A Hybrid Machine Learning Framework for River Water Quality Prediction under Data Uncertainties

Thesis submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Civil Engineering

by

RAJESH MADDU
(2019900041)
rajesh.maddu@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad – 500 032, INDIA
April 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**A Hybrid Machine Learning Framework for River Water Quality Prediction under Data Uncertainties**" by **Rajesh Maddu**, has been carried out under my supervision and has not been submitted elsewhere for a degree.

Dr. Shaik Rehana
Associate Professor
Lab for Spatial Informatics
International Institute of Information Technology
Hyderabad, Telangana, India – 500032

Date: .05.2024

Dedicated to

My Family Members

Acknowledgements

I would like to sincerely thank my esteemed guide, Dr. Shaik Rehana Madam, for her invaluable supervision, support, and guidance during my Ph.D. Her genius, encouragement, and technical discussions helped a lot to improve myself in the research thought process and driven me towards excellence. Being her Ph.D. student has been a privilege.

I would like to thank Prof. Pradeep Kumar Ramancharla, IIITH, Director, CSIR-Central Building Research Institute (CBRI) for his invaluable suggestions, encouragement, and continuous support throughout the course of PhD.

I would like to express my sincere gratitude to Prof. Balaji Rajagopalan (University of Colorado, USA) for his expertise, insightful comments, and suggestions as an examiner and as the Associate Editor for the Water Resources Research publication.

I would like to thank Dr. Shailesh Kumar Singh (National Institute of Water & Atmospheric Research Ltd (NIWA), New Zealand) for the discussions and evaluation of the PhD Thesis. I would like to thank Dr. Sagar Rohidas Chavan, Head of Civil Engineering Dept. (IIT Ropar) for the valuable suggestions.

I would like to thank Prof. K. S. Rajan sir, Head, Lab of Spatial Informatics, for his inputs regarding my research.

I am extremely thankful to my wonderful Hexagon managers and colleagues, Chandrasekharam Somanchi, Shiva Kumar Krishnamurti, Prashanti Seri, Rajya Lakshmi Vulava, Suresh Samudrala, Siddartha Thota, Arun Akella, Pramod Reddy Valipireddy for their motivation to join in Ph.D. and supported continually.

I would like to thank my family for their unconditional love and support. I would not have achieved my Ph.D. degree if not for their support. They have been my greatest strength during my highs and lows. For my wonderful sons Tanish Reddy and Aarush Reddy, for their affection. I am extremely grateful to my wife, Mrs. Sandhyarani, for her love, affection, and support during my Ph.D. tenure.

I would also like to extend my gratitude to Satish Kumar Mummidivarapu and Avantika Latwal (research scholars in LSI lab, IIITH) for their invaluable support and for making my IIIT-Hyderabad days joyful and memorable. I thank Sri Y Kishore, Smt. Pushpalatha, thank you for your valuable support throughout the course of PhD.

(Rajesh Maddu)

ABSTRACT

The impact of climate change on water quality variables is an essential topic for sustainable river water quality management in a warming environment and is a great environmental concern worldwide. River Water Quality (RWQ) models aim to simulate the behavior of various water quality variables in response to pollutants, land use changes, and climate change. However, these water quality models suffer from sparse data leading to data uncertainty. In the past decades, different models have been successfully used for RWQ modeling under different spatial and temporal scales. To simulate RWQ variables, physically based water quality models can be used, but they require large amounts of site-specific detailed data, including stream geometry, meteorological variables, and hydraulic properties of the river, which are unavailable for many river systems globally. However, unlike process-based models, statistical models possess many advantages. Additionally, statistical models do not require a large number of input variables, which are unavailable for many ungauged river systems. However, accurately describing the nonlinear characteristics of a data series is a significant shortcoming of this approach. To overcome such limitations, artificial intelligence algorithms, i.e., Machine Learning (ML) techniques, are widely used to address a range of nonlinear prediction problems. Such models are suited for information extraction from sequential data in RWQ modeling, and they serve functionalities to build models using a reduced number of variables with more accurate simulation.

Machine Learning (ML) has been increasingly adopted due to its ability to model complex and nonlinearities between river water quality (RWQ) variables and their predictors (e.g., Air Temperature, AT, streamflow). To simulate RWQ parameters using data-driven algorithms, more input variables are required, which are unavailable for many ungauged river systems. Climatic variables that are readily available are the maximum, minimum, and average AT to build RWQ models with more accurate simulation and higher computational efficiency. In this context, most of these ML approaches have been applied without any detailed sensitivity analysis to identify the most influencing variables to be considered in the prediction of RWQ variables. Furthermore, the development of systematic models combined with ML under minimum data input variables has not been intensively studied in predicting RWQ variables. To address these, the present study first 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 RWQ variables estimates. Air Temperature (AT) changes can affect River Water Temperature (RWT) under anthropogenic climate change, the primary variable that influences water quality. Therefore, the present study selected RWT as a water quality variable prediction with a tropical river system of India, Tunga-Bhadra River, as a case study. Further, the proposed ML approaches have been combined with the Ensemble Kalman Filter (EnKF) data assimilation (DA) technique to improve the predicted values based on the measured data. Overall, the study concluded that the SVR has been noted as the most robust ML model when coupled with a global sensitivity algorithm and DA techniques to predict RWT at a monthly time scale compared to daily and seasonal. Also, the study concluded that the SVR model is a strong choice for smaller datasets and is less sensitive to outliers in the data compared to some other models. The SVR is generally less computationally expensive than the ML models.

Another data uncertainty is the lack of availability of long-time series data to capture interannual variability and consistent water quality measurement datasets in RWQ modeling. Generally, RWQ data availability is on a monthly scale and is burdened with a large number of missing values with limited durations. In this context, the selection of appropriate model inputs, development of models under limited data, processing of non-stationary data, seasonality scenarios, and different potentially influenced relevant lags of variables have not been intensively investigated in the literature, especially in the case of estimation of RWQ variables. Given the missing, limited, and non-stationary data scenarios, the present thesis developed hybrid models for RWQ variables prediction using Long Short-Term Memory (LSTM), integrated with (i) k-nearest neighbor (k-NN) bootstrap resampling algorithms (kNN-LSTM) to address the data-limitations and (ii) discrete wavelet transform (WT) approach (WT-LSTM) to address the time-frequency localized features. To demonstrate the prediction of RWQ variables and to assess the impact of climate change on the river water quality parameters, this study considered the two most important water quality variables, i.e., River Water Temperature (RWT) and saturated Dissolved Oxygen (DO) concentrations, and AT and lag variables as predictors. When WT and k-NN bootstrap resampling algorithms were included, LSTM outperformed the conventional models; hence these hybrid models are the new promising frameworks for RWQ prediction under data-sparse regions. Bayesian

optimization is applied to optimize the hyperparameters of all applied ML models. The hybrid kNN-LSTM has effectively predicted RWT for five catchment sites (i.e., Narmada, Cauvery, Musi, Godavari, and Ganga) out of seven catchment sites (i.e., Narmada, Cauvery, Sabarmati, Tunga-Bhadra, Musi, Godavari, and Ganga) at monthly time scales under data limitations and outperformed the standalone LSTM, WT-LSTM, and hybrid 3-parameter version of Air2Stream models (physical based RWT prediction model). Also, this thesis presents the combined effects of streamflow and AT in the prediction of RWT using the kNN-LSTM model, LSTM model, a modified nonlinear regression model, and an 8-parameter version of Air2Stream when applied to three major river systems of India (Tunga-Bhadra, Musi, and Ganga). Results revealed that the kNN-LSTM model could predict RWT more accurately than the LSTM model, a modified nonlinear regression model, and an 8-parameter version of the Air2Stream model for all three catchment sites. Overall, the study concluded that hybrid models consistently outperformed standalone models in addressing uncertainty due to data sparsity.

The study assessed the climate change impacts on river water quality variables using an Ensemble of National Aeronautics Space Administration (NASA) Earth Exchange Global Daily Downscaled Projections (NEX-GDDP) with Representative Concentration Pathways (RCP) scenarios 4.5 and 8.5 for seven major polluted river catchments of India. For this assessment, the best performance hybrid kNN-LSTM model has been used for future predictions. The RWT increase for Tunga-Bhadra, Musi, Ganga, and Narmada basins are predicted as 3.0, 4.0, 4.6, and 4.7 ^{o}C, respectively for 2071-2100. Overall, RWT over Indian catchments is likely to rise by more than 3.0 °C for 2071-2100.

While river water temperatures (RWTs) are increasing under climate change signals, how climate change affects DO saturation levels in response to RWT has not been intensively studied. This thesis examined the direct effect of rising RWTs on saturated DO concentrations for seven major polluted river catchments of India at a monthly scale. The RWT reaches close to 35 ^{o}C, and decreases DO saturation capacity by 2%–12% for 2071–2100. Also, in this thesis evaluated the effect of climate change on DO saturation levels with respect to RWT and streamflow using the kNN-LSTM model forced with nine hypothetical climate change scenarios for three polluted catchments of India (Tunga-Bhadra, Musi, and Ganga). The largest DO decreases (13.22 %) were found in the Ganga catchment for selected

climate change scenarios relative to the historical values. Overall, for every 1 ^{o}C RWT increase, there will be about 2.3 % decrease in DO saturation level concentrations over Indian catchments under climate signals.

Overall, the study demonstrates how hybrid ML methods can be coupled with a global sensitivity algorithm, DA techniques, bootstrapping algorithms, and wavelets to generate accurate RWQ variables prediction under data uncertainties. Although the focus of our study has been limited to climate change impacts on RWT and DO saturations, the proposed hybrid ML modeling frameworks are generic and have the potential to incorporate other water quality parameters as well to make better decisions towards river water quality management.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
ARIMA	Autoregressive Integrated Moving Average Model
AT	Air Temperature
BO	Bayesian Optimization
BOD	Biochemical Oxygen Demand
CPCB	Central Pollution Control Board
CWC	Central Water Commission
DA	Data Assimilation
DL	Deep Learning
DO	Dissolved Oxygen
DT	Decision Tree
EnKF	Ensemble Kalman Filter
GA	Genetic Algorithm
GCM	General Circulation Model
GPR	Gaussian Process Regression
GSA	Global Sensitivity Analysis
KGE	Kling–Gupta Efficiency
KJ	Kilojoule
KNN	K-nearest neighbors
k-NN	k-Nearest Neighbor
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NEERI	National Environmental Engineering Research Institute
NEX-GDDP	NASA Earth Exchange Global Daily Downscaled Projections
NSE	Nash-Sutcliffe Efficiency
Q	Discharge

R^2	Coefficient of Determination
RCP	Representative Concentration Pathway
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RR	Ridge regression
RSR	RMSE observations Standard deviation Ratio
RWQ	River Water Quality
RWT	River Water Temperature
SVR	Support Vector Regression
SWAT	Soil and Water Assessment Tool
T_w	River Water Temperature
T_a	Air Temperature
UN	United Nations
WT	Wavelet Transformation

Chapter 1
INTRODUCTION

1.1. General

The surface of our planet contains freshwater bodies such as rivers, lakes, and ponds that are suitable for human consumption (Dugan 1972). The development of major human civilizations was largely dependent on freshwater sources like rivers and lakes (Lundqvist 2009). It may be noted that man's early habitation and civilization sprang up along the banks of rivers (Dugan 1972; Webb 1992). For several centuries, these water bodies have been supplying fresh water to several household, agricultural, and industrial purposes (Paily et al. 1974; Mohseni et al. 1998; van Vliet et al. 2011). So, the survival of human society is largely dependent on the availability of clean freshwater resources like rivers and lakes, and freshwater resources have, therefore, become increasingly crucial to our society (Penchev 1972; Tennant 1976; Shiklomanov 1993). River water quality is a great environmental concern worldwide, and the evaluation of water quality is fundamental to the study and use of water for various purposes such as drinking water supply, Irrigation, ecosystem health, Industrial purposes, etc. (Chapra 1998).

Water quality models are important tools in analyzing the spatiotemporal extent of pollutants and identifying the state of the environment (Thomann and Mueller 1987; van Vliet et al. 2021). River Water Quality (RWQ) models aim to simulate the behavior of various water quality variables in response to pollutants, land use changes, and climate change (Edinger et al. 1968). The input data and information that would be needed for water quality modeling to simulate RWQ variables include initial and boundary concentrations, source of pollutants, baseline conditions, flow characteristics, meteorological and flow data, and the geometry of the modeled waterbodies. In the past decades, different models have been successfully used for RWQ modeling under different spatial and temporal scales (Mohseni et al. 1998; van Vliet et al. 2013; Piccolroaz et al. 2016; Feigl et al. 2021). In general, the model selection depends not only on the study requirements, namely the output timescale but also on the availability and quality of the input data (Almeida and Coelho 2022). In this context, process-based models are strongly rooted in scientific theory (Hilborn

and Mangel 1997), which represent physical processes controlling RWQ variables (Feigl et al. 2021), and their implementation based on the heat exchange dynamics between the water body and the surrounding environments (Sinokrot and Stefan 1993; Du et al. 2018). To simulate RWQ variables, physically based water quality models (e.g., Delft3D model, Soil and Water Assessment Tool (SWAT) model, QUAL2K, dynamical surface water quality model (DynQual), etc.), statistical models (e.g., Air2Stream model) were most widely used (Piccolroaz et al. 2016; van Vliet et al. 2021; Shrestha and Pesklevits 2022; Wang et al. 2022). As opposed to statistical models relying heavily on observed water quality data, physical models simulate the emission and transport of pollutant loadings along the river network based on site-specific detailed data, including stream geometry, meteorological and hydraulic properties of the river (Piccolroaz et al. 2016). However, these water quality models suffer from sparse data leading to data uncertainty (Pohle et al. 2019). In many regions of the world the monitoring and the information collected on water quality variables are limited, primarily due to the shortage of financial resources (Tabari and Hosseinzadeh Talaee 2015; Jackson et al. 2017). Furthermore, most of the river systems are burdened with data limitations and form a significant challenge for implementing process-based RWQ models at various spatial and temporal scales (Read et al. 2019; Pohle et al. 2019). Although process-based models could give very accurate results, they require large amounts of site-specific detailed data, including stream geometry, a complete set of meteorological variables, and the hydraulic properties of the river, that are unavailable for many river systems, along with in-depth knowledge of the field (Piccolroaz et al. 2016). However, given a large amount of data and complex algorithms, process-based models are always time-consuming (Jackson et al. 2018; Piotrowski and Napiorkowski 2019). According to Dugdale et al. (2017), these models first calculate energy fluxes to or from the river and then determine the changes in RWQ variables. Furthermore, these models sometimes have complex practical implementation issues in the large spatial domain of interest (Feigl et al., 2021; Rehana, 2019). Another data uncertainty is the lack of availability of long-time series data to capture interannual variability and consistent water quality measurement datasets in RWQ modeling. Consequently, measurement errors or discontinuity might lead to noisy data (Graf et al. 2019; Virro et al. 2021). The loss of data can have different causes, both instrumental and administrative. Sensors collecting data in situ are susceptible to technical errors, failing to

record observations and/or anomalies, and, therefore, data gaps (Kermorvant et al. 2021). Sparse and often irregular sampling of RWQ data makes the statistical analyses (such as trend analysis) problematic (Hirsch et al. 2015; Shrestha et al. 2019). Incorporating erroneous values into RWQ models could result in wrong conclusions that might be costly to the environment or humans (Rangeti et al. 2015). The lack of water quality data in rivers can make it difficult to manage river ecosystems, particularly in reference to biodiversity, oxygen conditions (and consequently for their self-cleaning capacity), and finally, their quality (Sojka and Ptak 2022).

In contrast to process-based models, statistical models cannot inform about energy transfer mechanisms within a river (Dugdale et al. 2017). However, unlike process-based models, statistical models possess many advantages, including simplification of the complex relations between water quality indicators and the identification of similar temporal and spatial characteristics patterns among water quality variables. Additionally, statistical models do not require a large number of input variables, which are unavailable for many ungauged river systems. However, accurately describing the nonlinear characteristics of a data series is a significant shortcoming of the approach because the statistical models are usually based on temporal linear correlations within the modeled dataset (Wang et al. 2020). To overcome this shortage, artificial intelligence algorithms, i.e., Machine Learning (ML) techniques are widely used to address a range of nonlinear prediction problems. Especially these ML techniques are suited for information extraction from sequential data in RWQ modeling, and they serves functionalities to build models using reduced number of variables with more accurate simulation, higher computational efficiency, greater representativeness of the indices, and reduced requirements for human involvement and expertise (Shen 2018; Qiu et al. 2020; Zhu and Piotrowski 2020). These advances contribute to improving the prediction of water quality variables, thus seeking to obtain one that represents the various phenomena that occur in river water basins more rapidly and coherently with the reality and social context of water resources.

Recently, to simulate RWQ variables, the Air2Stream model (Shrestha and Pesklevits 2022; Zhu et al. 2022), Temperature Duration Curve (TDC) (Ouarda et al. 2022), Process-based models (e.g., Delft3D model, Soil and Water Assessment Tool (SWAT) model, etc.), have been used (Wang et al., 2022). However, such approaches require large amounts of site-

specific detailed data at daily time scales, including stream geometry and meteorological and hydraulic properties of the river (Piccolroaz et al. 2016). However, globally, RWQ data availability is on a monthly scale and is burdened with a large number of missing values with limited durations. The use of data-driven algorithms, such as ML models using minimum data inputs (such as AT), can be robust in addressing data sparsity in simulating RWQ models. Therefore, this study tries to address the data sparsity and uncertainties using data-driven algorithms to predict RWQ variables.

1.2. Climate Change Impact on River Water Quality

Climate change will affect river hydrologic and thermal regimes, directly impacting freshwater ecosystems and human water use (van Vliet et al. 2013). Increased evaporation, resulting from higher temperatures, together with regional changes in precipitation characteristics, has the potential to affect the runoff, frequency, and intensity of floods and droughts, soil moisture, water quality, and water demands (Intergovernmental Panel on Climate Change (IPCC) 2001; Cunderlik and Simonovic 2005). Many studies (Rajagopalan et al., 2019; Yang et al., 2017) have investigated the impacts of regional and global climate change and variabilities on river water quantity, but relatively very less attention has been given to river water quality. River water quality parameters are directly affected by changes in climate under anthropogenic greenhouse gases in the atmosphere, which in turn increases the risk of deterioration of the river ecosystem in terms of decreased Dissolved Oxygen (DO) levels under the decrease of stream flows and increase in River Water Temperature (RWT) (Rehana and Dhanya 2018). Climate change has adversely impacted RWQ under intensification and alterations in various hydro-climatic variables (e.g., AT, precipitation, streamflow, etc.) (Stefan & Sinokrot, 1993; van Vliet et al., 2013; Webb et al., 2003). RWT is the critical RWQ variable, which is intensified under various climatological variables such as AT under anthropogenic climate change effects (Mohseni et al. 1999; van Vliet et al. 2011) and due to net changes of heat flux at the air-water interface. The pronounced increase in RWT drives the rates of biological and chemical processes, affecting the reaction kinetics. Intensification of RWT will have adverse impacts in terms of a decrease in river DO saturation levels, directly affecting the river system self-purification capacity. Another dominant factor affecting the RWQ is streamflow, which defines the pollutant transport and dilution of nutrients and pollutant loads. Changes in streamflow directly affect the dilution

capacity of a river receiving pollutants. Furthermore, stream flow, which is a smoothened force caused by rainfall, also play a vital role in RWQ (van Vliet et al. 2011; Zhu et al. 2019f; Feigl et al. 2021), especially in Indian rivers impacted by low flows during summer seasons, leading to water quality deterioration. Therefore, the major RWQ parameters are RWT, DO, and turbidity, which are directly affected under climate changes, which further increases the risk of deterioration of the river ecosystem (Webb and Walling 1993; Mohseni et al. 1998; van Vliet et al. 2011). Higher RWTs and changes in extremes, including floods and droughts in the future, are projected to influence water quality and intensify many forms of water pollution (Bates et al. 2008).

In this context, assessment of climate change impacts on river water quality is vital in terms of assessing the variability of various RWQ parameters such as RWT, DO, Turbidity, etc. (Rehana & Mujumdar, 2011; van Vliet et al., 2011). To this end, various methods for estimating the impacts of climate change on hydrological behavior, as implemented in a number of earlier studies, are (1) using high resolution Regional Climate Models (RCMs) (e.g., Malmaeus et al. (2006)); (2) using General Circulation Models (GCMs) through statistical downscaling techniques (e.g., Wilby & Wigley (1997)); and (3) using hypothetical scenarios as input to hydrologic models (e.g., Jiang et al. (2007)) and river water quality models (Rehana & Mujumdar, 2014).

The preferred climate scenarios are usually those derived from the GCMs, which consider the natural and anthropogenic greenhouse gas emissions in the atmosphere. GCMs are developed to evaluate the plausible responses of the climate system to the changes in the behavior of natural and human systems, either separately or together (Houghton et al. 1996). GCMs can simulate Earth's climate with different climate variables, initial and boundary conditions, and structure. GCMs are increasingly being employed to solve or assess regional/local issues. GCMs are formulated on the principles of movement of energy, the momentum of a particle, and conservation of mass (Wilby et al. 2009). Forecasting future climate projections will be helpful for efficient planning in order to mitigate and adapt to changing climate. GCMs are widely used for impact assessment under climate change, but the common practice is to employ the output of a single GCM or single scenario, which ultimately results in various uncertainties (Srinivasa Raju and Nagesh Kumar 2018). The accuracy of GCMs, developed for coarse grid resolution, decreases with an increase in finer

spatial and temporal scales, rendering them unable to replicate sub-grid scale features. However, features at the sub-grid scale are important to hydrologists and water resources planners (Wang et al. 2004; Mujumdar and Nagesh Kumar 2012). GCMs are climate models designed to simulate a time series of climate variables globally, accounting for the greenhouse gases in the atmosphere in current and future scenarios (Rehana, 2019). Downscaling models are the statistical techniques to bridge the spatial and temporal resolution gaps between the GCMs and impact assessment studies. The study aimed to include such climate change projections in predicting river water quality variables. The limitations of the prediction models include difficulty in interpreting, accuracy heavily relying on the quality and quantity of the data, and overfitting. Such models carry assumptions that input features are independent features, models might identify correlations between features and the target variable, but they may not represent true causal relationships, and many models assume stationarity data.

1.3. Water Quality Variables

There are certain quality standards set up by international organizations like the World Health Organization (WHO) and the Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document “Parameters of Water Quality”, EPA mentions a total of 101 parameters that affect water quality in one way or another. The critical water quality parameters are Temperature, Dissolved Oxygen (DO), Biochemical oxygen demand (BOD), Ammonia, Nitrite, Nitrate, and pH, and important water quality parameters are Alkalinity/Hardness, Salinity, CO_2, and Solids (Loucks and van Beek 2017a). Out of all RWQ parameters, many of the physical, biological, and chemical characteristics of river water and its life cycle depend on the RWT, and that reduces the saturation concentration for DO. Fish and other aquatic organisms have not evolved the ability to adapt to rapid temperature fluctuation. Also, RWT is the first variable directly impacted by the intensification of air temperature due to climate change. Additionally, the RWT is involved in the self-purification capacity of the river in response to the pollutants and climate. So, it is evident that the RWT is a significant water quality parameter. The RWT exerts a major influence on biological activity and growth, affects water chemistry, can influence water quantity measurements, and governs the types of organisms that live in water

bodies (United States Geological Survey (USGS) 2018a). Another critical water quality parameter is DO, which measures the quantity of oxygen in milligrams per liter of water - the amount of oxygen available to living aquatic organisms can reveal a lot about water quality. A small amount of oxygen, up to about ten oxygen molecules per million of water, is dissolved in water. This dissolved oxygen is breathed by fish and zooplankton and is needed by them to survive (United States Geological Survey (USGS) 2018b). All forms of aquatic life use DO in river water; therefore, this constituent is typically measured to assess the health of rivers. Wastewater from human activities, decaying aquatic vegetation, polluted stormwater discharges, sewage effluent, and decaying aquatic vegetation all lower DO levels as they are decomposed by micro-organisms present in river water. There are two main routes for oxygen input in surface waters: transfer of oxygen directly from the atmosphere (a process called reaeration), and from plants as a result of photosynthesis (Federal Interagency Stream Restoration Working Group (FISRWG) 2014). Therefore, the study focused on two critical river water quality variables, i.e., RWT and DO, and assessed the impacts of these two variables on river water quality management under climate change. There are standards set by the Bureau of Indian Standards (BIS) (BIS10500:1991) and Central Pollution Control Board (CPCB) Standards for river water quality variables. According to BIS10500:1991, for example, RWT's acceptable range is 25 °C maximum, and the minimum tolerance limit of DO is 6 mg/l.

