES

Chapra, S. C. (1998). Surface Water Quality Modelling. *McGraw Hill Kogakusha Ltd. New York.*

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

CPCB. (2019). Guidelines for water quality management. *Delhi, India*. https://cpcb.nic.in/wqm/Designated_Best_Use_Water_Quality_Criteria.pdf

Feigl, M., Lebiedzinski, K., Herrnegger, M., & Schulz, K. (2021). Machine learning methods for stream water temperature prediction. *Hydrology and Earth System Sciences Discussions*, 1–35. https://doi.org/10.5194/hess-2020-670

Gavahi, K., Mousavi, S. J., & Ponnambalam, K. (2019). Adaptive forecast-based real-time optimal reservoir operations: Application to Lake Urmia. *Journal of Hydroinformatics*, *21*(5), 908–924. https://doi.org/10.2166/hydro.2019.005

Ouarda, T. B. M. J., Charron, C., & St-Hilaire, A. (2022). Regional estimation of river water temperature at ungauged locations. *Journal of Hydrology X*, *17*, 100133. https://doi.org/10.1016/j.hydroa.2022.100133

Pedregosa, F., Varoquaux, G., et al. 2011. Scikit-learn: ML in Python. *Journal of ML Research*, 12.

Qiu, R., Wang, Y., Wang, D., Qiu, W., Wu, J., & Tao, Y. (2020). Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River. *Science of The Total Environment*, *737*, 139729. https://doi.org/10.1016/j.scitotenv.2020.139729

Rajesh, M., & Rehana, S. (2021). Prediction of river water temperature using machine learning algorithms: A tropical river system of India. *Journal of Hydroinformatics*, *jh2021121*. https://doi.org/10.2166/hydro.2021.121

Rajesh, M., & Rehana, S. (2022). Impact of climate change on river water temperature and dissolved oxygen: Indian riverine thermal regimes. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-12996-7

Rehana, S. (2019). River Water Temperature Modelling Under Climate Change Using Support Vector Regression. In S. K. Singh & C. T. Dhanya (Eds.), *Hydrology in a Changing World: Challenges in Modeling* (pp. 171–183). Springer International Publishing. https://doi.org/10.1007/978-3-030-02197-9_8

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

Shrestha, R. R., & Pesklevits, J. C. (2022). Modelling spatial and temporal variability of water temperature across six rivers in Western Canada. *River Research and Applications*. https://doi.org/10.1002/rra.4072

Sinokrot, B. A., & Stefan, H. G. (1993). Stream temperature dynamics: Measurements and modeling. *Water Resources Research*, *29*(7), 2299–2312. https://doi.org/10.1029/93WR00540

Vapnik, V., Golowich, S. E., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Proceedings of the 9th International Conference on Neural Information Processing Systems*, 281–287.

Wang, L., Xu, B., Zhang, C., Fu, G., Chen, X., Zheng, Y., & Zhang, J. (2022). Surface water temperature prediction in large-deep reservoirs using a long short-term memory model. *Ecological Indicators*, *134*, 108491. https://doi.org/10.1016/j.ecolind.2021.108491

Zhu, S., & Piotrowski, A. P. (2020). River/stream water temperature forecasting using artificial intelligence models: A systematic review. *Acta Geophysica*, *68*(5), 1433–1442. https://doi.org/10.1007/s11600-020-00480-7