1.4. River Water Temperature Heat Transfer Process

The RWT is formed under the influences of various hydrometeorological, topographical, and geophysical factors operating over any drainage area. Among the factors, one should include solar radiation, the temperature of the air (Penchev 1972; Tsachev et al. 1982; Marinov 1990), convective heat exchange between the free water surface and the atmosphere (Dingman 1972), the intensity and duration of sunshine (Arnell 1996), the character of river feeding, etc. (Webb, 1992). Several other natural factors influence river water temperature, such as geological structures, presence of karst or deep artesian water, type, and state of soil-vegetation cover, altitude, afforestation and position of the watershed, etc. (Edinger et al. 1968; Gu and Li 2002; Nelson and Palmer 2007). The diverse anthropogenic activities in the watersheds also exert their effect, causing sharp and permanent changes in water temperature

(Webb, 1992). Usually, RWT samples collected 30 cm below the water surface from the point of interest (Central Water Commission 2018).

Different heat transfer processes are the principal contributors to the total surface heat exchange: (1) the net short-wave radiation hitting the water; (2) the net long-wave radiation leaving the water; (3) evaporation; (4) conduction and (5) melting of snow (Figure 1.1). The melting heat of water is abnormally high at about - 333.7 kiloJoule (KJ). Paily et al. (1974) have given a detailed description of the different components of the heat transfer processes, shown in Figure 1.1.

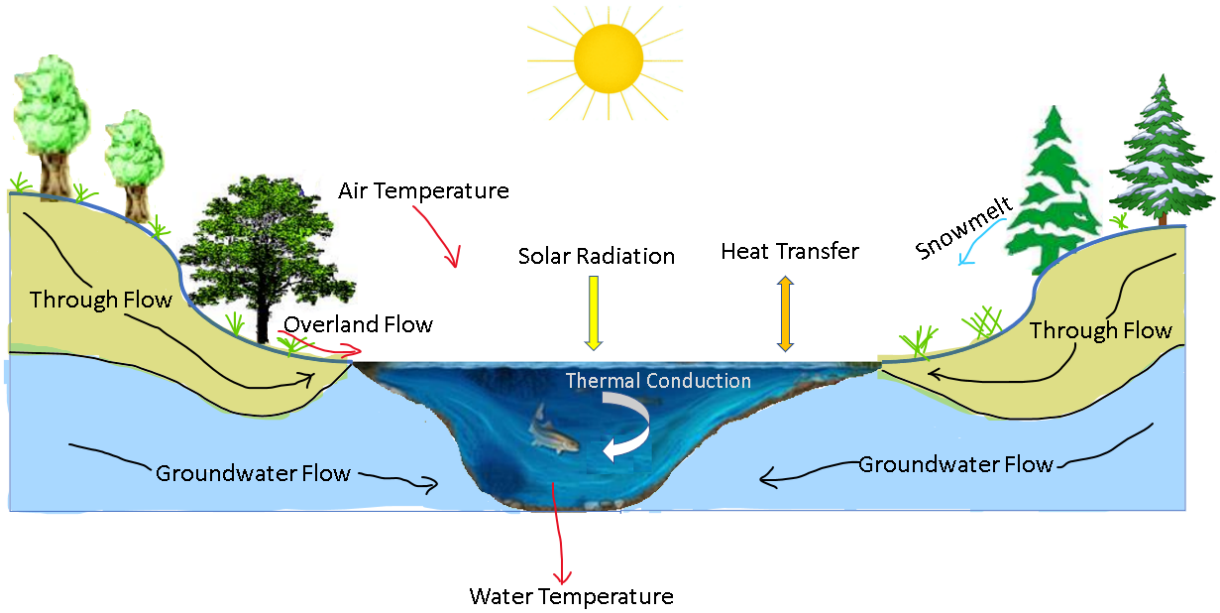


Figure 1.1. River Water Temperature Contributing Factors.

1.4.1. Air Temperature

Air temperature is the main factor affecting the river water temperature (Arnell 1996). The river water temperature (T_w) constantly seeks to achieve equilibrium with the surrounding air (T_a) at a rate that is proportional to the difference between the two temperatures. The heat transfer is determined by the equation:

$$Q = q(T_w - T_a) \quad (1.1)$$

where Q = rate of heat loss from the surface in calories per square cm day

q = energy exchange coefficient in calories/square cm day ^{o}C.

Equilibrium temperature (T_e) of the water surface is a state at which the net energy

exchange with the atmosphere stops (Dingman 1972). The heat is gained by the body of water when the temperature of the water surface (T_{ws}) is smaller than T_a and is lost when $T_{ws} > T_e$.

Temperature can further be defined as a measurement of the average thermal energy of a substance. Thermal energy is the kinetic energy of atoms and molecules, so the temperature in turn, measures the average kinetic energy of the atoms and molecules. This energy can be transferred between substances as the flow of heat. Heat transfer, whether from the air, sunlight, another water source, or thermal pollution, can change the temperature of water (Dingman 1972; Paily et al. 1974; Tsachev et al. 1982).

1.4.2. Stream Flow

For a given level of solar radiation or heat, stream temperature is inversely proportional to stream discharge. This relationship is illustrated in Brown (1972) prediction equation:

$$\Delta T = \frac{AN}{Q} \times 0.000167 \qquad (1.2)$$

according to it the expected change in stream temperature ΔT (°C) increases with stream surface area A (m^2) and net radiation load N (cal cm^{-2} min^{-1}), inversely with stream flow Q (m^3 s^{-1}).

The discharge Q is an important factor (agent) in the thermal feature of river water temperature because the lower the discharge is, the lower the capacity of the stream for heat storage. Thus, the temperature of small streams typical for headwater regions may increase significantly in summer by any rise in the air temperature.

1.4.3. Depth of the Water

Shallow waters are usually warmer than deep-water courses because they require less time to warm up. The temperature also changes from the surface to the bottom of the water column. The water column in deep lakes and other stagnant water is usually stratified because in summer, the water at the surface can get very warm from the sunshine and low thermal conductivity of water, while at the bottom, it remains cold (Arnell 1996).

1.4.4. Other Factors

There are several other factors that have a direct impact on RWT, such as Solar radiation, that have the greatest impact because they determine the heat exchange and fluxes taking place at the surface of the river (Sahoo et al. 2009), Flow regulation and construction of reservoirs (Webb and Walling 1993; Lowney 2000; Qiu et al. 2021), River geometry (Gu and Li 2002), and canopy cover (DeWeber and Wagner 2014). Land use characteristics that are directly or indirectly related to water temperature and useful in geographically diverse basins and over large spatial extents for better modeling. Especially for identifying the relationships between water temperature and each land cover type, local riparian forest cover, as several studies have shown the importance of shade from riparian vegetation on nearby temperatures (Nelson and Palmer 2007; DeWeber and Wagner 2014). Thermal pollution is any discharge that will dramatically alter the temperature of a natural water source and commonly comes from municipal or industrial effluents (Edinger et al. 1968; Webb and Nobilis 2007).

1.5. Prediction of Water Quality Variables

River water quality parameters such as RWT, DO, BOD, Total Dissolved Solids (TDS), Electrical Conductivity (EC), etc., form vital signs for defining the health of a river water body's ecosystem (Chapra et al. 2021). Accurate water quality prediction has an essential role in improving water management and pollution control (Nouraki et al. 2021). Water quality modeling studies demonstrated using various parameters, namely BOD, Chemical Oxygen Demand (COD), DO, Electrical Conductivity (EC), Nitrate-Nitrogen (NO_3-N), Nitrite-Nitrogen (NO_2-N), Phosphate (PO_4^{3-}), the potential for Hydrogen (pH), Sodium (Na), Temperature, Total Dissolved Solids (TDS), and Turbidity (TUR) for various basins globally (Asadollah et al., 2021; Azad et al., 2019; Danladi Bello et al., 2017; Du et al., 2019; Ficklin et al., 2013; Heddam et al., 2022; Nouraki et al., 2021; Rehana & Mujumdar, 2012, 2011; Santy et al., 2020; Ubah et al., 2021). Such studies modeled based on regression models (Santy et al. 2020) and process-based models such as QUAL2K (Ficklin et al. 2012, 2013; Du et al. 2019), and ML-based models(Azad et al. 2019; Nouraki et al. 2021; Heddam et al. 2022). In this study, the water quality parameters to demonstrate the RWQ modeling

were RWT and DO saturation levels. Below sections are the brief discussion about RWT and DO saturation levels estimation:

1.5.1. River Water Temperature Prediction

For many environmental, hydrology, and ecology applications, accurate prediction and assessment of RWT have become the key problem (Zhu et al. 2019e, a). RWT is a complex process to predict hydro climatological and river morphological parameters (Zhu and Piotrowski 2020). In this context, modeling RWT under different spatial and temporal scales is based on conceptual processes created on thermal advection-dispersion models (Sinokrot and Stefan 1993), equilibrium temperature-based models (Mohseni et al. 1999), statistical or machine learning (ML) models (Feigl et al., 2021; Isaak et al., 2017; Pike et al., 2013; Heinz G. Stefan & Preud'homme, 1993) and hybrid models (Toffolon and Piccolroaz 2015). Due to the simplicity of implementation, regression models have been improved using the relationship between air and water temperatures (e.g., Erickson Troy R. & Stefan Heinz G., 2000; Neumann David W. et al., 2003; Pilgrim et al., 1998; Rehana & Mujumdar, 2011; Heinz G. Stefan & Preud'homme, 1993). Deterministic models apply energy budget approaches to predict RWT (Sinokrot and Stefan 1993; Du et al. 2018; Zhu et al. 2019b), while statistical and ML models are grouped into parametric approaches, including regression (Mohseni et al. 1999; van Vliet et al. 2012) and stochastic (Ahmadi-Nedushan et al., 2007; Caissie, 2006; Mohseni et al., 1999; Rabi et al., 2015; Heinz G. Stefan & Preud'homme, 1993; Webb et al., 2003), and non-parametric approaches based on ML algorithms (Feigl et al. 2021). Unlike process-based models, ML models do not require many input variables, which are unavailable for many ungauged river systems and have been widely used in RWT modeling in recent years (Zhu and Piotrowski 2020). Particularly, river water quality management models which can predict RWT accounting for the hydroclimate and ambient meteorological variables of rivers based on such data-driven algorithms are prominent for river water quality control.

Artificial Neural Networks (ANN), which belong to the statistical group, gained much attention in the literature due to their ability to capture and represent complex nonlinear relationships between air and water temperature (Chenard and Caissie 2008; Sahoo et al. 2009). ANN has proven to be a promising mathematical tool for predicting the nonlinear

relationships and their applications in RWT predictions (Chenard and Caissie 2008; Sahoo et al. 2009; Hadzima-Nyarko et al. 2014; DeWeber and Wagner 2014; Rabi et al. 2015; Piotrowski et al. 2015; Temizyurek and Dadaser-Celik 2018; Zhu et al. 2018, 2019d, c; Qiu et al. 2020). Toffolon & Piccolroaz (2015) developed the air2stream hybrid model for RWT prediction. The air2stream model has been used in a variety of hydrological research over a variety of catchment sites, and results were usually better than ML models (Piccolroaz et al. 2016; Yang and Peterson 2017; Piotrowski and Napiorkowski 2018, 2019; Zhu et al. 2019f; Tavares et al. 2020).

The linear and non-linear regression models, as well as traditional ML models for RWT prediction, have some limitations, i.e., large modeling errors, particularly when it comes to non-stationary data processing (Graf et al. 2019). Because air temperature (AT) is an ecosystem "master variable", denoising the AT data could translate directly into improved prediction results (Magnuson et al. 1979). In this regard, wavelet transform (WT), a time-frequency localization method that can extract periodicities and trends, has frequently been coupled with ML methods due to their complementarity and can yield better performances compared to conventional forecasting models in various hydrological applications (Zhu et al. 2019c). Some studies have also demonstrated that hybrid techniques, which incorporate various ML techniques in different stages of the model construction, can be better than standalone ML since specific patterns in the time series (e.g., transients or trends) can be well encapsulated by various methods (Zhu et al. 2019c; Graf et al. 2019; Stajkowski et al. 2020). WT has been extensively applied in hydrology (Ebrahimi and Rajaee 2017; Sang et al. 2018; Roushangar et al. 2018; Honorato et al. 2018; Shoaib et al. 2019), their applications for the prediction of RWT have been very limited (Piotrowski et al. 2015; Zhu et al. 2019c; Graf et al. 2019).

ANNs have a large number of free parameters and thus require vast data sets, and therefore face slow training issues (Shen 2018). The rapid growth of data and advances in computation have led to powerful empirical tools such as deep learning (DL) (LeCun et al. 2015; Shen 2018). In contrast to ANN, one of the state-of-the-art deep learning architectures, the Recurrent neural network (RNN) (Rumelhart et al. 1986) is a neural network family for processing sequential data. This is achieved by having internal (hidden) states. The RNNs possess important features over ANN, namely, events from the past that can be retained and

used in current computations (Nagesh Kumar et al. 2004). The most widely known RNN is Long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997). LSTMs have recently been applied in a wide range of hydrological studies and showed promising results for time-series prediction tasks (Nagesh Kumar et al. 2004; Kratzert et al. 2018; Zhang et al. 2019; Kratzert et al. 2019; Xiang et al. 2020; Li et al. 2020b). RNNs also set accuracy records in multiple applications related to RWT prediction (Stajkowski et al. 2020; Feigl et al. 2021; Qiu et al. 2021).

Overall, RWT is the basic river water quality variable that has direct impacts under climate change due to alterations in hydro-climatic variables (e.g., air temperature, precipitation, streamflow, etc.). Further, RWT predictions have the direct implementation to assess the fish aquatic habitat (Centre for Climate Change Research (CCCR) 2017), to assess the deterioration of freshwater ecosystems and human water use (e.g., thermoelectric power and drinking water production, fisheries, and recreation) (van Vliet et al. 2013), and to assess saturated DO concentrations with respect to RWT (Ficklin et al. 2013). To this end, the assessment of RWT is of much relevance for Indian river systems due to minimum flows and higher temperatures during non-monsoon seasons, which is crucial for river water quality management. Furthermore, the considerable lack of measuring water quality data is a significant issue in many parts of the world and most Indian river systems for implementing process-based RWT models. There are only a few studies on the prediction of RWT for Indian river case studies, which are mainly on linear regression models (Chaudhary et al., 2019; Rehana & Mujumdar, 2011; Santy et al., 2020), Support Vector Regression (SVR) (Rehana, 2019).

1.5.2. Dissolved Oxygen (DO) Prediction

Global warming climates have also shown an adverse impact on RWT under intensification of various climatological defining variables, majorly Air Temperature (AT) (Stefan & Sinokrot, 1993; van Vliet et al., 2013; Webb et al., 2003). Intensification of RWT will have adverse impacts in terms of a decrease in river DO saturation levels, where most of the river water quality standards are defined based on such saturation levels (van Vliet et al. 2013). Precisely, saturation DO is a prominent indicator of river water quality and is considered a standard measure to define the pollutant extent (Central Water Commission 2019). The

influence of climate change on DO in relation to RWT can lead to water quality degradation and ecological distortion (Null et al. 2013; El-Jabi et al. 2014; Bayram et al. 2015; Lee and Cho 2015; Svendsen et al. 2016; Danladi Bello et al. 2017). RWT is inversely related to DO concentration that every change in RWT affects the river's ability to self-purify by lowering the amount of oxygen that can be dissolved and utilized for biodegradation (Intergovernmental Panel on Climate Change 2007; Khani and Rajaee 2017; Kauffman 2018). Hence, climate change impacts on RWT and saturation oxygen content are prominent to understanding the projected river water quality and possible alterations in quality standards under climate change warming signals. To this end, the assessment of DO saturation rates with respect to RWT is relevant for Indian river systems due to minimum flows and higher temperatures during non-monsoon seasons.

1.6. Significance of Study

Studying RWQ variables under climate change has become crucial for many environmental and hydrology applications (Feigl et al. 2021). The following implications and applications emphasize the significance of studying and predicting RWQ variables under climate change:

Ecosystem Health: Rivers and their associated ecosystems depend on water quality for their health and functionality. Changes in AT, precipitation, and runoff due to climate change can influence RWQ parameters such as RWT, DO, pH, nutrients, etc. For instance, the rate of chemical reactions generally increases at higher RWT under pronounced AT. The increased RWT may decrease DO levels and subsequently can interrupt aquatic habitats. Furthermore, RWT is a prominent variable for RWQ and aquatic habitat affecting DO concentrations, algal metabolism, fish growth, and production in aquatic systems (Wilby and Johnson 2020). The knowledge of RWT is of great scientific and practical importance for determining evaporation losses (Marinov 1990). Besides, the RWT is a prominent variable in the context of climate change as it is a function of climatic variables such as AT, humidity, solar radiation, and wind speed. For example, temperature fluctuations affect water density and hence water transport (Thomann and Mueller 1987). Also, RWQ variables are involved in the self-purification capacity of the river in response to pollutants and climate.

Public Health: RWQ and public health are closely related, especially when it comes to sources of drinking water and recreational activities. It is imperative to predict the potential

effects of climate change on waterborne diseases, algal blooms, and pollution levels in order to protect public health.

Industry and Agriculture: River water is frequently used in industry and agriculture for a variety of purposes, including irrigation and cooling. Stakeholders can adjust their practices to ensure sustainable use by anticipating changes in RWQ.

Scientific Research: Researching the connections between RWQ variables and climate change advances our knowledge of intricate environmental systems. This information can help interdisciplinary collaborations and direct future research.

To summarize, the ability to anticipate RWQ variables in the context of climate change is essential for risk management, ecological protection, public health, and scientific research(Chapra 1998).

1.7. Research Questions

Understanding the riverine pollution extent and impact assessment under climate and anthropogenic influences is challenging due to sparse spatiotemporal river water quality data (Read et al. 2019). Most of the studies evaluate water quality modeling with spatially dense data globally, but the development of systematic models combined with ML under data uncertainties context has not been intensively studied for the prediction of RWQ variables (Zhu et al. 2018; Isaak et al. 2020; Zhu and Piotrowski 2020).

To simulate RWQ parameters using data-driven algorithms, more input variables are required, which are unavailable for many ungauged river systems. For example, in RWT simulation, influencing variables are groundwater flow, snowmelt, through flow, solar radiation, heat transfer, river geometries, and discharge. However, this physical data may not be available for many ungauged basins. Climatic variables which are readily available, are the maximum, minimum, and average AT to build RWQ models with more accurate simulation and higher computational efficiency. In this context, most ML approaches have been applied without any detailed sensitivity analysis.

Therefore, our first question is:

Q1: *"How can sensitivity analysis reveal a deeper understanding of the underlying processes governing water quality in the river systems? How can a sensitivity analysis be*

coupled with ML approaches to select the most suitable and effective variables for predicting river water quality variables? How can we assimilate theory-driven understanding of rich processes with data-driven approaches to improve the predicted values based on the measurement data?"

To answer this, the global Sobol global sensitivity analysis (GSA) was performed to consider the most sensitive variables for given river water quality variable to be predicted. This study demonstrates how ML methods can be coupled with sensitivity analysis to predict RWT. Due to data availability, only three parameters (maximum, minimum, and average of AT) were considered, and tried to identify which one is more sensitive. Further, it is shown how ML methods can be coupled with data assimilation (DA) techniques to generate accurate RWT.

Another sort of data uncertainty is the lack of availability of long-time series data and consistent water quality measurement datasets in RWQ modeling. Generally, RWQ data availability is at monthly scales and is burdened with a large number of missing values with limited durations. Given the missing, limited, and non-stationary data scenarios, the present thesis proposes, the following second question as:

Q2: *"How to infer the relationships between river water quality indicators and hydroclimatic variables (e.g., Air Temperature (AT), streamflow)? How do different potentially influence lagged variables as additional predictive power features in river water quality modeling to improve the model performance under sparse, non-stationary data, and seasonality scenarios?"*

The river water quality data time series should be long enough to capture interannual variability, but there might be measurement errors, which leads to noisy data. To overcome the limited data scenarios, processing of non-stationary, and the noisiness of river water quality data, a methodology was developed using bootstrap resampling algorithms and wavelet approaches to predict river water quality variables. To demonstrate, the hybrid ML models were developed in this study to predict RWT under data uncertainties. Also, validated whether one variable time series at time *t*-lag provides important information helping to predict values of another variable time series at time *t* by using Granger Causality Analysis test.

However, global warming climate has shown an impact on river water quality variables in terms of changes in RWT and river flows. The present study aims to study the

climate change impact on RWQ parameters using the developed hybrid data-driven algorithm with the following third question as:

Q3: *"How do the climate change variables (e.g., temperature, precipitation) impact key physical processes within a river system (e.g., biological activities, dilution), ultimately influencing river water quality variables?"*

Climate change caused by anthropogenic greenhouse gases in the atmosphere directly impacts the quality of river water, which raises the possibility of the river ecosystem degrading in terms of decreased DO saturation levels under the decrease of stream flows and increase in RWT. For this reason, it is crucial to research how climate change will affect the thermal processes (e.g., RWT) and other self-purification capacity defining variables such as DO of river system. To demonstrate, this study assessed the climate change impacts on RWT and DO saturation levels. Furthermore, RWT and DO are the basic river water quality variables that directly impact under climate change due to alterations in hydro-climatic variables (e.g., air temperature, precipitation, streamflow, etc.). The projected changes in RWT and DO saturation levels are quantified using an ensemble of the NEX-GDDP dataset with RCP 4.5 and 8.5 experiments for 2071–2100, and hypothetical climate change scenarios.

1.8. Objectives

Based on the questions identified in the above discussion, the following objectives have been considered for the present research:

1. To apply various classical ML models (RR, KNN, RF, SVR) coupling with the sensitivity analysis for river water quality variables prediction.
2. To develop hybrid ML modeling framework for the prediction of river water quality variables under the limited and non-stationary data scenarios of given river water quality data.
3. To assess the climate change impacts on river water quality variables using an ensemble of the NEX-GDDP dataset with RCP 4.5 and 8.5 experiments for 2071–2100, and hypothetical climate change scenarios using hybrid data driven models.

1.9. Contributions of the Thesis

In this thesis, the hybrid ML models are proposed for RWQ predictions under data uncertainty scenarios by considering the RWT and DO saturation levels as explanatory variables for demonstration in order to answer questions Q1, Q2, Q3, and Q4 and accomplish the above objectives. The reason for choosing the RWT and DO is that out of all RWQ parameters, RWT is directly impacted by AT under climate change, which affects the characteristics of river water and its life cycle, and DO can reveal a lot about the health of rivers. The following contributions are made in each chapter, specifically:

- Presented literature on predicting RWQ variables, i.e., RWT and DO saturation levels, using various techniques, process-based deterministic models, regression-based models, and ML models. The chapter concludes with the robust and hybrid ML approaches to RWQ modeling under data uncertainties required, as an accurate simulation of RWQ variables plays an important role in water quality management (Chapter 2).
- Demonstrated how new ML approaches, such as RR, KNN regressor, RF regressor, and SVR, can be coupled with Sobol' GSA to predict accurate RWT estimates with the most appropriate form of AT (e.g., maximum, minimum, and average) in cases when there is a lack of data. The chapter is concluded with the hybrid models are the new promising frameworks by coupling with a global sensitivity algorithm for accurate RWT predictions under lack of data scenarios and may deserve further study in the field of hydrology and water resources (Chapter 3).
- Demonstrated hybrid models using LSTM, integrated with (i) kNN bootstrap resampling algorithm (kNN-LSTM) to address the lack of availability of long-time series data of RWT prediction, (ii) WT approach (WT-LSTM) to address the time-frequency localized features of RWT prediction under data uncertainties. Compared the performance results of hybrid models with LSTM, air2stream. Assessed the climate change impacts on RWT using an ensemble of the NEX-GDDP GCM dataset under RCP scenarios 4.5 and 8.5 for seven major polluted river catchments of India at monthly time scale. It is concluded that the hybrid models yielded better performance results than standalone LSTM and air2stream forecasting models at a monthly scale under data uncertainties, i.e., when WT and k-NN bootstrap resampling algorithms

were included. Also used the Granger Causality Analysis test to assess whether one variable at time *t*-lag causes another variable at time *t*. (Chapter 4).

- Demonstrated (i) the combined effects of streamflow and AT in ML models for prediction of RWT variables under sparse data scenarios, (ii) compared the performance results of the kNN-LSTM model with standalone LSTM, nonlinear regression model, and 8-parameter version Air2Stream in the prediction of RWT, (iii) the assessment of climate change impacts on the rate of change of oxygen saturation with respect to RWT, and streamflow using an ensemble of the NEX-GDDP GCM dataset and hypothetical climate change scenarios. The chapter is concluded with the assessment of the individual contribution of RWT rise on depletion of saturated DO levels (Chapter 5).
- Presented the review of the results of this thesis, discussed some limitations, and suggested a number of future directions that would further improve the methodology of this thesis (Chapter 6).

Chapter 2
LITERATURE REVIEW

2.1. Introduction

This chapter presents a brief review of the literature related to past research on climate change impacts on river water quality variables and prediction of river water quality variables, i.e., RWT and DO saturation levels using the process based, regression based, hybrid, and ML based models. Few conclusions are drawn from the literature on the prediction of river water quality variables. These conclusions provide relevance to the current research. In the following section, a detailed review of relevant literature has been included.

2.2. River Water Quality Modeling and Prediction

River water quality modeling and predictions are useful for monitoring of current and assessment of future water quality scenarios resulting from different management strategies (Loucks and van Beek 2017b). Water quality predictive models include both mathematical expressions and expert scientific judgment. They include process-based (mechanistic) models and data-based (statistical) models. River water quality parameters are directly affected by any changes in climate, which in turn increases the risk of deterioration of the river ecosystem (Rehana and Dhanya 2018). Many previous water quality modeling studies have been carried out to assess the impact of climate change on river water quality, i.e., increased air temperatures and reduced summer flows may further exacerbate water temperature increases (Chapra et al., 2021; Chaudhary et al., 2019; Du et al., 2019; Ficklin et al., 2013; Isaak et al., 2012; Daniel J. Isaak et al., 2010; Nelson & Palmer, 2007; Rehana & Mujumdar, 2012; Santy et al., 2020; van Vliet et al., 2013; van Vliet & Zwolsman, 2008). Rehana & Mujumdar (2012) used an empirical joint probability distribution of monthly average streamflow and river water temperatures to estimate the risk of low water quality for a given DO threshold. Ficklin et al. (2013) assessed the impacts of climate change on water quality variables using SWAT and observed that 10% decrease in DO by 2100 at Sierra Nevada in California, USA. van Vliet et al. (2013) assessed the global daily simulations of river flow and RWT under climate change based on a physically based hydrological-water temperature modeling framework and concluded that the impact of discharge changes generally increases during dry warm periods when rivers have a lower thermal capacity. Chaudhary et al. (2019)

assessed the impacts of climate change and dry seasons on water quality indicators, i.e., DO, BOD using QUAL2K water quality model for the Yamuna and Bhadra rivers in India. The research findings of the study show that current river flow conditions are ineffective at maintaining acceptable water quality criteria and proposed possible scenarios for Eflows to improve the water quality standards. Santy et al. (2020) assessed the impacts of climate change and land cover change on water quality indicators, i.e., DO, BOD, etc., using the QUAL2K water quality model for the heavily industrialized stretch of the Ganga River in India and found that DO of the critical points is reduced with RWT increase due to increased reaction kinetics at higher temperature in the climate change scenarios, while RWT is modelled using linear regression. Heddam et al. (2022) studied the river nitrate concertation predictions using ML models with different water quality variables namely, RWT, DO, specific conductance, water turbidity, water pH, and river discharge, a case study of Willamette River at Portland, Oregon, USA, and concluded that water quality variables contribute significantly for the improvement of the performances of ML models in accurate prediction of nitrate concentrations. Ubah et al. (2021) forecasted the water quality parameters namely pH, Total Dissolved Solids (TDS), Electrical Conductivity (EC), and Sodium (Na) using artificial neural network, a case study of Ele River, Anambra State for irrigation purposes, and showed that TDS, EC and Na were above the permissible standard for irrigation during dry seasons while the pH was normal all through the season. Nouraki et al. (2021) predicted the water quality parameters namely TDS, sodium absorption ratio (SAR) and total hardness (TH) using ML models, a case study of the Karun River, Iran and showed that ML models could satisfactorily estimate the TDS, SAR and TH for all stations. Azad et al. (2019) forecasted the EC, TDS, SAR, carbonate hardness (CH), and TH water quality parameters using Artificial intelligence models, a case study of the Zayandehrood River, in Iran and showed that adaptive neuro fuzzy inference system is the best ML model for different water quality parameters prediction. Asadollah et al. (2021) used the monthly input water quality data including BOD, COD, DO, EC, Nitrate-Nitrogen (NO_3-N), Nitrite-Nitrogen (NO_2-N), Phosphate (PO_4^{3-}), potential for Hydrogen (pH), Temperature and Turbidity (TUR) for building the water quality prediction models, a case study of Lam Tsuen River in Hong Kong and showed that Extra Tree Regression model performed best for water quality prediction.

Recently, to simulate RWQ variables, such as RWT, the Air2Stream model (Shrestha and Pesklevits 2022; Zhu et al. 2022), Temperature Duration Curve (TDC) (Ouarda et al. 2022), Process-based models (e.g., Delft3D model, Soil and Water Assessment Tool (SWAT) model, etc.), have been used (Wang et al., 2022). However, such approaches require large amounts of site-specific detailed data at various time scales, including stream geometry and meteorological and hydraulic properties of the river (Piccolroaz et al. 2016).

This thesis demonstrates the importance of climate-driven changes in hydrology as fundamental to understanding changes in the local water quality. In particular, the focus on changes in RWT, and DO saturation levels. In the following sections, a detailed review of RWT and DO saturation levels literature has been included.

2.2.1. Review of Studies on River Water Temperature Prediction

2.2.1.1. Process based Deterministic Models

For many environmental, hydrology, and ecology applications, accurate prediction, and assessment of RWT has become the key problem (Zhu et al. 2019e, a). In this context, process-based RWT models have been evolved based on heat advection-dispersion transport equations (Stefan and Sinokrot 1993) and net heat transfer processes at the surface based on thermal equilibrium concepts (Mohseni et al., 1999; Rehana & Mujumdar, 2012).

United States Environmental Protection Agency (USEPA) developed the QUAL2K modeling framework to simulate water quality indicators (Chapra and Pelletier, 2003). In the QUAL2K model, the river is divided into reaches and each reach is further divided into a series of equally spaced elements. The governing equations of QUAL2K model are the advection-dispersion-reaction equations with external sources and sinks. The model permits the input of wastewater discharges, tributary flows, incremental flows and withdrawals. QUAL2K can simulate total 24 water quality constitutes including RWT, DO, BOD, pH, etc. The drawback of QUAL2K model includes its one-dimensionality and high input data requirements.

Ficklin et al. (2012) developed the hydroclimatological stream temperature model within the SWAT model to consider hydrology and air temperature's impact in simulating the water-air heat transfer process. In this new model, RWT is determined as a function of three components (i) temperature and amount of local water within the subbasin (snowmelt contribution, groundwater contribution, surface runoff, lateral soil flow), (ii) temperature and

inflow volume from upstream subbasins(s) (iii) air-water temperature transfer during the streamflow travel time in the subbasin. Du et al. (2018) modified the hydroclimatological model (Ficklin et al. 2012) by including the equilibrium temperature approach to model heat transfer processes at the water-air interface, which reflects the influences of air temperature, solar radiation, wind speed, and streamflow conditions on the heat transfer process. It is a computationally expensive model for large-scale simulations.

Toffolon & Piccolroaz (2015) developed the Air2Stream model for predicting RWT, which combines a physical based structure with a stochastic parameter calibration. Air2stream uses the inputs AT and streamflow and was derived from simplified physical relationships expressed as ordinary differential equations for heat budged processes.

Overall, process-based models are mathematical representation of the underlying physics, and it provides exact results. However, such process-based models are complex in nature and large amount of detailed and computationally intensive data is required.

2.2.1.2. Regression based Models

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 R. & Stefan Heinz G. 2000; Neumann David W. *et al.* 2003; Rehana & Mujumdar 2011). The usual illustrations are linear regression models (Morrill et al. 2005; Krider et al. 2013), nonlinear 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 and Piccolroaz 2015; Piccolroaz et al. 2016) have been developed successfully for data relating to different time scales in the past years.

van Vliet et al. (2013) assessed the global daily simulations of river flow (Q) and water temperature (Tw) under climate change based on a physically based hydrological-water temperature modeling framework. The basic idea is to assess the impact of climate change on daily river discharge and water temperature on a global scale using a physically based hydrological and water temperature modeling framework forced with an ensemble of daily bias-corrected GCM output.

Islam et al. (2019) investigated the impact of both air temperature and streamflow changes on river water temperatures using the Air2Stream model on Canada's Fraser River Basin (FRB). The basic idea is the quantification of climate change impacts on the thermal regimes of rivers in British Columbia (BC) using the Air2Stream model. Implement and evaluate a hybrid water temperature model at 17 river sites in the FRB.

2.2.1.3. Machine Learning based Models

ANN has proven to be a promising mathematical tool for predicting nonlinear relationships and their applications in RWT predictions (Chenard and Caissie 2008; Sahoo et al. 2009; Hadzima-Nyarko et al. 2014; DeWeber and Wagner 2014; Rabi et al. 2015; Piotrowski et al. 2015; Temizyurek and Dadaser-Celik 2018; Zhu et al. 2018, 2019d, c). In recent years Zhu *et al.* (2018, 2019a, b) 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 nonlinear 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 (Heddam & Kisi, 2018; F. Huang et al., 2017; Komasi et al., 2018; Rasouli et al., 2012; Rehana, 2019; W. Wang et al., 2013). Random Forest (RF) models have been used extensively in hydrology (Balk and Elder 2000; Tehrany et al. 2013; Li et al. 2020a) and few researchers have applied for RWT modeling (Lu and Ma 2020). The K-nearest neighbors (KNN) approach has been used in many hydrology applications (Souza and Lall 2003; Beersma and 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).

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 and 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 and 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 relationship between average air and water temperatures (Piccolroaz et al. 2016; Rehana and Dhanya 2018; Zhu et al. 2018, 2019a; Rehana 2019; Graf et al. 2019). However, at maximum air temperatures, 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 appropriate AT variables (average, maximum, and minimum) has not been intensively studied in the literature. To our 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.

Most Indian River systems are burdened with data limitations and form a significant challenge for implementing process based RWT models. There are only a few studies on the prediction of RWT for Indian river case studies, which are mainly on linear regression models (Chaudhary et al., 2019; Rehana & Mujumdar, 2011; Santy et al., 2020), Support Vector Regression (SVR) (Rehana, 2019). Therefore, given the limitations over data availability for Indian river systems, developing hybrid models of DL is the viable solution to produce more accurate results. In this context, ANN, a statistical group, has proven to be a viable technique for RWT forecasting (Chenard and Caissie 2008; Sahoo et al. 2009; Hadzima-Nyarko et al. 2014; DeWeber and Wagner 2014; Rabi et al. 2015; Zhu et al. 2018, 2019d, c; Qiu et al. 2020). However, ANNs have many free parameters, and they require a significant amount of data and are therefore burdened with slow training issues (Shen 2018). In contrast to ANN, one of the state-of-the-art DL architectures, Recurrent neural network (RNN) (Rumelhart et al. 1986), which possesses important features over ANN, namely, events from the past can be retained and used in current computations (Nagesh Kumar et al. 2004). The most well-known RNN is Long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997). Recently, LSTM has been successfully used in hydrological modeling applications and exhibited promising results (Li et al., 2020; Nagesh Kumar et al., 2004;

Xiang et al., 2020; Zhang et al., 2019). RNNs also set accuracy records in multiple applications related to RWT prediction (Stajkowski et al. 2020; Feigl et al. 2021; Qiu et al. 2021).

Feigl et al., (2021) propose a novel six different machine learning models: stepwise linear regression, Random Forest, eXtreme Gradient Boosting (XGBoost), Feedforward neural networks (FNN), and two types of Recurrent neural networks (RNN) to improve the performance of RWT prediction. To make the results comparable to previous studies, two widely used benchmark models have been applied additionally: linear regression and air2stream. The basic idea is most studies mainly use air temperature and discharge as inputs for water temperature prediction. The evaluated input data sets include combinations of daily means of air temperature, runoff, precipitation, and global radiation. The applied data preprocessing consists of feature engineering (i.e., deriving new features from existing inputs) with lags of all variables for the four previous days are computed and used as additional features. Machine learning models are generally parameterized by a set of hyperparameters that have to be chosen by the user to maximize the performance of the model. Depending on the model, hyperparameters can have a large impact on model performance. To optimize the hyperparameters of nearly all machine learning models in this study with the Bayesian optimization method.

Zhu, Nyarko, Hadzima-Nyarko, et al. (2019) proposes a different versions of feedforward neural network (FFNN), Gaussian process regression (GPR), and decision tree (DT) models were developed to estimate daily river water temperature using air temperature (Ta), flow discharge (Q), and the day of the year (DOY) as predictors. Modeling results were compared with the air2stream model. The Basic idea is in this research, ANN, GPR, and DT models were developed for eight river stations characterized by different hydrological conditions using Ta, flow discharge (Q), and day of the year (DOY) as predictors. When the day of the year was included as model input, the performances of the three ML models dramatically improved. Including flow discharge instead of the day of the year as an additional predictor, provided a lower gain in model accuracy, thereby showing the relatively minor role of flow discharge in RWT prediction. However, an increase in the relative importance of flow discharge was noticed for stations with high altitude catchments which are influenced by cold water releases from hydropower or snow melting, suggesting the

dependence of the role of flow discharge on the hydrological characteristics of such rivers.

Zhu et al. (2018) propose the air-water temperature relationship of the Missouri River is investigated by developing three different machine learning models (ANN, Gaussian Process Regression (GPR), and Bootstrap Aggregated Decision Trees (BA-DT)). The basic idea is although many factors influence the prediction of river water temperature, the objective of this study was to estimate the daily water temperature of the Missouri River with the aid of only the mean air temperature.

Rehana (2019) adopted Support Vector Machine (SVR) to analyze the predictability performance of RWT. The proposed machine learning algorithm of SVR is applied with air temperature and streamflow as predictors to estimate the RWT at the Shimoga river water quality checkpoint along Tunga-Bhadra, a tributary of Krishna River, India. The future RWT projections were analyzed using trained and tested models with the downscaled projections of air temperature and streamflow. In this context, there are limited studies for testing the predictability of RWT with SVR in the literature. This study revealed that the SVR model had been identified as the best prediction performance compared to linear regression models.

Graf et al. (2019) propose a hybrid model that couples discrete wavelet transforms (WT) and ANN for forecasting water temperature. Four mother wavelets, including Daubechies, Symlet, discrete Meyer, and Haar, were considered to develop the WT-ANN hybrid model and examined the importance of the choice of decomposition level. The basic idea is traditional ANN models for river water temperature modeling frequently have limitations, especially in the processing of non-stationary data, which most hydrological time-series datasets. In this regard, wavelet transform, as a good pre-processing method for non-stationary data, can be a potential complement to traditional methods to improve performance. Choices of the appropriate mother wavelet and the decomposition level are the two main issues in the applications of DWT. In this paper, a hybrid model based on coupling discrete WT and ANN for daily river water temperature forecasting was proposed. Compared with previous studies, four widely used mother wavelets were evaluated: Daubechies (Db), Symlet (Sym), discrete Meyer (dMey), and Haar. This study revealed that model performances improved with an increase in the decomposition level in the wavelet transform, and the discrete Meyer (dMEY) mother wavelet performed the best.

Qiu et al. (2021) explored the potential of a long short-term neural network (LSTM),

a type of deep learning method, to forecast daily river water temperatures and quantify temporal variations in thermal regimes induced by changes in climate and by dam construction. The basic idea is data-driven methods have not explored fully the extent to which river WT forecasting can be improved through the application of deep learning methods. This study evaluated the forecasting performance of an LSTM model to predict mean daily river temperatures. To evaluate the influence of the dam on water temperatures. The construction of the reservoir strongly influenced water temperature variations, producing the strongest cooling effect from mid-April to mid-May when it produced cooling of about 4 ˚C and the greatest warming effect in late December and early January when it produced a warming of about the same amount.

Stajkowski et al. (2020) adopted a genetic algorithm (GA)-optimized LSTM technique to predict river water temperature (WT). The basic idea is an LSTM model consists of several parameters, such as the number of hidden layers, number of epochs, batch size, learning rate, number of units, and window size (previous time steps). In LSTM, the largest difficulty arises due to the selection of the window size and the number of units. Therefore, this research investigated this problem through a genetic search. The optimal window size and the number of units based on the lowest RMSE was determined in this search using a genetic algorithm, and the best window size and the number of units were fed into an LSTM model.

The robustness of any DL-based forecasting algorithms, i.e., ANN or RNN, depends on the extensive input data. These models can extract time characteristic information with long time series data without errors and missing data points. But the availability of hydrological data in India is limited, with small temporal resolutions.

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, understanding-rich processes with data-driven discovery processes (Babovic 2005). Recent progress in ML inspires the idea of learning DA models directly from 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 2005). The DA method can be categorized into four groups (WMO 1992; Babovic

2005): (a) updating input parameters; (b) updating model parameters; (c) updating state variables, and (d) 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 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 (Li et al., 2013; Liu et al., 2010; Mehrparvar & Asghari, 2018; Wang et al., 2016, 2017; X. Wang & Babovic, 2016). In RWT forecasting, only a few studies addressed the use of DA (Morrison and Foreman 2005; Yearsley 2009; Pike et al. 2013; Ouellet-Proulx et al. 2017).

Overall, a visual illustration of RWT prediction model milestones is shown below:

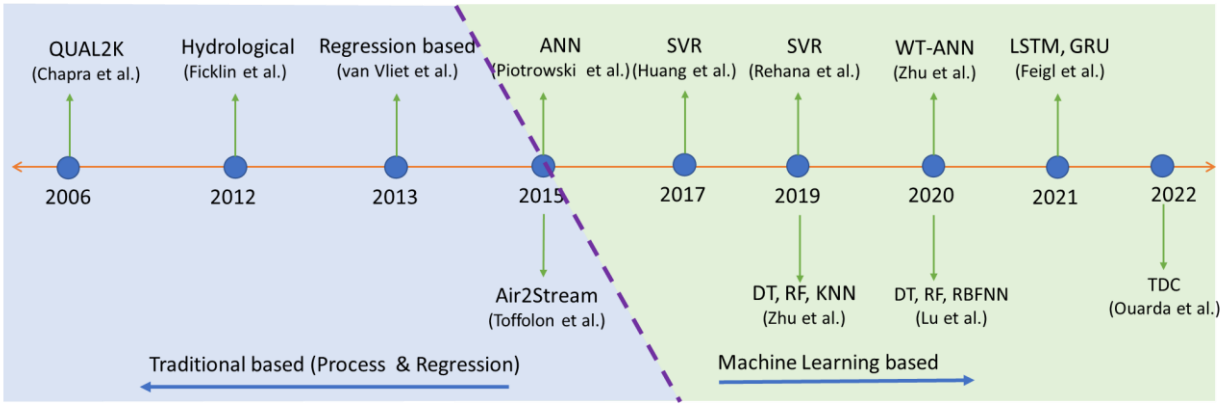


Figure 2.1. River Water Temperature model milestones

2.2.2. Review of Studies on Dissolved Oxygen Saturation Levels Prediction

Water quality modeling studies predicted depletion of DO under streamflow, RWT, and land use changes for various basins globally (Chapra et al., 2021; Cox & Whitehead, 2009; Danladi Bello et al., 2017; Du et al., 2019; Ficklin et al., 2013; Harvey et al., 2011; Rehana & Mujumdar, 2012, 2011; Santy et al., 2020). Harvey et al. (2011) assessed the influence of AT on RWT and the Concentration of DO in Newfoundland Rivers in Canada using regression models and determined that the exponential model was found to be better suited to modeling low DO concentrations at higher RWTs in a temperate climate. Ficklin et al. (2013) assessed the effects of climate change on RWT, DO, and sediment concentration in the eastern and western watersheds of the Sierra Nevada Mountain range in California, USA using the Soil

and Water Assessment Tool (SWAT) with a newly developed stream temperature model, which simulates RWT and associated water quality parameters based on AT and the effects of local hydrology (Ficklin et al. 2012). Results show that RWT increases by up to 6 ^{o}C for summer, reaching close to 30 ^{o}C in the lower-elevation reaches, decreases DO by 2%–12%, and overall decreases in sediment concentrations. Danladi Bello et al. (2017) predicted the impact of climate change on RWT and DO in tropical rivers in the Skudai Watershed located in the southern part of the Malaysia peninsular using the Hydrological Simulation Program FORTRAN (HSPF) model. Results show that an increase in AT will have little effect on RWT and DO concentrations. Also concluded that high to moderate stream flows lower RWT and increase the DO concentration. Du et al. (2019) assessed climate change impacts on RWT in the Athabasca River Basin, Canada using the SWAT equilibrium temperature model, and findings show that annual RWT is expected to increase by 1.6 to 3.1 °C and DO concentrations on the basin average scale will decrease by 0.72 mgO_2/L under RCP 8.5 scenario for 2061-2100. Chapra et al. (2021) assessed how river oxygen levels will be influenced by rising RWTs due to global warming, and findings shows that freshwater saturation concentrations at 5 °C increments between 20 and 35 °C are indicated because this range would encompass the present and future summer RWT's in most of the world's rivers over the next 50 years. In India, most river systems are burdened with data limitations and form a significant challenge for implementing process based RWT models to predict DO. Only a few studies used the regression models to assess DO variations for Indian river case studies (Chaudhary et al., 2019; Rehana & Mujumdar, 2011; Santy et al., 2020). Chaudhary et al. (2019) assessed the impacts of climate change and the dry seasons on water quality indicators, i.e., DO, BOD using the QUAL2K water quality model for the Yamuna and Bhadra rivers in India, and the findings show that current river flow conditions are ineffective at maintaining acceptable water quality criteria. To assess this, average monthly RWT data were obtained from the CPCB for the Yamuna stretch, and the linear regression equation developed by Rehana & Mujumdar (2011) was adopted for the Bhadra stretch. Santy et al. (2020) assessed the impacts of climate change and land cover change on water quality indicators, i.e., DO, BOD, etc. under hypothetical scenarios using QUAL2K water quality model and RWT is modeled using linear regression. Results show that DO of the critical points is reduced with RWT increase due to increased reaction kinetics at higher

temperature in the climate change scenarios. However, such studies are basin or river stretch specific, data intensive, and limits application for data sparse and ungauged locations with an emphasis on simulated DO levels in response to streamflow, RWT, and land use (Ficklin et al., 2013; Rehana & Mujumdar, 2012, 2011; Santy et al., 2020). However, DO saturation level, which serves as a baseline to measure oxygen-based water quality by determining the oxygen concentration of unpolluted water depending on RWT, salinity, and oxygen partial pressure (Chapra et al. 2021) and prominent in defining the maximum permissible limits and standards for various river usages (CPCB, 2019; Rehana & Mujumdar, 2009), has not been assessed under climate change. Specifically, while some recent studies have looked at how climate change affects RWTs, the question of how climate change affects saturation DO has yet to be answered. More specifically, the direct integration of RWT predictions in the assessment of DO saturation concentration levels under climate change signals has not been quantified.

2.3. Summary of Literature

Most of the studies assessed water quality modeling with spatially dense data globally, however, the development of systematic models paired with ML models under the context of data uncertainties has not been intensively examined for the prediction of RWQ variables (Zhu et al. 2018; Isaak et al. 2020; Zhu and Piotrowski 2020).

To simulate RWT, the Air2Stream model (Shrestha & Pesklevits, 2022; Zhu et al., 2022), Temperature Duration Curve (TDC) (Ouarda et al., 2022), Process-based models (e.g., Delft3D model, Soil and Water Assessment Tool (SWAT) model, etc.), have been used (Wang et al., 2022). However, such approaches require large amounts of site-specific detailed data at daily time scales, including stream geometry, meteorological and hydraulic properties of the river (Piccolroaz et al., 2016). However, globally, RWT data availability is at monthly scales and is burdened with large number of missing values with limited durations. Most of the ML studies (van Vliet et al. 2013; Zhu et al. 2018; Zhang et al. 2019; Graf et al. 2019; Stajkowski et al. 2020; Feigl et al. 2021; Qiu et al. 2021; Ouarda et al. 2022) assessed water quality modeling with spatially dense data at daily time scale. The Air and water temperature series should be long enough to capture interannual variability but the availability of hydrological data in India is limited with small temporal resolutions, hence we are unable to apply these models to rivers in India.

To simulate RWQ parameters using data-driven algorithms, more input variables are needed, which are not available for many ungauged river systems. For example, in RWT simulation, influencing variables are discharge, groundwater flow, snowmelt, through flow, solar radiation, heat transfer, river geometries, and wind speed. However, this physical data may not be available for many ungauged basins. Climatic variables which are readily available, are the maximum, minimum, and average AT to build RWT models with more accurate simulation and higher computational efficiency. In this context, most ML approaches have been applied without a detailed sensitivity analysis.

Another data uncertainty is the lack of availability of long-time series data and consistent water quality measurement datasets in RWQ modeling. Generally, RWQ data availability is at monthly scales as most of the measurement/sampling time intervals are monthly, which are further burdened with a large number of missing values with limited durations. In this context, the selection of appropriate model inputs, development of models under limited data, processing of non-stationary data, seasonality scenarios, and relevant lags of variables have not been intensively investigated in the literature, especially in the case of estimation of RWT.

The most influencing RWQ variable affecting under increased RWT is Saturated DO, defining the self-purification capacity and forming a basis for the standards or permissible limits (CPCB, 2019; Rehana & Mujumdar, 2009). Furthermore, it is essential to quantify the climate change impacts on thermal processes (e.g., RWT) and other self-purification capacity defining variables such as the DO of the river system. In this context, how climate change affects saturation DO with respect to the RWT has yet to be answered.

RWT is directly influenced by multiple parameters, including streamflow, solar radiation, wind speed, river geometry, groundwater inputs, slope, water depth, etc., which are not considered as inputs in the ML models for RWT prediction. In this context, building ML model with multiple input variables and validating these results with conceptual models are yet to be explored. Finally, the usefulness of machine learning models in RWQ modeling needs to be validated. Given the limitations over data uncertainties for Indian river systems, hybrid models need to be developed to produce more accurate results.

Chapter 3
PREDICTION OF RIVER WATER QUALITY VARIABLES USING CLASSICAL MACHINE LEARNING ALGORITHMS BY INTEGRATING THE SENSITIVITY ANALYSIS

3.1. Introduction

The majority of the river systems are burdened with data limitations, which make it difficult to implement process-based RWQ models at various spatial and temporal scales (Read et al. 2019; Pohle et al. 2019). Though process-based models have the potential to produce accurate results, they also require extensive amounts of site-specific detailed data (Piccolroaz et al. 2016). Statistical models have numerous benefits over process-based models, such as simplifying the complex relations between water quality indicators and the identifying patterns of comparable temporal and spatial characteristics among water quality variables. However, a major limitation of the statistical approach is its inability to accurately describe the nonlinear characteristics of a data series. To overcome this shortage, Machine Learning (ML) techniques are widely used to address a range of nonlinear prediction problems. To simulate RWQ parameters using ML algorithms, more input variables are required. For example, in RWT simulation, influencing variables are solar radiation, snowmelt, discharge, etc. However, this physical data may not be available for many ungauged basins (Zhu and Piotrowski 2020). The use of ML models using minimum data inputs (such as AT) can be robust in addressing data sparsity when simulating RWQ models. In this context, most ML approaches have been applied without any detailed sensitivity analysis.

For RWT modeling, in most studies, more than a single model is used to assess model performance (Zhu and Piotrowski, 2020). The Support vector regression (SVR) model, which is based on structural risk minimization to avoid overfitting (Vapnik et al. 1996), has been widely used in hydrology (Heddam and Kisi, 2018; Huang et al., 2017; Komasi et al., 2018; Rasouli et al., 2012). As per Zhu and Piotrowski (2020) only two studies have applied SVR models for river temperature prediction (e.g., Lu and Ma, 2020; Rehana, 2019). Random Forest (RF) is a variant of the bagging ensemble technique, which has been frequently used in hydrology (Balk and Elder, 2000; Li et al., 2020; Tehrany et al., 2013) and has been applied by few researchers in water temperature modeling (Lu and Ma, 2020). The KNN approach has been used in many hydrology applications (Beersma and Buishand, 2004;

Leander et al., 2005; Souza and Lall, 2003) and can be a proper choice for RWT predictions (Muluye 2012; Antunes et al. 2018; Gavahi et al. 2019). Hence, for uniqueness, globalization of the solution, working with less data, being less computationally expensive, and avoiding overfitting, the current work focused on SVR, RF, KNN, and Ridge regression. Therefore, the study considered these four ML models along with sensitivity and data assimilation algorithms to demonstrate how RWT model performances can be improved.

The present study proposed a Global Sensitivity Algorithm variance based on Sobol' method (Sobol 1990; Sobol′ 2001) to predict more influencing AT variable by selecting highly sensitive features 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). This study proposed an integrated modeling framework with ML, GSA and DA approach to improving the predicted values based on the measurement data. Changes in AT can affect RWT, the primary variable that influences water quality. Therefore, to demonstrate in this study, the RWT has been selected as water quality variable for prediction. Due to data availability, only three parameters (maximum, minimum, and average of AT) were considered, and tried to identify which one is more sensitive. The proposed algorithm has been demonstrated with a river gauging station daily temperature data of Shimoga station along Tunga River, a tributary of Tunga-Bhadra River, a major tributary of Krishna River, India.

In summary, the objectives of the this chapter are to (i) identify the most influencing AT variable by GSA algorithm (ii) apply various classical ML models (Ridge regression (RR), K-nearest neighbors (KNN) regressor, Random Forest (RF) regressor, Support Vector Regression (SVR)) with the best selected AT for RWT prediction (iii) applying Ensemble Kalman Filter (EnKF) with each ML model (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.

3.2. Methodology

The overview of the proposed modelling framework is shown in Figure 3.1 and 3.2. 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 timescale. Figure 3.1 shows the architectural flow diagram proposed for the prediction of RWT using sensitivity and ML. Figure 3.2 shows the ML model and EnKF data assimilation method's architectural flow diagram to improve the ML model's efficiency in each simulation step.

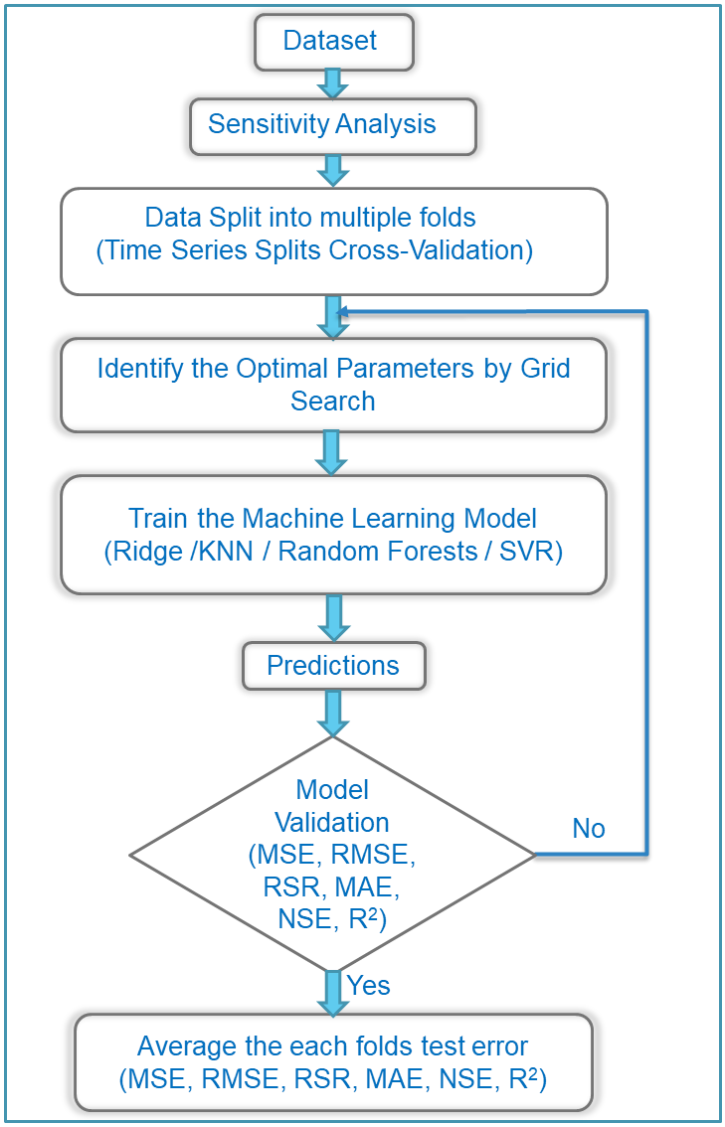


Figure 3.1. Architectural flow diagram for ML regression models.

3.2.1. 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 Sensitivity analysis environmental models are available in the literature (Saltelli et al. 2010; Yang 2011), based on variance decomposition. The variance-based Sobol' method is a sensitivity analysis 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 (SI's).

3.2.2. Sobol' Sensitivity Analysis Method

The method of Sobol' is an advanced, global, model-independent sensitivity analysis method that is based on variance decomposition. It can handle nonlinear 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 $X = (X_1, \dots\dots, X_M)$. For further developments, $f_{Xi}(x_i)$ and $f_x = \Pi_{i=1}^{M} f_{Xi}(x_i)$ refer to their marginal probability density function (PDF) and the corresponding joint PDF of a given set. The sensitivity model can be defined as:

$$Y = f(X) = f(X_1, \dots\dots, X_M) \tag{3.1}$$

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

$$f(X_1, \dots\dots, X_M) = f_0 + \sum_{i=1}^{M} f_i(X_i) + \sum_{i=1}^{M} \sum_{j=i+1}^{M} f_{ij}\left(X_i, X_j\right) + \cdots + f_{1,\dots,M}(X_1, \dots, X_M) \tag{3.2}$$

If the input factors are independent of each term in equation (Equation 3.2) 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

$$V(Y) = \int_{\Omega^M} f^2(X)dX - f_0^2 \quad (3.3)$$

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 (3.2) as

$$V_{i_1 \ldots i_s} = \int_0^1 \ldots \int_0^1 f_{i_1 \ldots i_s}^2 \left(X_{i_1}, \ldots, X_{i_s}\right) dX_{i_1} \ldots dX_{i_s} \quad (3.4)$$

Where $1 \leq i_1 \leq \cdots \leq i_s \leq$ M and $s = 1, \ldots, M$. Assuming that the parameters are mutually orthogonal, Equation (3.5) results for the variance decomposition.

$$V(Y) = \sum_{i=1}^{M} V_i + \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} V_{ij} + \cdots + V_{1,\ldots,M} \quad (3.5)$$

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 ration of partial variance to the total variance, the Sobol' sensitivity indices:

First-order SI: $S_i = \frac{V_i}{V}$ (3.6)

Second order SI: $S_{ij} = \frac{V_{ij}}{V}$ (3.7)

Total SI: $S_{Ti} = S_i + \sum_{j \neq i} S_{ij} + \cdots$ (3.8)

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 (3.6) 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_{Ti} can be base on variance $V_{\sim i}$that results from the variation of all parameters, except X_i(Homma and Saltelli 1996).

$$S_{Ti} = 1 - \frac{V_{\sim i}}{V} \quad (3.9)$$

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 analyzing 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 (3.4) into an integral of the product of $f(X_{j_1}, \ldots., X_{j_{k-s'}}, X_{i_1}, \ldots. X_{i_s})$ and $f(X'_{j_1}, \ldots., \mathrm{X}'_{j_{k-s'}}, X_{i_1}, \ldots., X_{i_s})$. Because environmental models are mostly complex and nonlinear, it is almost impossible to calculate the variances using analytical integrals. The SI's can be calculated by performing Monte-Carlo simulations.

3.2.3. The Evaluation of the Sensitivity Analysis

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 Sobol' method. To analyze sensible parameters, maximum, minimum, and average air temperature 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 and 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 cross-sampling scheme creates a total of N * (2D + 2) parameter sets, where D is the number of input parameters, 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 $\pm$ 10% changes of air temperature

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 3.1 as defined by Lenhart *et al.* (2002). Normalized Sensitivity Indices for RWT model inputs parameters are listed in Table 3.4.

Table 3.1. Sensitivity Index Categories (Lenhart et al. 2002).

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

3.2.4. Ridge Regression

The method of ridge regression proposed by Hoerl & Kennard (1970). Ridge regression is a linear regression extension where the loss function is modified to minimize the model's complexity (Equation 3.11). 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 (3.10) represents the 2- norm or L2 norm.

$$\|w\|_2 = (w_1^2 + w_2^2 + \cdots + w_N^2)^{\frac{1}{2}} \tag{3.10}$$

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

Objective = RSS (Residual Sum of Squares) + λ * (sum of the square of coefficients)

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2 \tag{3.11}$$

3.2.5. K-nearest Neighbors (KNN) Regressor

KNN is a simple algorithm (Cover and 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 K nearest neighbors:

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (3.12)$$

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

3.2.6. Support Vector Regression (SVR)

Dibike *et al.* (2001) firstly applied the support vector machines (SVM) approach for accurate simulation of rainfall-runoff processes in hydrology. The Support Vector Machine (SVM) is a kernel function learning machine, which follows the structural risk principle (Vapnik et al. 1996). When the training data of $\{(x_1, y_1), \ldots\ldots (x_n, y_n)\}$ 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 variables, X will be mapped into a higher dimensional feature space using a nonlinear mapping function φ.

$$f(x; w) =< W, \varphi(x) > +b \quad (3.13)$$

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. SVR uses the ∈-insensitive error to measure the error between $f(x)$ and the observed values of y. Where ∈ is the hyper-parameter.

$$|f(x; w) - y|_{\in} = \begin{cases} 0, & if|f(x; w) - y| < \in \\ |f(x; w) - y| - \in, & otherwise, \end{cases} \quad (3.14)$$

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

$$F = \frac{C}{N} \sum_{i=1}^{n} |f(x_i, w) - y_i|_{\in} + \frac{1}{2}||w||^2 \quad (3.15)$$

Where C and ∈ 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:

$$f(X) = \sum_i K(X, X_i) + b \quad (3.16)$$

Based on previous studies (Dibike et al., 2001;Rehana, 2019), 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).

3.2.7. Random Forest (RF) Regressor

RF is designed to produce output by 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:

$$f(x) = f_0(x) + f_1(x) + f_2(x) + \cdots \tag{3.17}$$

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

3.2.8. Ensemble Kalman Filter (EnKF)

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

$$X_{t+1} = F(X_t) + W_t(N(0, Q)) \tag{3.18}$$

$$Y_t = H(X_t) + V_t(N(0,R)) \tag{3.19}$$

Where *F* is a nonlinear 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:

$$X_{t+1}^b = F(X_t^a) + W_t \tag{3.20}$$

$$P_{t+1}^b = E[(X_{t+1}^b - \bar{X}_{t+1}^b)(X_{t+1}^b - \bar{X}_{t+1}^b)^T] \tag{3.21}$$

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 (3.21), 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 -

$$X_{t+1}^a = X_{t+1}^b + K(D - HX_{t+1}^b) \tag{3.22}$$

$$P_{t+1}^a = E[(X_{t+1}^a - \bar{X}_{t+1}^a)(X_{t+1}^a - \bar{X}_{t+1}^a)^T] \tag{3.23}$$

$$K = P_{t+1}^b H^T (HP_{t+1}^b H^T + R)^{-1} \tag{3.24}$$

where $(D - HX_{t+1}^{b})$ is called the residual or measurement innovation. The Kalman gain K in Equation (3.24) 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 data set D and H as the observation operator.

3.2.9. Ensemble Kalman Filter (EnKF) Model Development

In this study, EnKF as a data assimilation 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 3.2 shows the ML and DA architectural flow diagram.

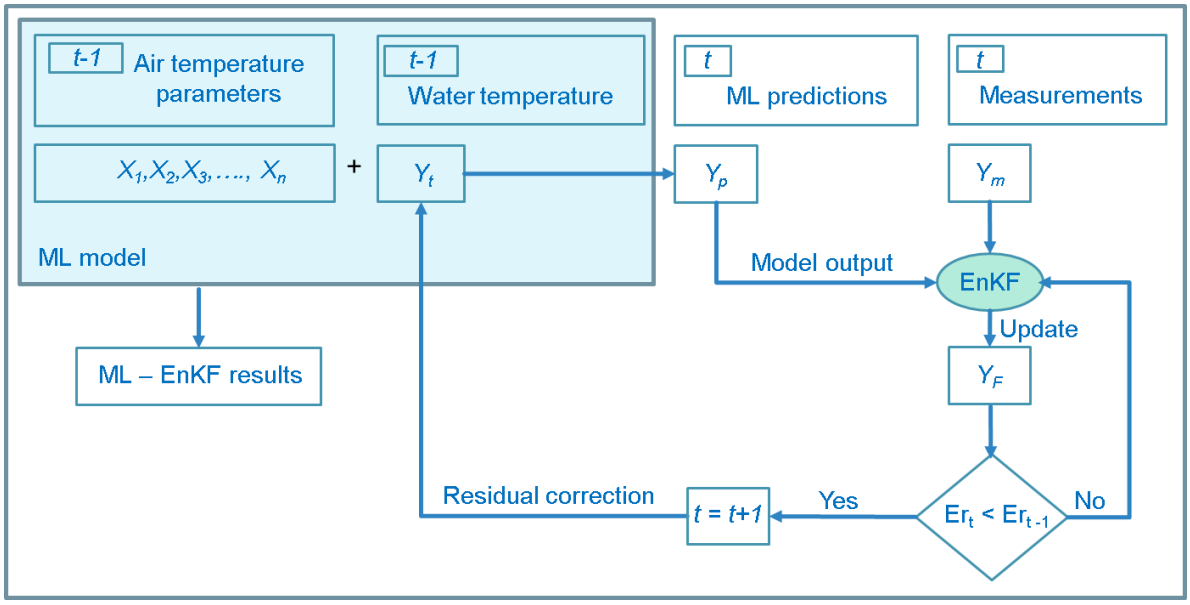


Figure 3.2. Architectural flow diagram of ML model and EnKF data assimilation method.

In Figure 3.2, Y_p is the results 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.

3.3. Study Area and Data

The river location considered for the RWQ modelling is Shimoga along the Tunga River, which confluences with Bhadra river to form the Tunga-Bhadra River, a major tributary of Krishna River basin, India (Figure 3.3). 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/sec. The variables used to fit a ML model are AT and RWT. Therefore, in this study, AT, and RWT are selected for the data analysis. The observed minimum, maximum and average air (water) temperature mean was noted as 19.66 ^oC, 29.74 ^oC, and 24.78 ^oC (27.54 ^oC) and standard deviation as 3.48 ^oC, 3.47 ^oC, and 2.77 ^oC (2.66 ^oC) 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 Shimoga city municipal effluent. The daily average RWT data and average, maximum and minimum air temperature data from 1st January 1989 to 1st January 2004 recorded at Shimoga station was 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 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.

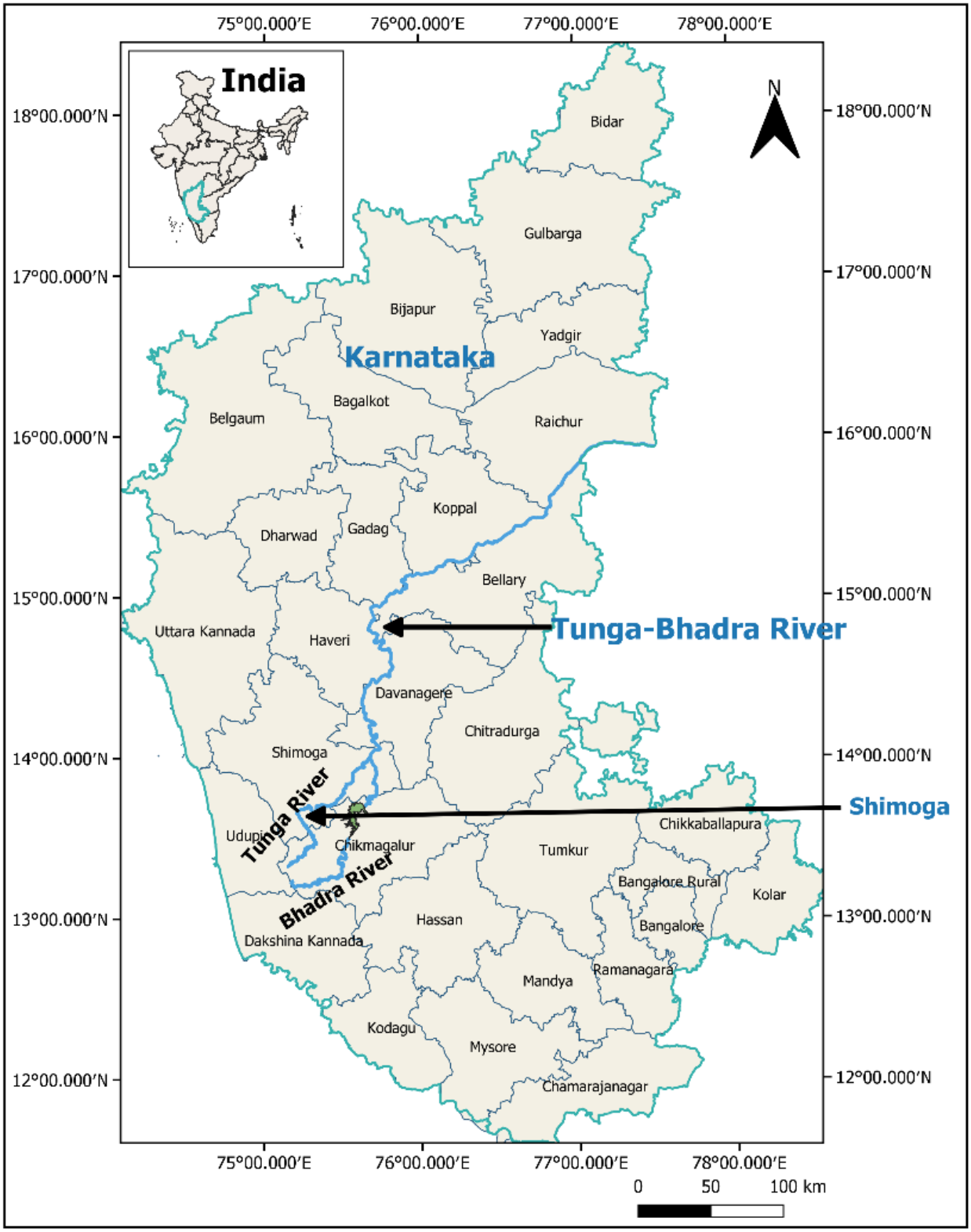


Figure 3.3. Location map of Tunga-Bhadra River and Shimoga station, India.

3.4. Model Evaluation

The accuracy of the applied ML models was evaluated using various good-ness of fit measures such as (Chadalawada and Babovic 2017): The coefficient of determination (R^2) (Equation 3.25), the mean squared error (MSE) (Equation 3.26), the root mean squared error (RMSE) (Equation 3.26), RMSE-observations standard deviation ratio (RSR) (Equation 3.27) (Moriasi et al. 2007), Nash-Sutcliffe efficiency (NSE) (Equation 3.28) (Nash and Sutcliffe 1970), the mean absolute error (MAE) (Equation 3.29), and Kling–Gupta efficiency (KGE) (Equation 3.30) (Kling et al. 2012). For assessment and comparison purposes, RSR and NSE can be ranked into the four classes found in Table 3.2 as defined by Moriasi et al. (2007).

$$R^2 = 1 - \frac{\sum(T_{w_{pred}} - T_{w_{obj}})^2}{\sum(T_{w_{pred}} - T_{w_{mean}})^2} \tag{3.25}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(T_{w_{pred}} - T_{w_{obj}})^2}{n}} \tag{3.26}$$

$$RSR = \frac{RMSE}{STDEV_{obj}} = \frac{\left[\sqrt{\sum_{i=1}^{n}(T_{w_{obj}} - T_{w_{pred}})^2}\right]}{\left[\sqrt{\sum_{i=1}^{n}(T_{w_{obj}} - T_{w_{mean}})^2}\right]} \tag{3.27}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(T_{w_{obj}} - T_{w_{pred}})^2}{\sum_{i=1}^{n}(T_{w_{obj}} - T_{w_{mean}})^2}\right] \tag{3.28}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}(T_{w_{pred}} - T_{w_{obj}}) \tag{3.29}$$

$$KGE = 1 - \sqrt{(1-r)^2 + (\gamma - 1)^2 + (\beta - 1)^2} \tag{3.30}$$

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

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

where $T_{w_{pred}}$ is the predicted daily river water temperature at time step *i* in °C; $T_{w_{obj}}$ is the observed daily river water temperature at time step i in °C; $T_{w_{mean}}$ is the average daily river water temperature at time step i in °C; $STDEV_{Obj}$ is the standard deviation of the observed daily river water temperature; 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; n is the number of data pairs in comparison.

Table 3.2. RSR and NSE performance ratings (Moriasi et al. 2007)

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

3.5. Results

The data used in this study consist of daily water temperature and corresponding daily minimum, maximum, and mean air temperature for the period from 1st January 1989 to 1st January 2004. The observed minimum, maximum and average air (water) temperature mean was noted as 19.66 °C, 29.74 °C, and 24.78 °C (27.54 °C) and standard deviation as 3.48 °C, 3.47 °C, 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 1st January 1989 to 1st January 2004. Spearman's correlation coefficients between RWT and maximum, minimum and average air temperatures were calculated. It is observed that RWT is highly significant with maximum, minimum, and average air temperatures (p-value < 0.001) (Table 3.3). Based on the statistical dependency measures, the maximum air temperature was positively correlated with daily RWT for the case study.

Table 3.3. Seasonal period Spearman's correlation coefficients between various air and water temperature variables

Season	RWT - maximum AT	RWT - minimum AT	RWT - average AT
Monsoon (June- September)	0.90	0.18	0.71
Post monsoon (October – November)	0.77	0.26	0.59
Winter (December - February)	0.84	0.20	0.62
Summer (March - May)	0.77	0.55	0.76
Annual	0.84	0.31	0.70

Furthermore, based on the SA (Table 3.4), it is observed that the maximum air temperature is highly sensitive, with a sensitivity index of 0.95 in the prediction of RWT compared to minimum and average air temperatures. The SA also supports the use of maximum air temperature as the most important independent variable to be considered in the prediction of RWT. To show the variability of maximum air temperature with RWT, the daily data from 1st January 1989 to 1st January 2004 has been compared, as shown in Figure 3.4. Most of the earlier studies considered average air temperature as the independent variable in RWT prediction. For example, Rehana and Mujumdar (2011) evaluated the average air temperature to predict the RWT for the Tunga-Bhadra river at 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 air temperature 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 air temperature and the average air temperature was compared with the linear regression model. The resulting R^2 value in RWT prediction was obtained as 0.58 and 0.83 with average and maximum air temperatures, respectively. Such improved performance of the RWT prediction model was convincing with an earlier study by Rehana and Mujumdar (2011), which used average air temperature as the predictor variable in RWT modelling.

Table 3.4. 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

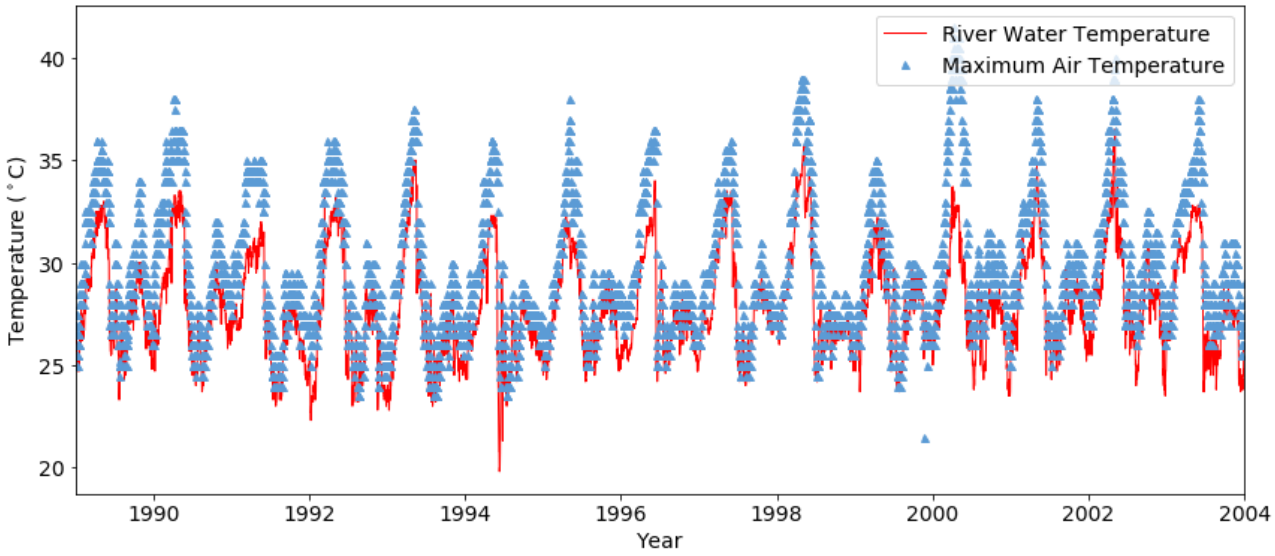


Figure 3.4. Time series of daily maximum air temperatures, water temperatures [1989-2004] of Tunga-Bhadra River at Shimoga station, India.

To understand the variability of air and water temperature changes for long term periods, the study estimated the linear trends of both variables (Figure 3.7(a), 3.7(b)). As can be observed, the long-term maximum air temperature and the RWT are varied during the period from 1989 to 2004 (Figure 3.5). The monthly seasonal dynamics of RWT and maximum air temperature based on 15 years averages at Shimoga station [1989-2004] are presented in Figure 3.6. It is shown that RWT and maximum air temperature give a strong seasonal pattern with larger values in summer and lower values in winter. As shown in Figure 3.7, the long-term air temperature and the water temperature increased during the period 1989-2004 at Shimoga station. Air temperature has been increased about 0.077 ^{o}C year^{-1}, while RWT increased about 0.062 ^{o}C year^{-1}. Such increasing trends of RWT has been investigated in many parts of the world. For example, the observed RWT has shown a growing trend of about 0.029 – 0.046 ^{o}C year^{-1} over China (Chen et al., 2016), over the USA of about 0.009–0.077 ^{o}C year^{-1} (Isaak et al., 2012; Rice & Jastram, 2015; van Vliet et al., 2013) and Europe as 0.006–0.18 ^{o}C year^{-1} (Albek and Albek 2009; Orr et al. 2015; Hardenbicker et al. 2017). Air temperature 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 river water temperatures and decrease of stream flows for the river of interest (e.g., Rehana and Mujumdar 2012; Rehana and Dhanya 2018).

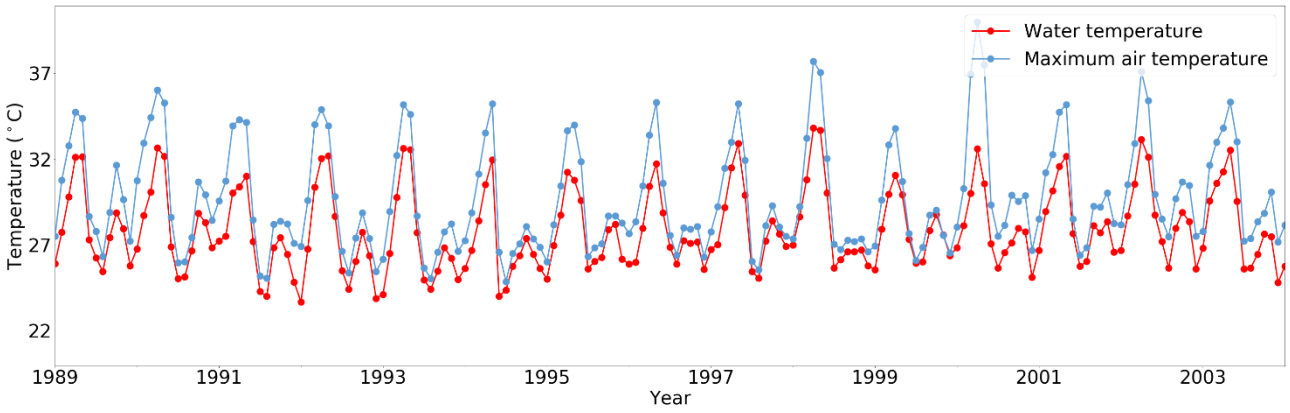


Figure 3.5. Time series of monthly mean maximum air temperature and water temperature for the period 1989-2004.

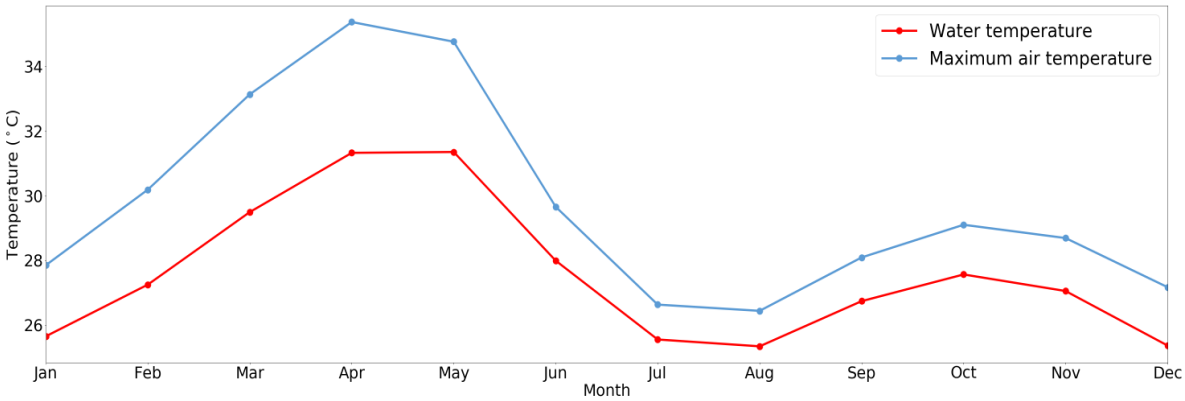


Figure 3.6. Monthly mean maximum air temperature and water temperature based on 15 years average at Shimoga station [1989-2004].

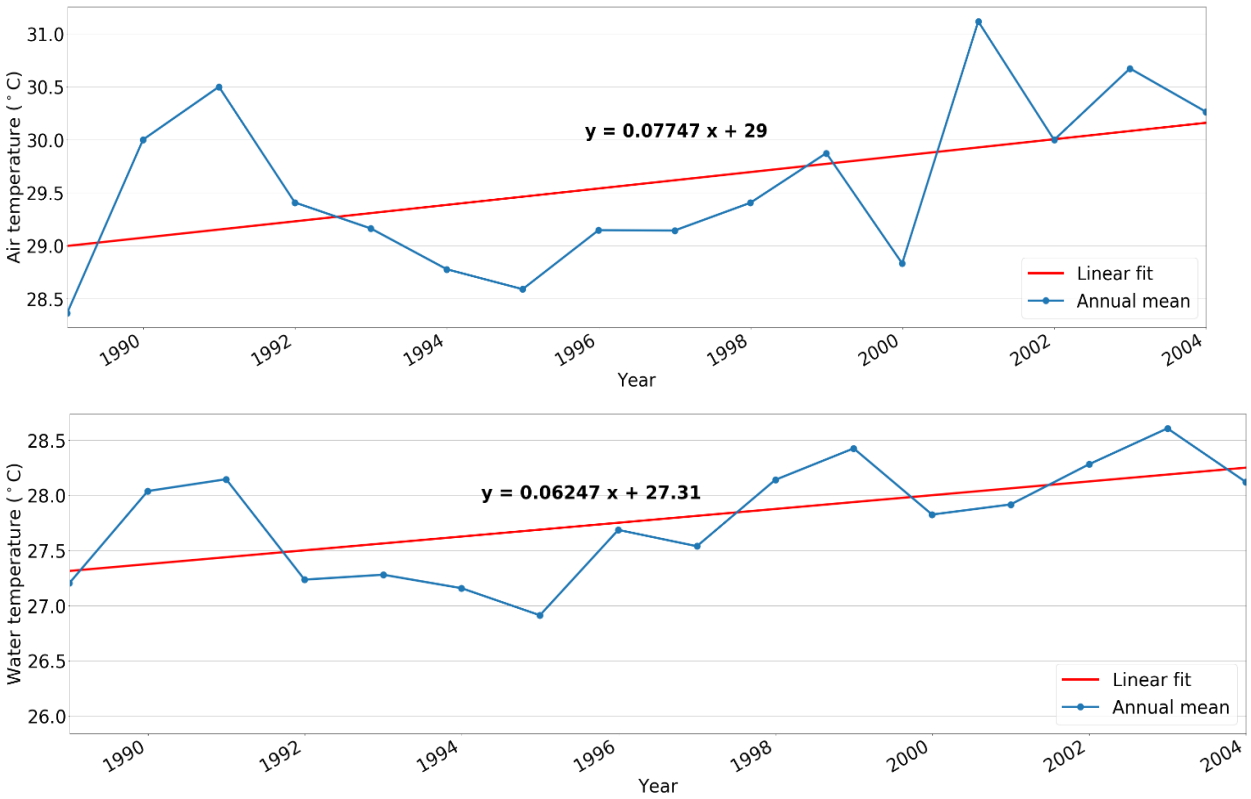


Figure 3.7. Time series of annual average (a) maximum air temperatures and (b) water temperatures for 1989-2004.

3.5.1. 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 3.1. 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 cross-validation (like k-fold) cannot be used since the adjacent data points are often highly dependent, so standard cross-validation will fail. To overcome these issues, the time-series splits cross-validation technique was used in the present study (Pedregosa et al. 2011; Scavuzzo et al. 2018). This cross-validation 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 cross-validation procedure provides an almost unbiased estimate of the true error (Varma and Simon 2006). The error on each split is

averaged in order to compute a robust estimate of model error, as shown in Figure 3.1. 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 goodnesses 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 3.8 show 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 °C and 28 °C with median RWT of 26 °C. According to Figure 3.8, 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 predicted value, in addition the lower and the upper quartile ranges predicted using these models were marginally varied compared to the observed data.

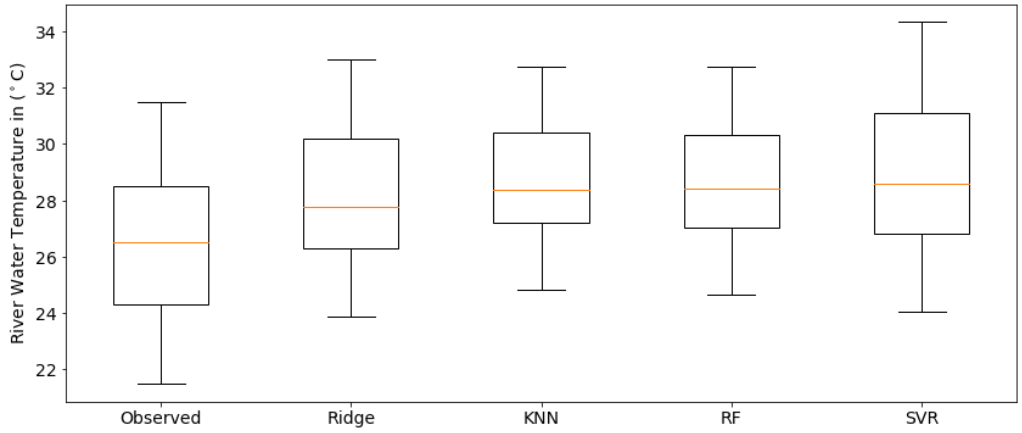


Figure 3.8. Box plots of observed and calculated river water temperature (°C) in the validation phase with the four ML models.

The performance of the Ridge, KNN, RF, and SVR models for daily data at Shimoga station is provided in Table 3.5 and Figure 3.9. Results showed that the seasonal variations of predicted RWT is almost synchronous and comparable with the observed values (Figure 3.9), but the Ridge model performed poor with overestimated values in high water temperature

period and performance statistics (R^2, MSE, RMSE RSR, NSE, and MAE) can be found in Table 3.5. From Table 3.5, 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, 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 3.2) with lower values of MSE and RMSE. The relationship between daily RWT and maximum AT at Shimoga station has a relatively strong correlated value for all four models (R^2 values). Based on RSR and NSE performance ratings (Moriasi et al. 2007) (Table 3.2), 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 3.5) for daily time scale.

Table 3.5. 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

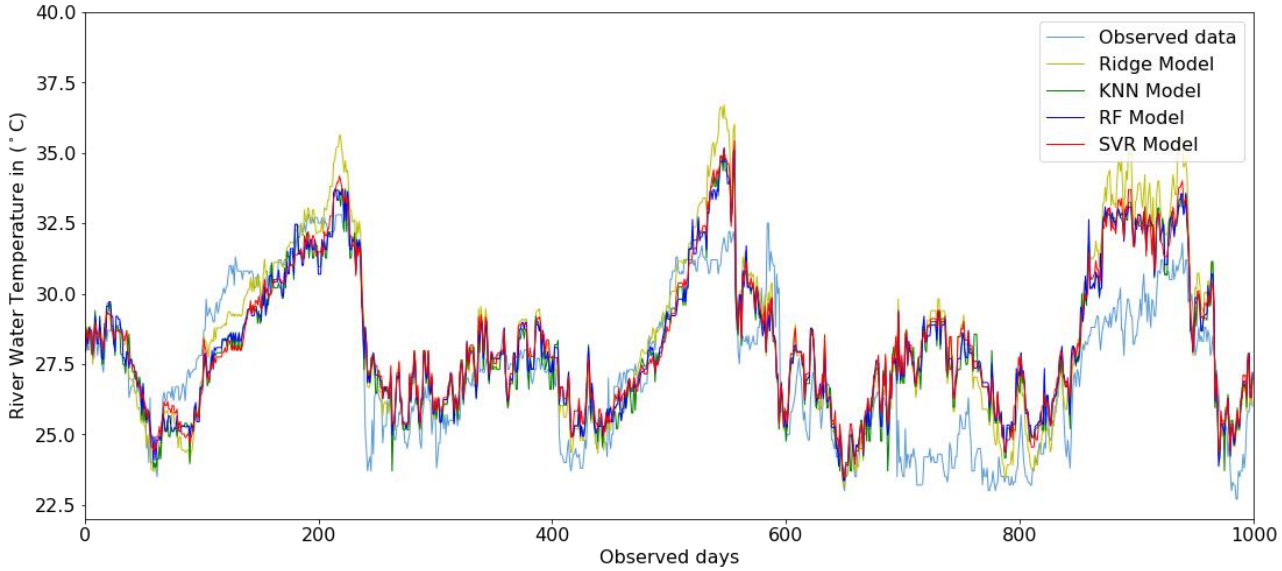


Figure 3.9. Comparison between the daily predicted values and observed values of river water temperature (°C) in the validation phase, with the four ML models.

A summary of the Ridge, KNN, RF and SVR model performances for monthly data was illustrated in Table 3.5 and Figure 3.10. ML results showed that the seasonal variations of predicted RWT is almost synchronous and comparable with the observed values (Figure 3.10), but the Ridge model performed poor with overestimated values in high water temperature period and performance statistics are given in Table 3.5. Compared to the four ML models, 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 to daily time scale (Table 3.5). 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 3.2).

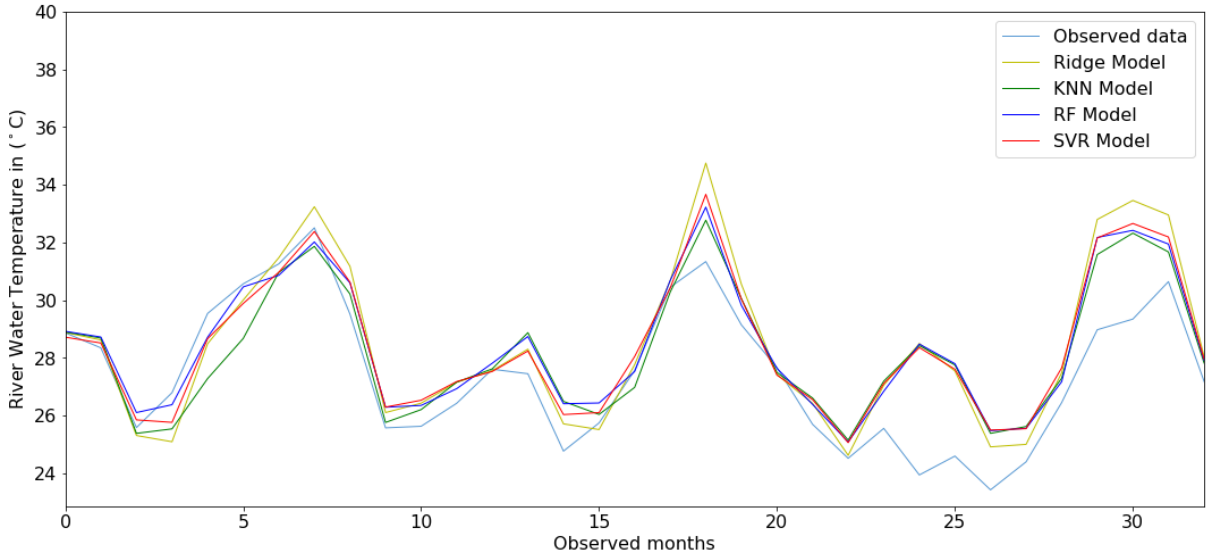
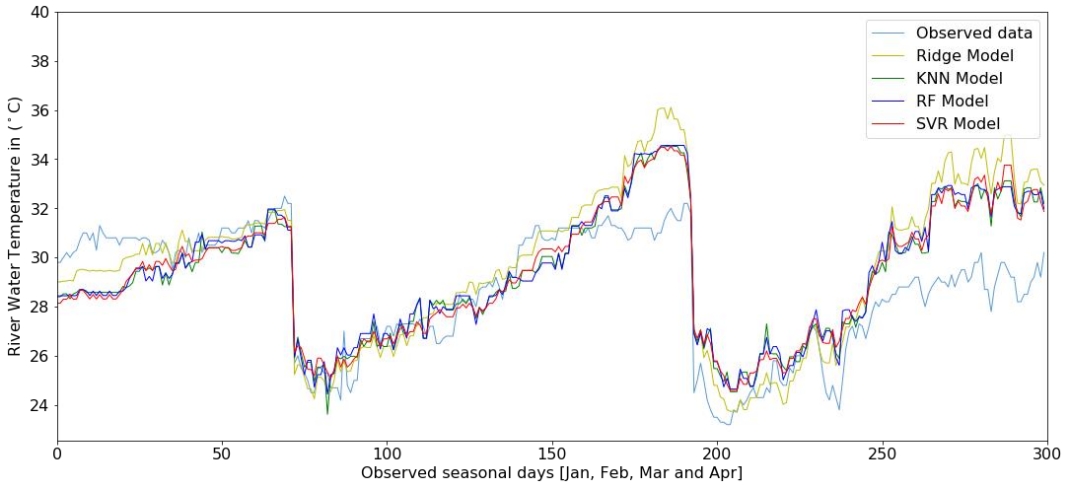


Figure 3.10. Comparison between the monthly predicted values and observed values of river water temperature (°C) in the validation phase, with the four ML models.

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. 2019a) is shown in Figure 3.11. Results showed that the seasonal variations of predicted RWT are almost in agreement with the observed values (Figure 3.11), but the Ridge model performed poorly with overestimated values in high water temperature periods and performance statistics are given in Table 3.5. From Table 3.5, 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 to the other two seasons, daily time scale and monthly time scale values.



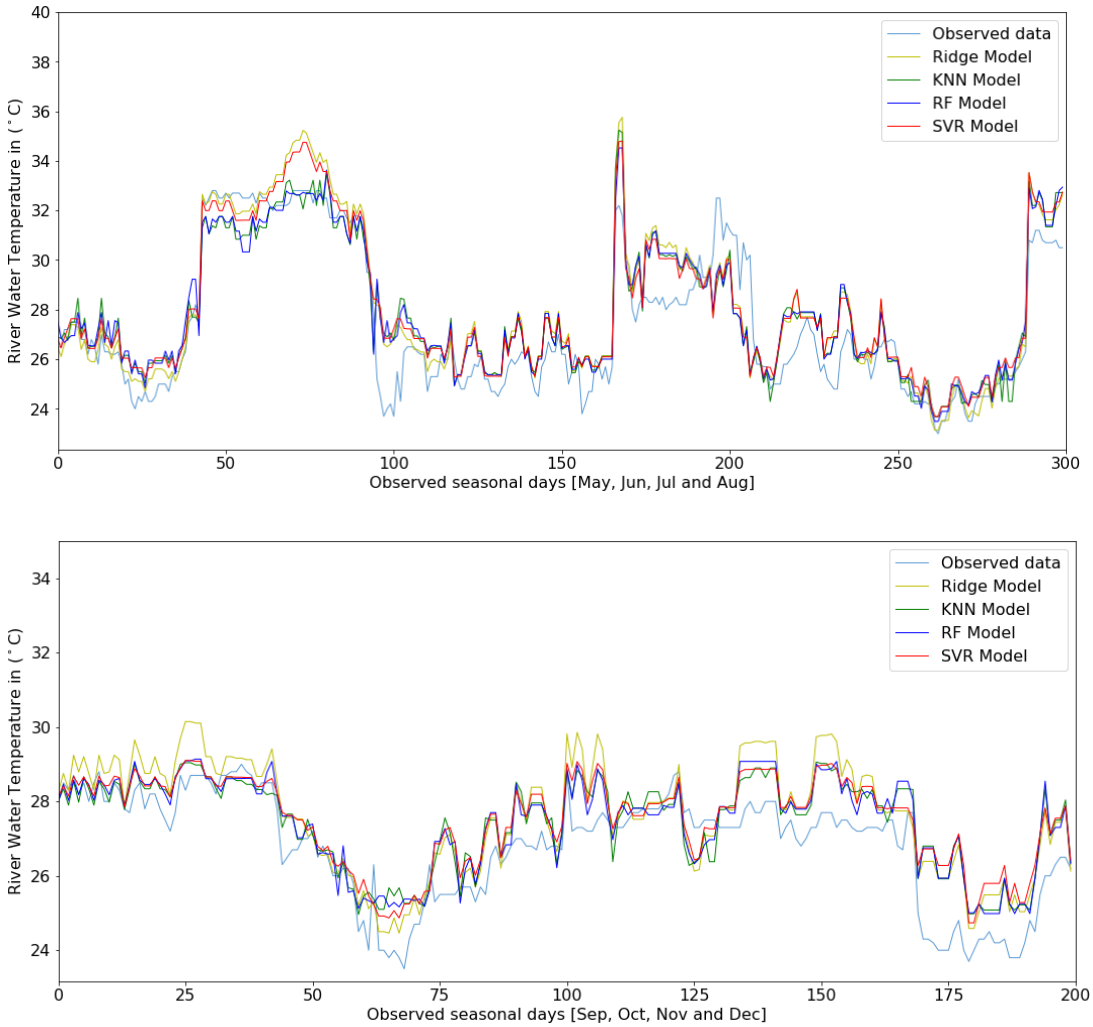


Figure 3.11. Comparison between the (a) Jan - Apr months (b) May - Aug months (c) Sep-Dec months seasonal predicted values and observed values of river water temperature (°C) in the validation phase, with the four ML models.

3.5.2. ML - EnKF Model Performance

In the next step in the prediction of RWT, the EnKF data assimilation technique is implemented to improve the efficiency of ML models in each simulation step. Table 3.6 shows the results of the ML-EnKF model at different simulation steps with the assimilated data. Table 3.6 shows that the blended data show the improved results from simulation-1 (1st January 2001 to 1st January 2002) to simulation-2 (1st January 2002 to 1st 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 3.6.

Table 3.6. Performances of different models with assimilated data in the prediction of RWT.

Data	Model	R^2	KGE	MSE	RMSE	RSR	NSE	MAE
Simulation -1 [1/1/2001 to 1/1/2002]	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 [1/1/2002 to 1/1/2003]	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

3.6. Discussion

This chapter presents new intuitions on the assessment of performance of a suite of ML models for RWQ variables prediction, such as RWT for the Tunga-Bhadra River, India, with the aid of the minimum and maximum air temperature at daily, monthly and seasonal time scales. The relationship between daily RWT and maximum AT at 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 3.5) 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 and Dadaser-Celik (2018) (2.10–2.64). Based on RSR and NSE performance ratings (Moriasi et al. 2007) (Table 3.2), 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 3.5) 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 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 SVR with maximum AT (NSE: 0.84 and RMSE: 0.99) as an independent variable compared to average AT (NSE: 0.61 and RMSE:1.69) (Rehana, 2019) for the same case study at daily

time scale. Table 3.5 shows that the four models constructed in this study may learn the RWT variation rules from the historical data and reproduce the seasonal dynamics of RWT. It can be noted that the improved ML model accuracy with monthly data compared to daily data is due to taking the daily values into monthly averages (i.e., averaging all the daily values, the errors will be distributed and get better results) and less data variability involved in the prediction. This case study demonstrates that integrating the scientific knowledge into ML tools promises to improve many important environmental variables predictions.

It can also be concluded that the ML-EnKF model can do a better job with assimilated data in RWQ variables 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 3.6. As section 3.2.9 states, the ML-EnKF model is designed to improve the ML model performance by a combination of both ML models and a data-assimilation approach to enhance the predicted values based on the measurement data. 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.

The present study demonstrated how a data-driven modelling framework could be scaled up and used for the prediction of RWQ variables. The data assimilation methods can also combine with ML models to improve the predicted values based on the measurement data. Overall data-driven modelling framework presented in the chapter indicated that all ML models were proven to be effective in RWQ variables 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 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 RWQ

variables prediction can always be implemented with newly updated data as demonstrated in the present study, which can be extended to other stations and other variables based on data availability. In this study, all the models proposed typically provide a single-point prediction, neglecting the inherent variability present in the data and the model itself. It was observed that inherent uncertainties from each of the ML models can accumulate and affect the final performance measures. These uncertainties can originate from different sources, such as noise, covariates considered, temporal discontinuity present in the original water quality sampled observations to the model parameters, type of ML algorithm used to predict RWT, and non-stationarity assumptions related to forecasting model parameters (Beven 2016). To estimate such model uncertainties originating from various ML models, techniques like bootstrap aggregation (bagging) and Monte Carlo methods can be adopted. Developing an ensemble model using a robust weighted voting regressor (VR) method to quantify forecasting uncertainty and to improve the model performance can be further research (Rajesh et al. 2022). By employing a lag-1 time series model as the null model, comparing skill scores of different ML models with various covariate sets are presented in the Chapter 4 by considering the seven majorly polluted catchments of India (CPCB 2015; National River Conservation Directorate (NRCD) 2018).

3.7. Chapter Summary

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 RWQ variables prediction under limited data input variables. To demonstrate, RWT was selected as water quality variable and predicted with the aid of the minimum and maximum AT at daily, monthly and seasonal time scales for the Tunga-Bhadra River, India. In this chapter, an attempt has been made to identify the most sensitive AT variable (average, maximum and minimum) using Sobol' sensitivity analysis method described in Section 3.2.2, which can serve as an input variable in the prediction of RWT. Furthermore, each model's configurable variable is optimized, and the performances of various ML models are analyzed to test the applicability of the data-driven models in the RWT being investigated. Further, the EnKF algorithm was described in Section 3.2.9, which is integrated with ML approaches to improve the predicted values based on the measurement data. Finally, this chapter was concluded with the following conclusions:

1. The results indicated that the maximum AT was the most important variable in the prediction of 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 method for RWQ variables prediction modelling.
2. The study revealed that ML model performance coefficients are improved in monthly data compared to the daily time scale. The seasonal time scale RWT prediction models also performed poorly compared to 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.
3. The SVR has been noted as the most robust ML model to predict RWT. The SVR model is a strong choice for smaller datasets and is less sensitive to outliers in the data compared to some other models. The SVR is generally less computationally expensive than the ML models. However, for highly complex relationships or very large datasets, ANN or DL architectures might be more suitable, considering their computational resources and potential for even higher accuracy.
4. The ML-EnKF model update of the prediction data with the observed data using the data assimilation method shows a better result. If the simulation steps continue, the ML-EnKF model is improved, and the simulation results are significantly improved.

Overall, this chapter demonstrated the prediction of RWQ variables using classical ML algorithms by integrating the sensitivity analysis and data assimilation techniques to improve the performance under limited data input variables. The proposed methods, demonstrated methodologies, frameworks in this chapter are generic, and can be implementable for any given RWQ variable. Further research into the robust and hybrid ML approaches is required and were presented in the next chapter to predict the RWQ variables under sparse, non-stationary data scenarios, as an accurate simulation of RWQ variables, which plays an important role in water quality management under data sparsity uncertainties.

Chapter 4
PREDICTION OF RIVER WATER QUALITY VARIABLES WITH SPARSE DATA USING HYBRID DEEP LEARNING METHODS

4.1. Introduction

The ML models presented in Chapter 3 is able to address limited data variables scenarios in predicting RWQ variables by integrating sensitivity analysis using minimal data inputs (such as AT). The present chapter addresses another sort of data uncertainty, namely the lack of availability of long-time series data to capture interannual variability and consistent water quality measurement datasets in RWQ modeling. Generally, RWQ data availability is at monthly scales and is burdened with a large number of missing values with limited durations. In this context, the selection of appropriate model inputs, development of models under limited data, processing of non-stationary data, seasonality scenarios, and relevant lags of variables have not been intensively investigated in the literature, especially in the case of estimation of RWQ variables.

To simulate RWQ variables, process-based models (e.g., Delft3D model, Soil and Water Assessment Tool (SWAT) model, etc.) are commonly used but such approaches require large amounts of site-specific detailed data at daily time scales, including stream geometry and meteorological and hydraulic properties of the river (Piccolroaz et al. 2016). The use of data-driven algorithms, such as DL models using minimum data inputs (such as AT), can be robust in addressing data sparsity in simulating RWQ models. However, the robustness of any DL-based forecasting algorithms, i.e., ANN or RNN, depends on the extensive input data to learn the dynamics of complex systems (Read et al. 2019). At the same time, RWQ datasets exhibit complex interrelationships among RWQ variables, serial dependence, data-limited context, stochastic nature, and seasonality (Rabi et al. 2015; Zhu et al. 2019f). Stochastic modeling approaches have been well developed in hydrology and hydro climatology to address these scenarios (Raseman et al. 2020). In literature, multiple simulation models have been implemented, and the most widely used approach is the k-NN bootstrap resampling simulation technique which is well suited to generate synthetic data (Lall and Sharma 1996; Rajagopalan and Lall 1999). The present study used the k-NN

algorithm to generate RWQ data, as high temporal resolution datasets are rarely available for Indian case studies.

In this chapter, developed hybrid models for RWQ variables predictions using Long Short-Term Memory (LSTM), integrated with (i) k-nearest neighbor (k-NN) bootstrap resampling algorithm (kNN-LSTM) to address the data-limitations, (ii) discrete wavelet transform (WT) approach (WT-LSTM) to address the time-frequency localized features. To build the kNN-LSTM model, the k-NN algorithm was adopted, which is described by Raseman *et al.* (2020) to create synthetic time-series data with realistic scenarios for predicting the RWQ variables. In water quality, RWT is a key element that affects the health of a freshwater ecosystem. Changes in AT can affect RWT, the primary variable that influences water quality. Therefore, to demonstrate the proposed hybrid models, the RWT has been considered as the water quality variable for prediction by using AT and lag variables as predictors. To study the pertinence of the k-NN algorithm and wavelets, the seven major Indian catchments were utilized with monthly datasets of air and water temperatures to simulate time-series data. One of the reasons for choosing the monthly time scale is most of the Central Pollution Control Board (CPCB), and Central Water Commission (CWC) samplings are at monthly time scales for RWQ monitoring (Central Water Commission 2018; CPCB 2020). Furthermore, the study compared the proposed WT-LSTM and kNN-LSTM monthly models with standalone LSTM, air2stream models (Toffolon and Piccolroaz 2015; Piccolroaz et al. 2016).

Furthermore, the study evaluated the effect of climate change on RWTs using Representative Concentration Pathways (RCP) scenarios 4.5 and 8.5 dataset outputs downscaled from the National Aeronautics Space Administration (NASA) Earth Exchange Global Daily Downscaled Projections (NEX-GDDP) dataset. Also, validated the causal linkages between the time series of data using Granger Causality Analysis (GCA) test to check if the results would be improved by the addition of lagged variables.

In summary, the objectives of this chapter are to (i) coupling the k-NN bootstrap resampling technique with the LSTM model (kNN-LSTM) to overcome the limited data scenarios of river water quality data, (ii) coupling the WT with the LSTM model (WT-LSTM) to yield better performance by overcoming both processing of non-stationary and the noisiness in data for RWQ variables prediction at monthly scale, (iii) compare the

performance results of WT-LSTM and kNN-LSTM models with LSTM, 3-parameter version air2stream in the prediction of RWQ variables when applied on seven major river systems of India, (iv) calculating the impacts of climate change on the rivers thermal processes in India and possible variability in RWT by using the kNN-LSTM model forced with an ensemble of 21 GCMs using the NEX-GDDP dataset under RCP scenarios 4.5 and 8.5 dataset output.

4.2. Model Development

The proposed model combines pre-processing data methods such as feature engineering, handling missing data, WT (see Sect. 4.2.1 for further information about WT), and an ensemble of simulated data by using k-NN bootstrap resampling algorithm and LSTM model with sufficient tests of performance measures of models at monthly timescale (Figure 4.1). In this study, WT was implemented to de-noise the historical data, and the k-NN algorithm was implemented to simulate the data from historical data for better performance of the LSTM models. In Figure 4.1a, yellow, blue, and green colored arrows represent the data workflows for LSTM, WT-LSTM, and kNN-LSTM models, respectively, for monthly water temperatures prediction at the seven catchment sites of India. The detailed flow charts of WT-LSTM and kNN-LSTM models in Figures 4.1b, and 4.1c. For comparison, the 3-parameter version air2stream model was used as a benchmark model with the original time series of ATs as predictor variables. For future RWT projections, RCP scenarios 4.5 and 8.5 down-scaled projections of AT data were fed into the kNN-LSTM monthly prediction model.

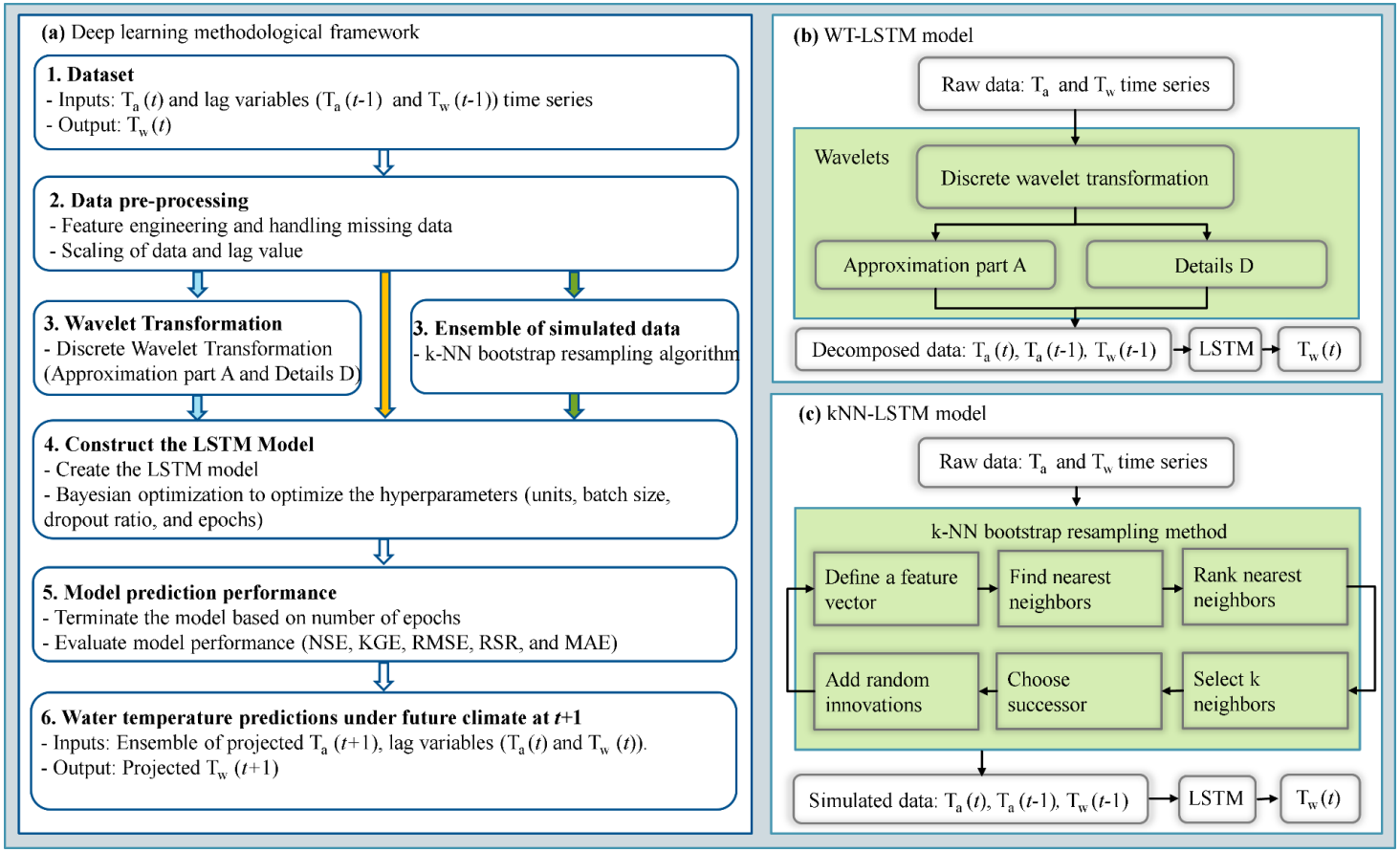


Figure 4.1. Overview diagram (a) the deep learning methodological framework of the proposed river water temperature forecasting models. Yellow, blue, and green colored arrows represent the data workflows for LSTM, WT-LSTM, and kNN-LSTM models, respectively, (b) the detailed flow diagram showing the steps of coupling Wavelet Transform (WT) and Long Short-Term Memory (LSTM) model (WT-LSTM), (c) the detailed flow diagram showing the steps of coupling k-NN bootstrap resampling algorithm, and Long Short-Term Memory (LSTM) model (kNN-LSTM). T_a is the average air temperature, and T_w is the water temperature.

4.2.1. Wavelet Transform (WT)

Wavelets transform is used in many scientific disciplines as a data pre-processing approach. Understanding the variability of hydrological processes is an essential and important scientific topic in hydrology studies, but it is also a difficult problem due to the complex stochastic nature of hydrological processes (Shoaib et al. 2014). Most of the hydrologic data are non-stationary in nature (Milly et al. 2008). Such a non-stationary time series consists of various components (e.g., seasonal, trend and abrupt components) occurring for varying durations which can be determined through time segmentation. Mathematically, time series (y_t) is stationary if, for all *t* (Huang et al., 1998).

$$E(y_t) = E[(y_{t-1})] = \mu \tag{4.1}$$

$$Var(y_t) = \gamma_0 < \infty \tag{4.2}$$

$$Cov(y_t, y_{t-k}) = y_t \tag{4.3}$$

where *E(.)* is the expected value defined as the ensemble average of the quantity, and *Var(.)* and *Cov(.)* are, respectively, the variance and the covariance functions. If these constraint conditions are violated, the time series would exhibit non-stationary characteristics, which is a major challenge for several fields (e.g., remote sensing, engineering, and hydrology). For this reason, several approaches are developed to analyze the non-stationary characteristics. However, the observed hydrologic series are usually complex and show non-stationary and multi-temporal scale characteristics in daily, monthly, annual, inter-annual, decadal and larger scales (Labat 2005; Sang et al. 2009, 2011). To deal with this problem, the WT has been frequently applied in hydrology as it has the superiority of addressing non-stationary variability of hydrological processes and identifying the significant level shifts, etc., in time-series data (Sang et al. 2015; Rahman et al. 2020).

Compared with the Fourier Transform (FT), WT has the advantage of simultaneously obtaining information on the time, location, and frequency of a signal, while the FT can only provide the frequency information of a signal (Daubechies 1990). WT has been widely used

to reveal information (signal) both over time and on a domain scale (frequency). It is thus a more powerful transformation for time-frequency analysis. To our knowledge, the present applications such as forecasting cannot be made using WT alone. Hybrid approaches, which combine DL with WT, currently offer the highest performance for time series analysis due to their complementarity (Wang et al., 2018). As a result, several researchers used the coupling of WT with ANN (Tan and Perkowski 2015; Komasi et al. 2018; Graf et al. 2019). WT has the benefit over the Fourier Transform (FT) in that it can collect information both over frequency and time, whereas the FT can only give the frequency data (Daubechies 1990). WTs are mainly parted into Continuous wavelet transforms (CWT) and Discrete wavelet transforms (DWT). Although the CWT can identify the complex characteristics of a time series under multitemporal scale, there is much repeated information (called data redundancy) in the continuous wavelet results of a time series, and the results get more affected by the boundary effects. These factors would affect the stability of the wavelet modeling structure and consequently increase uncertainty (Sang et al. 2016). In this work, DWT was utilised to decompose the original raw ATs and RWTs data into approximation part (A) and details (D) (Figure 4.1b and 4.7) since prior researchers have demonstrated that DWT has advanced efficacy and is easier to use (Montalvo and García-Berrocal 2015). Then, the transformed forms of the AT and RWT data were used as model inputs in the LSTM training and testing.

The most important considerations in DWT are picking the suitable wavelet decomposition level (L) and mother wavelet. There are multiple mother wavelets, i.e., Daubechies (Db), Haar, discrete Meyer, Symlet, etc. In this study, a Db wavelet was employed. The L can be calculated by the below approach (Nourani et al. 2009):

$$L = int[\log(N)] \tag{4.4}$$

Where N represents the length of data.

4.2.2. Long Short-term Memory (LSTM)

The LSTM is a variant of RNN where the previous step's output is fed as input to the present step (Hochreiter and Schmidhuber 1997). The RNN is preferred over ANN as it can represent time series data, allowing each sample to be presumed to rely on the prior one. To overcome gradient vanishing and exploding problems, RNNs can be improved using the gated RNN

architectures such as LSTM (Hochreiter and Schmidhuber 1997) and GRU (Cho et al. 2014). The current work considered the most widely known RNN architecture of LSTM to predict the RWT due to the superiority of using backpropagation through time and overcome the vanishing gradient problem and capable of learning long-term dependencies (Hochreiter and Schmidhuber 1997; Hochreiter 1998; Hochreiter et al. 2001; Greff et al. 2017). The LSTM consists of different memory blocks called cells (Figure 4.2). Each memory cell has an input gate, an output gate, and an internal state that feeds back into itself unaffected over time steps, which learns when it's time to forget about prior hidden states (h^{t-1}), when to update hidden states given new data and be used to learn complex temporal sequences. The LSTM architecture avoids the problem of vanishing gradients by introducing error gating (Hochreiter 1998). The following are the LSTM equations (Equations 4.5-4.10) (Olah 2015):

$$f_t = \sigma(x_t W_{xf} + h_{t-1} W_{hf} + b_f) \quad (4.5)$$

$$i_t = \sigma(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \quad (4.6)$$

$$o_t = \sigma(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \quad (4.7)$$

$$\hat{s}_t = tanh\left(x_t W_{xg} + h_{t-1} W_{hg} + b_g\right) \quad (4.8)$$

$$s_t = f_t \odot s_{t-1} + i_t \odot \hat{s}_t \quad (4.9)$$

$$h_t = o_t \odot tanh(s_t) \quad (4.10)$$

where h is the hidden units, $f_t,\ i_t, and\ o_t$ denotes the forget, input, output gates, respectively, $\hat{s}_t$ denotes candidate of cell state, s_t, h_t denotes cell, hidden states, respectively, and W_x , W_h, b are trainable weights. σ is the sigmoid function, $\odot$ is the element-wise multiplication, x_t is a vector of d input features, and $tanh$ is the hyperbolic tangent activation function.

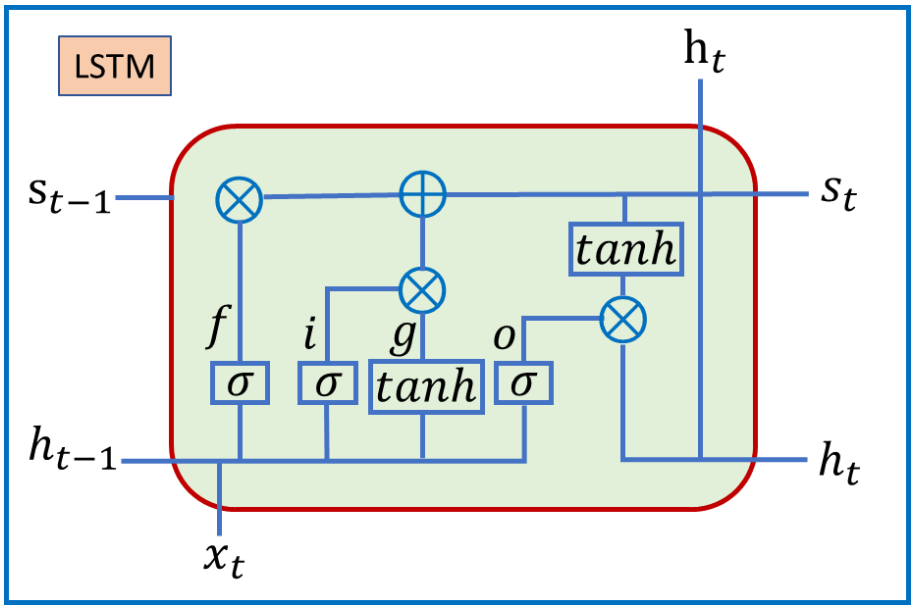


Figure 4.2. Overview diagram of Long short-term memory neural network (LSTM). Where f, i, and o denotes the forget gate, input gate, and output gate, h_t denotes hidden state, s_t denotes cell state, σ is the sigmoid function, *tanh* is the hyperbolic tangent activation function.

4.2.3. k-NN Bootstrap Resampling Algorithm

The k-NN bootstrap resampling algorithm was used for generating simulated time-series data from historical data based on Sharif and Burn (2007) and Raseman *et al.* (2020). The 1st month of simulated data is the user-defined month (e.g., January) and randomly chosen year. The steps for the subsequent months are listed below:

- **Feature vector:** Define a X_t ‘‘feature vector’’, dimension $d = qL$, where q and L are the variables and numbers of lags considered in the model, respectively. In this study, a lag-1 dependence or $L = 1$ was discovered and $q = 2$ has been used, i.e., AT and RWT variables to simulate the values.
- **Calculate Mahalanobis distance and determine the nearby neighbors:** For the present time step, i, the feature vector, X_i, is created. To identify which neighbors are nearby to X_i, the Mahalanobis distance (Mahalanobis 1936) (Equation 4.11) is used (Sharma and O’Neill 2002; Yates et al. 2003):

$$d_i = \sqrt{(X_t - X_i)^T C^{-1} (X_t - X_i)} \quad (4.11)$$

where C is a q x q matrix which defines the covariance between X_i and X_t, d_i represents the N-dimensional distance vector. N denotes the total number of years.

- **Rank nearby neighbors and choose k neighbors**: The nearby k neighbors are then selected from the ascending order list, and the successor is selected from among them. In this study, used $k = \sqrt{N}$, which is recommended by Lall and Sharma (1996).
- **Select successor**: To choose neighbors among the k nearby neighbors in a probabilistic manner, X_t^{kNN} (2 x k matrix, a subset of X_t), the discrete kernel K is used to define a weighting function, defined in Lall and Sharma (1996).
- **Random innovations to successor:** This step of the algorithm is divided into (1) create modified successors, (2) bound variables, and (3) check the bounds, and if bounded, repeat steps 1 and 2 until they produce a non-negative value.

After simulating the modified successor, $\tilde{x}_i$, it calculates the new current timestep's new value. The above steps are repeated for the next months till the simulated values match the historical record in length. Detailed information on the k-NN algorithm may be found in Raseman *et al.* (2020). A detailed flow diagram showing the steps of coupling the k-NN bootstrap resampling algorithm and LSTM model (kNN-LSTM) is provided in Figure 4.1c.

4.2.4. Air2stream

The air2stream is a hybrid model which combines a physically based structure with a stochastic parameter calibration used for RWT prediction developed with a limited computational complexity by Toffolon & Piccolroaz (2015). The air2stream model is based on a single ordinary differential equation linearly dependent on discharge, air, and water temperature. The 8-parameter version equation (Equation 4.12) for the air2stream model is given as follows:

$$\frac{dT_w}{dt} = \frac{1}{\theta^{a_4}} \left[a_1 + a_2 T_a - a_3 T_w + \theta \left(a_5 + a_6 \cos\left(2\pi \left(\frac{t}{t_y} - a_7 \right) \right) - a_8 T_w \right) \right] \quad (4.12)$$

where θ is the dimensionless flow discharge; t is the time; t_y is the number of time steps

over a year; $a_1 - a_8$ are model parameters; T_w is the water temperature (°C); T_a is the air temperature (°C).

A further form of the model can be found by simplifying Equation (4.12), also imposing θ = 1, and putting the constant and proportional terms together to T_w, a 5-parameter model is given as follows (Toffolon and Piccolroaz 2015):

$$\frac{dT_w}{dt} = a_1 + a_2 T_a - a_3 T_w + a_6 \cos\left(2\pi\left(\frac{t}{t_y} - a_7\right)\right) \quad (4.13)$$

A further form of the model can be found by ignoring the 2nd term from the Equation (4.12) and imposing that the θ impact may be approximated with a constant value, a 3-parameter model form is given as follows (Toffolon and Piccolroaz 2015):

$$\frac{dT_w}{dt} = a_1 + a_2 T_a - a_3 T_w \quad (4.14)$$

The original air2stream concept relies on a 20-year old inefficient method named Particle Swarm Optimization with inertia weight. This work used the updated 3-parameter air2stream version, which uses the covariance bimodal differential evolution (CoBiDE) optimization method developed by Piotrowski and Napiorkowski (2018). The source code of the original air2stream can be obtained from https://github.com/spiccolroaz/air2stream. On the Journal of Hydrology web page, the air2stream model's MATLAB code and the selected calibration process (CoBiDE) are available as Supplementary Material (Piotrowski and Napiorkowski 2018).

4.2.5. Climate Change Scenarios

Using the historical and simulated time-series data of AT and RWTs, the k-NN bootstrap resampling algorithm has been developed. The kNN-LSTM model was trained with the current AT and lag variables (AT [t-1] and RWT [t-1]) assessed from partial autocorrelation plots (Figure 4.6) and then forced with bias-corrected monthly outputs of NEX-GDDP downscaled projections of AT data from RCP scenarios 4.5 and 8.5 to produce predictions of monthly RWT for the 21st century. The first month's water temperature is calculated based on the catchment mean from the historical record which serves as the input for the next month's

prediction. The prediction of subsequent months proceeds as follows:

$$T_{t+1}^{w} = f(T_{t+1}^{a},\ T_{t}^{a},\ T_{t}^{w}) \quad (4.15)$$

Where T_{t+1}^{w} is the future RWT prediction at time t+1 month; f is a non-linear function which is generated by the kNN-LSTM monthly model; $T_{t+1}^{a}$is the future AT at time t+1 month; T_{t}^{a} is future AT at time t month; $T_{t}^{w}$is the predicted water temperature value at time t month.

For the analyses, the catchment's observed data periods were focused (Table 4.1) and future periods 2021-2050 and 2071–2100, followed the 30 years for a climatological standard normal (WMO 1989).

4.2.6. Granger Causality

The notion of Granger causality was introduced by Granger (1969) and soon found application in many fields (e.g., economics, and hydrology) because of its simplicity and robustness. Granger causality relates to a situation where the data concerning past values of one time series provide important information helping to predict values of another series not included in the information about its past values (Graf 2018). Granger causality assesses whether one variable at time *t*-lag causes another variable at time *t*. In this study, first order Granger Causality Analysis (GCA) was used to validate the time sequence of the causal linkages between the time series of data (i.e., whether cause precedes the effect relations for "water–air" and "air–water" directions of influence at a monthly scale) when applied on seven major river systems of India. Granger causality tries to answer the question of how much of the current variable can be explained by the past values of different values and whether adding lagged values can improve such an explanation (Kirchgässner et al. 2013). In the current research question is the time series AT Granger-causes time series RWT? Are the patterns in AT are approximately repeated in RWT after some time lag? In another words, the ability to predict the future values of a RWT time series using prior values of AT time series needs to be validated. One main assumption to test Granger causality is the stationarity of the time series. Granger causality between two stationary time series (*X* and *Y*) is formulated as follows:

$$Y_t = \sum_{j=1}^{m} a_t Y_{t-j} + \sum_{j=1}^{m} b_j X_{t-j} + \varepsilon_t \quad (4.16)$$

where *a* and *b* are coefficient (b ≠ 0) and ϵ is white noise. In this case, variable *X* Granger causes variable *Y*.

Granger causality between two stationary time series (*AT* and *RWT*) is formulated as follows:

$$RWT_t = \sum_{j=1}^{m} a_t RWT_{t-j} + \sum_{j=1}^{m} b_j AT_{t-j} + \varepsilon_t \quad (4.17)$$

In this case, variable *AT* Granger causes variable *RWT*.

4.3. Study Area and Data Setting

4.3.1. Study Area

For this study, seven majorly polluted catchments of India (CPCB 2015; National River Conservation Directorate (NRCD) 2018) were selected to predict RWQ variables with various physiographic features and studied the impact of climate change on water quality. To assess the pertinence and efficacy of the presented models, selected study sites with diverse properties. The seven river gauging stations are situated in India and are shown in Figure 4.3, and their main characteristics with training and testing periods are outlined in Table 4.1. Two data sources were used to compile the models, with one being global, one regional. The data from Global Freshwater Quality Database (GEMSTAT) were used for Narmada, Cauvery, Sabarmati, and Godavari catchments. The data from CWC, India were used for Tunga-Bhadra, Musi, and Ganga catchments. The Global Freshwater Quality Database GEMStat (Färber et al. 2018) is hosted by the International Centre for Water Resources and Global Change (ICWRGC) and provides inland water quality data within the framework of the GEMS/Water Programme of the United Nations Environment Programme (UNEP). Approximately 500 water quality parameters were available in the global GEMSTAT database, out of which water temperature was used in this study for Narmada, Cauvery, Sabarmati, Godavari catchments when compiling models. The gauging stations are run by the CWC, India, and measure water temperature (T_w) over a period of time (monthly mean of ten samples) (Central Water Commission 2018). The meteorological data used in this work are monthly minimum (T_{min}), and maximum (T_{max}) air temperatures. T_{min}, T_{max} was available from the India Meteorological Department (IMD) data on a 1^o Latitude x 1^o Longitude grids spatial resolution from 1951 to 2018. Using linear interpolation, the AT observations have

been spatially interpolated to the RWT gauging locations. To get the monthly mean AT as widely used literature (Yang & Peterson, 2017), T_{min} and T_{max} were averaged. Table 4.1 shows the catchment means for all variables.

This study used the subset of the National Aeronautics Space Administration (NASA) Earth Exchange Global Daily Downscaled Projections (NEX-GDDP) dataset to assess the impact of climate change on RWTs for seven catchments of India. The NEX-GDDP is made up of downscaled climate scenarios for the entire world produced from the GCM runs undertaken as part of the Coupled Model Intercomparison Project Phase 5 (CMIP5) and spanning two of the four greenhouse gas emissions scenarios known as Representative Concentration Pathways (RCPs) (Centre for Climate Change Research (CCCR) 2017). The NEX-GDDP dataset was created using the Bias-Correction Spatial Disaggregation (BCSD) technique, a statistical downscaling algorithm specifically developed to address the issue of regionally biased statistical characteristics (i.e., mean, variance, etc.) in global GCM outputs (Wood et al. 2002; Maurer and Hidalgo 2008; Thrasher et al. 2012). The ensemble mean of the NEX-GDDP dataset contains RCP 4.5 and RCP 8.5 downscaled projections from the 21 GCMs models and scenarios, and each climate projection has daily maximum temperature, minimum temperature, and precipitation for 1950 through 2100. The dataset has a spatial resolution of 0.25 degrees (~25 km x 25 km). This study retrieved the daily T_{min} and T_{max} values, converted them into a monthly scale, and averaging them to obtain the monthly mean AT for future RWT predictions. To perform the local scale validation, the historical AT values were compared with NEX-GDDP dataset RCP 4.5 values for Jan-2006 to Dec-2008. Results revealed that the RCP 4.5 projections are synchronous and equivalent to historical AT time series (Figure 4.4), and they have good statistical metrics (NSE: 0.92, RMSE: 1.57).

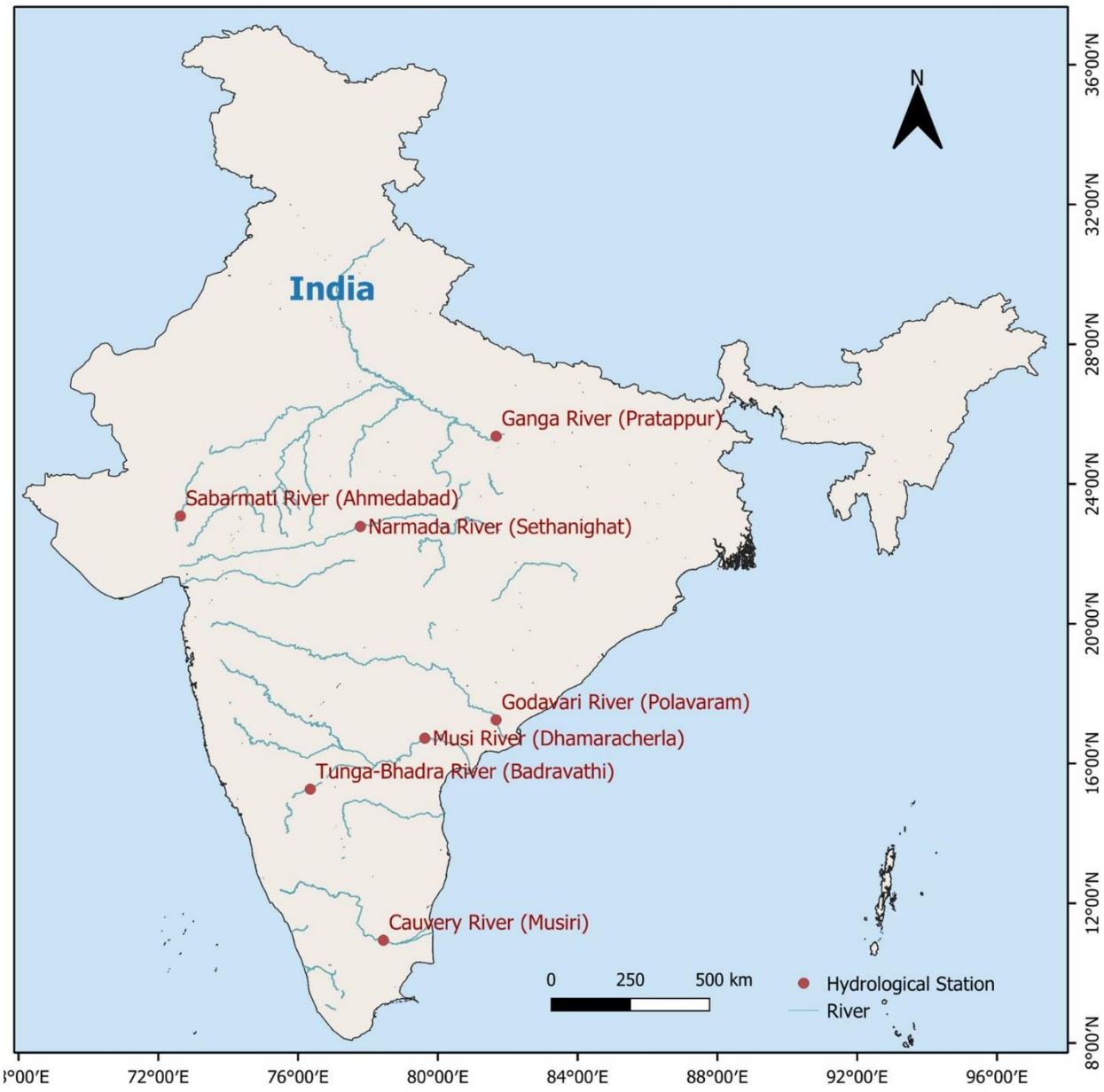


Figure 4.3. Location map of study sites in India. All catchments and gauging station information are summarized in Table 4.1.

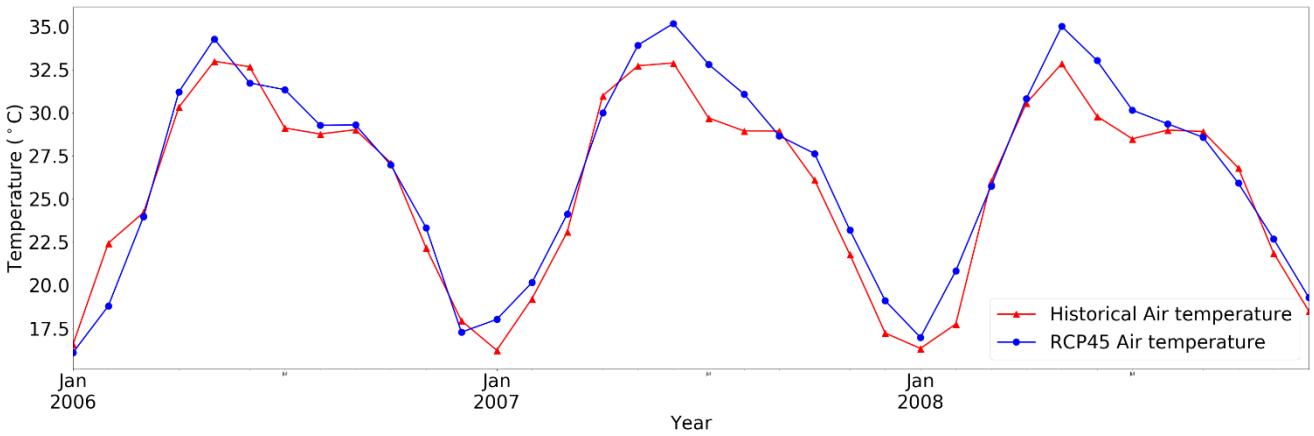


Figure 4.4. Comparison of the monthly historical and NEX-GDDP with RCP 4.5 air temperatures for the year Jan 2006 – Dec 2008.

4.3.2. Data Pre-processing

In this study, no observations were omitted from the time series for all seven catchments of India as outliers and uncertainties in observation data help troubleshoot potential issues at the modeling stage. It is also observed that the majority of the time series retrieved from the source datasets (GEMSTAT and CWC) are discontinuous. To perform both LSTM algorithms and k-NN bootstrap resampling algorithm simulation, a complete dataset is necessary. To build an entire data record (Figure 4.5), the na.interp() method in R's forecast library was utilized to interpolate the missing observations using the STL (Seasonal and Trend decomposition using Loess) decomposition (Hyndman et al. 2018). The applied data pre-processing consists of aggregating multiple data sources data and feature engineering. ML model's performance can be significantly improved by computing new features from a specified data and thus having more data representation (Bengio et al. 2013). Earlier research by Webb et al. (2003) demonstrated that lag information has a significant relationship with water temperature over time and can help models perform better. The data's autocorrelation and partial autocorrelation functions (ACF and PACF) were examined to account for the time-lag information. These functions suggest that the one-month time lag is significant in the observed record (Figure 4.6). The autocorrelation functions measure the strength of the linear relationship between successive values of a time series depending on the time lag between them. Thus, the lags of both AT and RWT variables for the one last month (AT[t-1],

RWT [t-1]) are calculated and used as additional features (inputs) for RWT predictions at time *t*.

Traditional ML and DL models for RWT modeling commonly have limitations, particularly when dealing with non-stationary data. At the same time, there might be measurement errors, which leads to noisy data. This study used the WT as a data pre-processing method to de-noise the time series data into its subcomponents to address these issues. WT transformed monthly average ATs and RWTs into approximation parts (A) and details (D). Then, the transformed forms of the monthly average ATs and RWTs were then used as model input in the LSTM model for training and testing (Figure 4.1b).

Table 4.1. Summary of study catchment information characteristics, catchment means of water temperature (T_w), air temperature (T_a), available data periods, training, and testing periods.

Catchment	**Gauging Station**	**Catchment area (km^2)**	**Time Period**	**Training Period**	**Testing Period**	**(Lat, Long)**	**T_w (^oC)**	**T_a (^oC)**
Narmada	Hoshangabad	97,410	1985-2008	1985-2003	2004-2008	22.76, 77.74	24.68	25.09
Cauvery	Musiri	81,155	1985-1999	1985-1995	1996-1999	10.94, 78.44	30.34	28.81
Sabarmati	Ahmadabad	21,674	1980-2008	1980-2004	2005-2008	23.08, 72.63	28.08	26.72
Tunga-Bhadra	Badravathi	2,58,948	2006-2017	2006-2014	2015-2017	15.27, 76.35	26.34	24.24
Musi	Dhamaracherla	11,212	1991-2005	1991-2002	2003-2005	16.74, 79.67	27.97	28.13
Godavari	Polavaram	3,12,812	1980-2008	1980-2004	2005-2008	17.25, 81.66	28.17	27.48
Ganga	Pratappur	8,61,404	2000-2015	2000-2011	2012-2015	25.37, 81.67	25.64	25.71

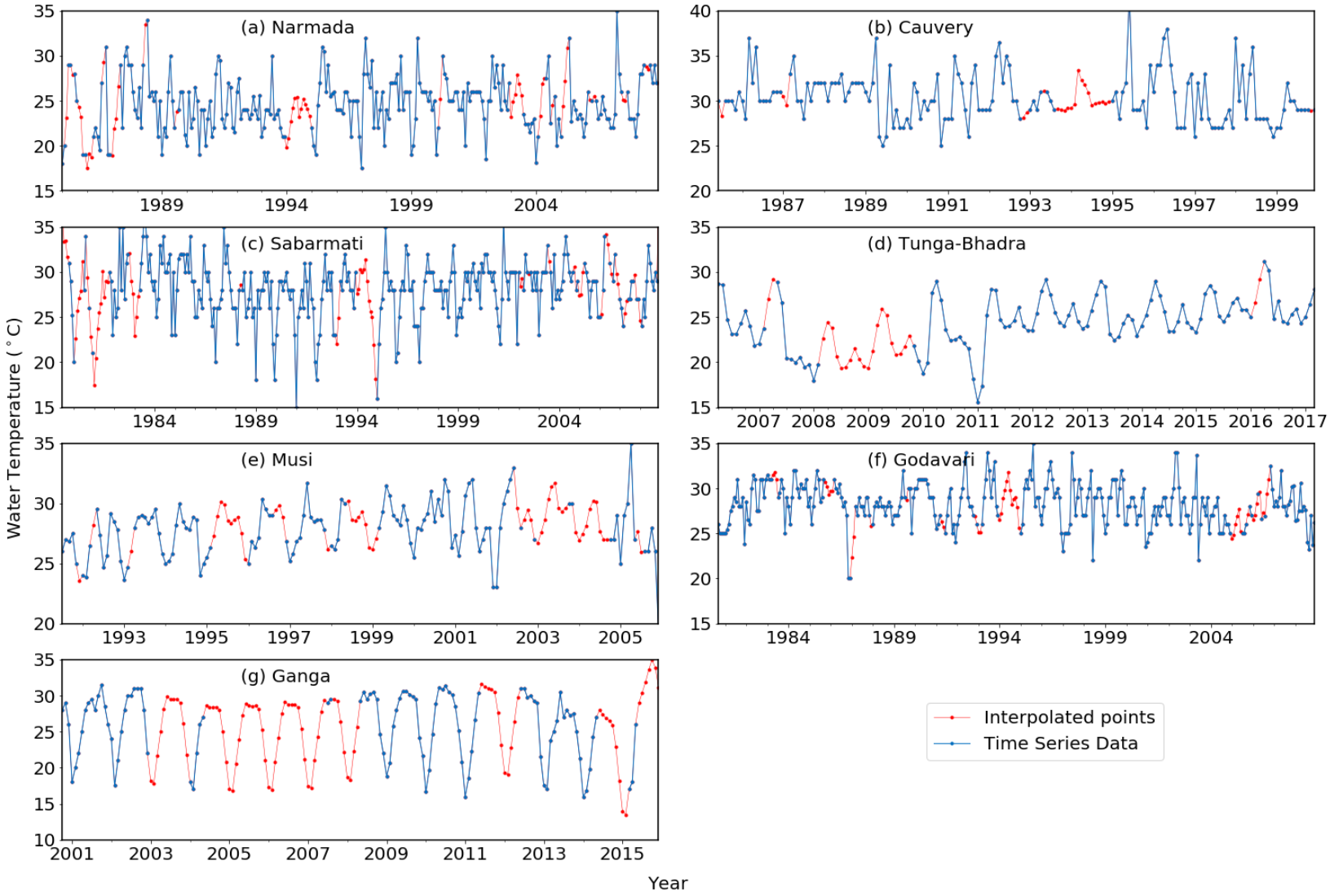


Figure 4.5. Monthly catchment sites time series data, red color represent missing values that were filled in using time series interpolation

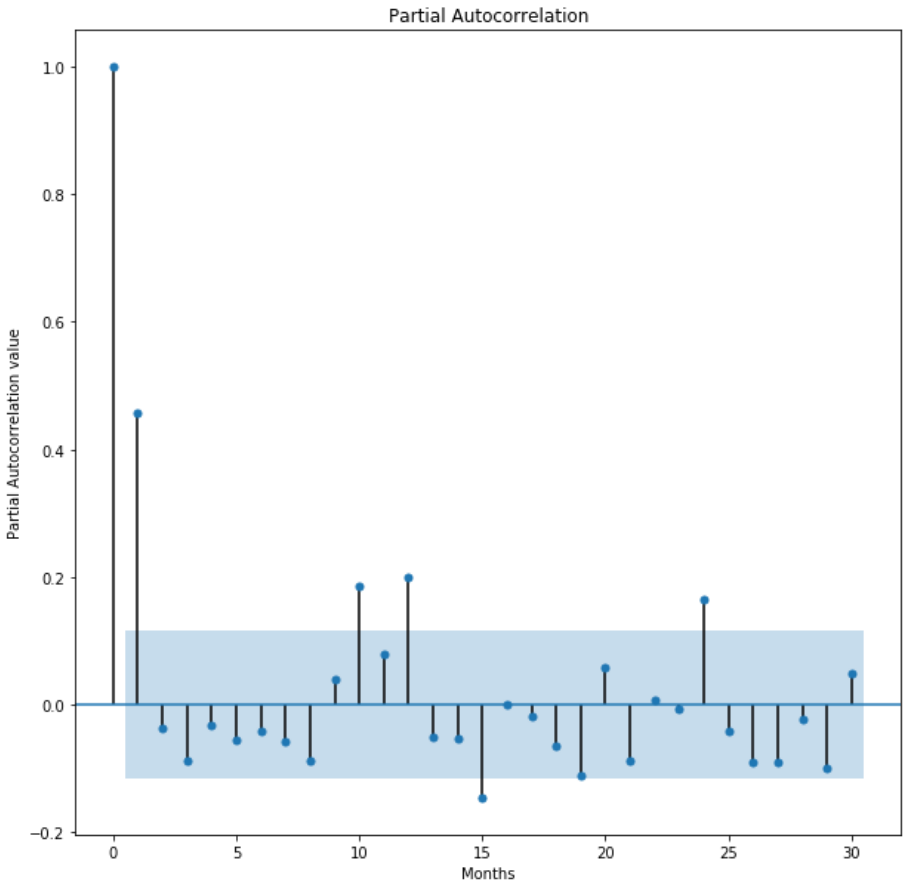


Figure 4.6. Partial autocorrelation function (PACF) for the monthly river water temperature time series at Narmada Catchment station.

4.3.3. Parameterization and Settings

The data sets were divided into two parts to compare all applied models objectively: the first 80% of the time-series data were utilized for training and the last 20% for testing. Overfitting is a significant issue with ML methods. For classical ML algorithms (SVR, RF, etc.), overfitting can be avoided by using the Cross-Validation (CV), and regularization techniques. A careful selection of a set of hyperparameters and early stopping is required for the DL algorithms to avoid overfitting (Feigl et al. 2021). Manual, grid search, random search, and Bayesian optimization (BO) are the standard methods of hyperparameter optimization for ML and DL to increase efficiency. Unlike random or grid search, BO is a global optimization method for Blackbox functions that keeps track of previous evaluation results. Kushner (1964) and Mockus (1989) originated the BO, which was later demonstrated

by Jones et al. (1998). After Snoek et al. (2012) research, it became notably well recognized for optimizing ML hyperparameters. To summarize, BO builds a surrogate model by employing a Gaussian process model to identify an appropriate next point at each iteration during optimization. In this study, while training a standalone LSTM, kNN-LSTM, and WT-LSTM model on a time series, all the possible combinations of LSTM hyperparameter sets (the number of LSTM hidden layers: 1-3, the total number of units per layer: 5-100, timesteps:1-12, the dropout ratio: 0-0.4, epochs: 50-100, and the batch size: 2-64) are evaluated using an emerging state-of-the-art BO approach to optimize the hyperparameters, and the topmost group is chosen to improve the model's performance. For hyper parameter optimization for all LSTM models was done by using a training split with 60 % data for training and 20 % data for validation. In this study, Daubechies wavelet of order 5 (DB5), which has been used at level 3, was chosen to train the WT-LSTM model as it is well-known in the literature, and its wavelet coefficients can capture the maximum amount of signal energy (Seo et al. 2015). The adoption of this approach in the current study resulted in a decomposition level of 3 for the seven Indian catchments. The decomposition level can be determined using the method provided by Nourani et al. (2009). Consequently, D1, D2, and D3 were detail time series, and A3 was the approximation time series (Figure 4.7). However, advances in the model performances slowed down when the decomposition level was greater than 3. This demonstrated that the WT-LSTM model might achieve significant accuracy by employing a decomposition level of 3.

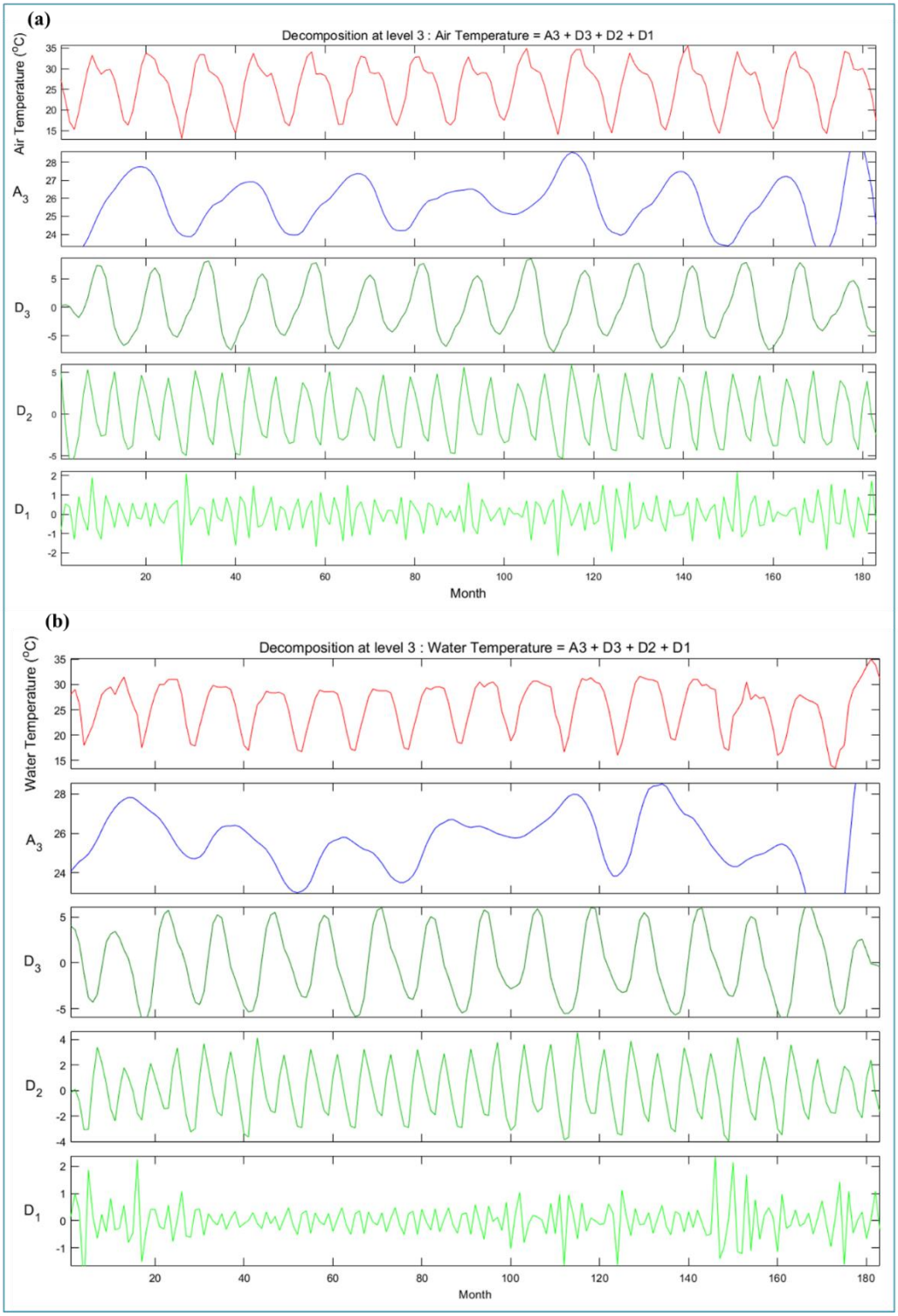
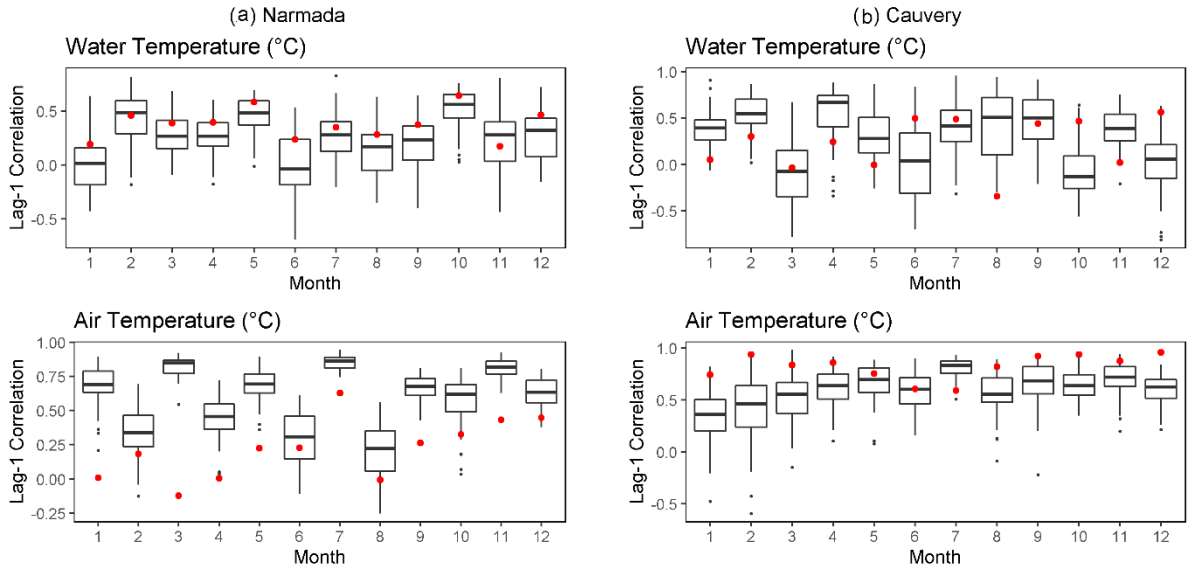


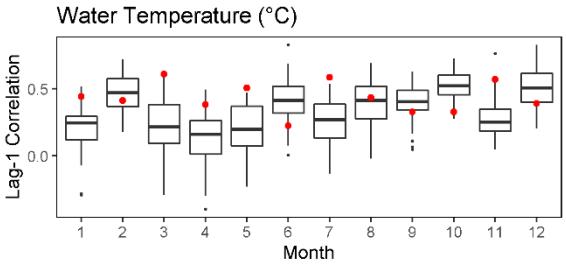
Figure 4.7. (a) Original and decomposed Air Temperature time series (A3, D1, D2, and D3) (b) Original and decomposed Water Temperature time series (A3, D1, D2, and D3) using db5 wavelet for the Ganga catchment station. A3 is the decomposed approximation part, and D1, D2, and D3 are the decomposed details.

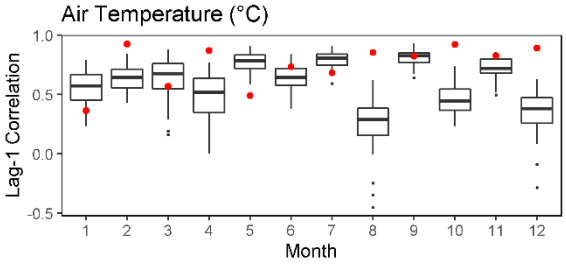
In the k-NN bootstrap resampling algorithm, “one simulation” is defined as a set of simulated values with a length equal to the observed dataset and chosen to generate 50 simulations. Following that, a comparison study of monthly statistics (maximum, minimum, mean, and standard deviation) was performed for both the historical and simulated ensemble records. Also, the lag-1 autocorrelation of the k-NN simulated RWT and AT was compared (Figure 4.8). The comparison (Figure 4.8) has revealed that the algorithm produced the applicable distributional statistics of the observed dataset, implying that the algorithm generates accurate and diverse conditions. The lag-1 autocorrelation represents the relationship between two consecutive time steps (e.g., x_t and x_{t-1}). When the lag-1 autocorrelation of the historical and simulated record is compared (Figure 4.8), the lag-1 autocorrelation’s seasonality is frequently reproduced. Then, the simulated values of the monthly average ATs and RWTs were then used as model input in the LSTM model (Figure 4.1c).

This study implemented the LSTM networks in Python using the Keras DL library to solve the RWT time-series prediction problem. Keras wraps the Theano and TensorFlow libraries to build DL models with less number of lines. These models will run on CPU and GPU; hence computation is speedy. All programming was done in Python, R, and MATLAB.

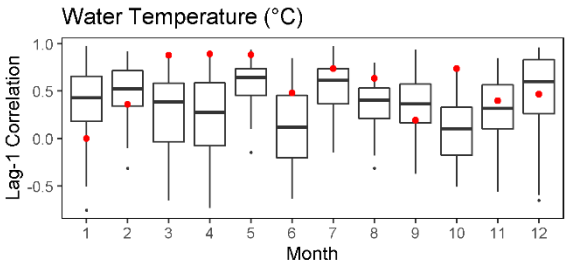


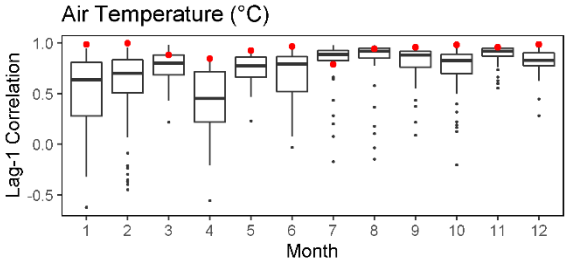
(c) Sabarmati



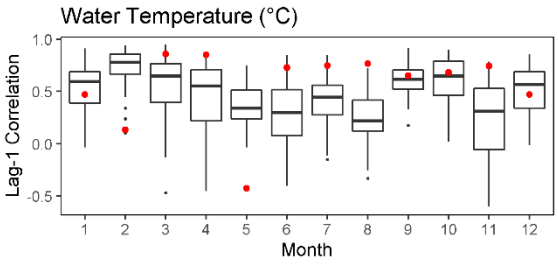


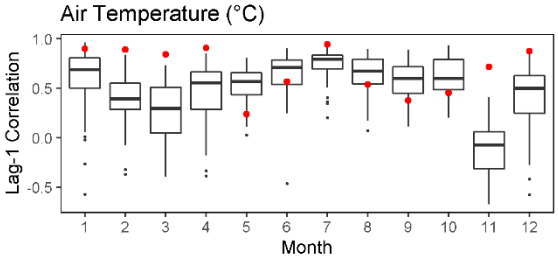
(d) Tunga-Bhadra



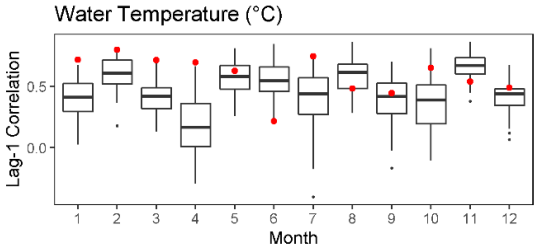


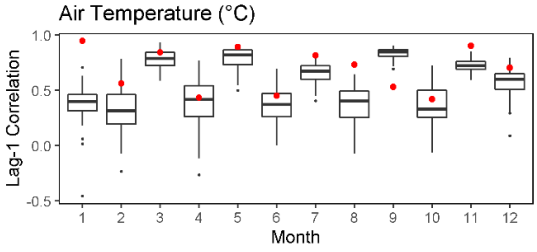
(e) Musi



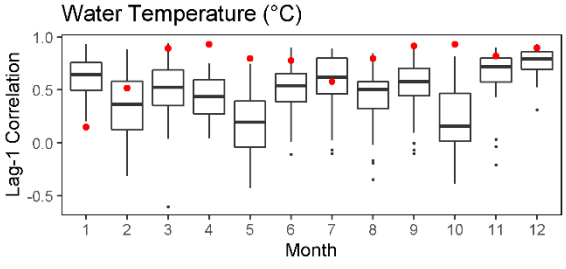


(f) Godavari





(g) Ganga



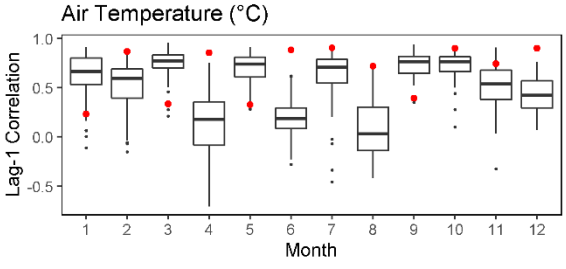


Figure 4.8. Observed (red points) and k-NN simulated (white box plots) lag-1 (i.e., month-to-month) autocorrelation for water temperature, and air temperature for seven catchments (a) Narmada (b) Cauvery (c) Sabarmati (d) Tunga-Bhadra (e) Musi (f) Godavari (g) Ganga.

4.3.4. Model Evaluation Metrics

To mathematically quantify the predictive performances of DL models, five statistical measures are calculated, such as the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970), Kling–Gupta efficiency (KGE) (Kling et al. 2012), RMSE-observations standard deviation ratio (RSR) (Moriasi et al. 2007), the root mean squared error (RMSE), and the mean absolute error (MAE). Detailed descriptions of these metrics can be found in section 3.4, chapter 3.

4.4. Results

The data used in this work comprises monthly average AT and the corresponding RWT for seven catchments of India. The seasons were defined by meteorological definitions as follows: monsoon=June, July, August, September; post-monsoon=October, November; winter = December, January, February; and summer = March, April, May (IMD 2021). The catchment means of RWT, AT for all seven catchments ranged between 24.68 ^{o}C, 30.34 ^{o}C, and 24.24 ^{o}C, 28.81 ^{o}C, respectively (Table 4.1). Spearman's correlation coefficients (SCC) for seven catchments were estimated to examine the statistical dependency between AT and RWT variables. According to the metrics from Table 4.2, the RWT was positively correlated with AT for the selected catchments on the annual scale. RWT was weakly correlated with AT during the winter months (December-February) for Tunga-Bhadra and Ganga catchments but positively correlated in other seasons (Table 4.2). All the catchments are positively correlated in the summer months except the Cauvery catchment.

To examine the variability of annually averaged AT and RWT changes, the study calculated the linear trends for seven catchments of India (Figure 4.9). The AT and the RWT increased during the studied period for all catchments except Cauvery, Godavari, and Ganga catchments (Figure 4.9). The RWT increasing rates are lower than those of AT in general.

Air temperature shows a rising trend except for Cauvery (-0.01 $^{\circ}$C/year) catchment,

and the rising rates range from 0.002 to 0.380 ˚C/year. RWT shows a rising trend except for Cauvery (-0.06 ˚C/year), Godavari (-0.03 ˚C/year), and Ganga (-0.07 ˚C/year) catchments, and the rising rates vary between 0.01 and 0.17 ˚C/year. Such RWT rising patterns have been explored in several locations throughout the world. The RWT, for instance, has been a rising trend vary between 0.009–0.077 ^{o}C year^{-1} over the USA (Isaak et al., 2012; K. C. Rice & Jastram, 2015; van Vliet et al., 2013), over the China of about 0.029–0.046 ^{o}C year^{-1} (Chen et al., 2016), British Columbia as ~0.036 ^{o}C year^{-1} (Islam et al. 2019), and Europe as 0.006–0.180 ^{o}C year^{-1} (Orr et al. 2015; Hardenbicker et al. 2017). Generally, RWT and AT follow similar variability, i.e., RWT increases are directly related to AT increases. However, for the Godavari and Ganga catchment, the water temperature has a decreasing trend (-0.03 ˚C/year and -0.07 ˚C/year respectively) with an increasing trend of AT (0.01 ˚C/year and 0.08 ˚C/year, respectively), which specifies that the temporal shifts of RWT may not be explained AT alone. RWT is directly influenced by multiple parameters, including streamflow (Sohrabi et al. 2017), river geometry, groundwater inputs, slope, water depth, etc. (Gu and Li 2002).

Table 4.2. Seasonal period Spearman's correlation coefficients between air and water temperature variables at different catchment areas

Catchment	**Correlation Coefficient (RWT - AT)**				
	Summer (March-May)	**Monsoon (June – September)**	**Post-monsoon (October – November)**	**Winter (December – February)**	**Annual**
Narmada	0.44	0.34	0.43	0.38	0.56
Cauvery	0.01	0.22	0.23	0.17	0.28
Sabarmati	0.29	0.39	0.39	0.18	0.56
Tunga-Bhadra	0.47	0.32	0.25	0.01	0.43
Musi	0.67	0.38	0.57	0.40	0.63
Godavari	0.38	0.16	0.21	0.18	0.42
Ganga	0.87	0.19	0.66	0.06	0.66

4.4.1. Seasonality Trends

Seasonality dynamics in RWTs in India were varied throughout the time and between

catchments, with no consistent temporal patterns among the locations (Figures 4.9, 4.10 and 4.11). Narmada, Sabarmati, Tunga-Bhadra, and Musi catchments had increasing winter RWTs (0.08, 0.12, 0.11, and 0.16 ˚C/year, respectively) over the years (Figures 4.5a, 4.5c, 4.5d, and 4.5e). Several catchments, e.g., Cauvery, Godavari, and Ganga, had decreasing trends (-0.04, -0.02, and -0.24 ˚C/year, respectively) in winter RWTs over the years (Figures 4.5b, 4.5f and 4.5g). Musi had the highest increasing winter RWT trend (0.16 ◦C/year), and Ganga had the highest decreasing winter RWT trend (-0.24 ◦C/year) compared to seven catchments of India. Narmada, Sabarmati, Tunga-Bhadra, and Musi catchments had increasing summer RWTs (0.13, 0.06, 0.35, and 0.22 ˚C/year, respectively) over the years (Figures 4.5a, 4.5c, 4.5d, and 4.5e). Other catchments, Cauvery, Godavari, and Ganga had decreasing trends (-0.15, -0.03, and -0.05 ˚C/year, respectively) in summer water temperatures over the years (Figures 4.5b, 4.5f, and 4.5g). Tunga-Bhadra and Musi had the highest increasing summer RWT trend (0.35, 0.22 ◦C/year respectively) compared to all the catchments, and Cauvery had the highest decreasing summer RWT trend (-0.15 ˚C/year). These patterns give some indication of the recent warming of RWTs during the summer seasons. A consistent seasonal trend in RWT was observed for the monsoon and post-monsoon seasons (Figure 4.11). Except for Sabarmati and Godavari catchments, RWTs are similar for monsoon and post-monsoon seasons (see Figure 4.11). In this work, Morlet CWT is used to show the wavelet power spectra to examine the long-term variability of temperature time series data from 2000 to 2015 for the Ganga catchment station. It has been examined if the CWTs could successfully identify variability in the annual scale. The local wavelet power spectrum measures the variance distribution of the time series according to time and periodicity; high variability is represented by red color, whereas blue indicates weak variability. The wavelet power spectrum showed that (Figure 4.12) displayed a consistent 1-year periodicity from 2000-2015. Air Temperature and Water Temperature showed stable 1-year periodicities for the whole duration of the time series and displayed a recurrent annual pattern. This steady behavior means that the system's complexity has not deteriorated so that the seasonality of this process has not perceptibly changed over time. The continuous wavelet analysis results, as demonstrated in this work, were found to agree with Alcocer et al. (2022), where the author performed the CWTs to examine the variability of water quality variables and concluded that the water quality variables display a recurrent annual cycle.

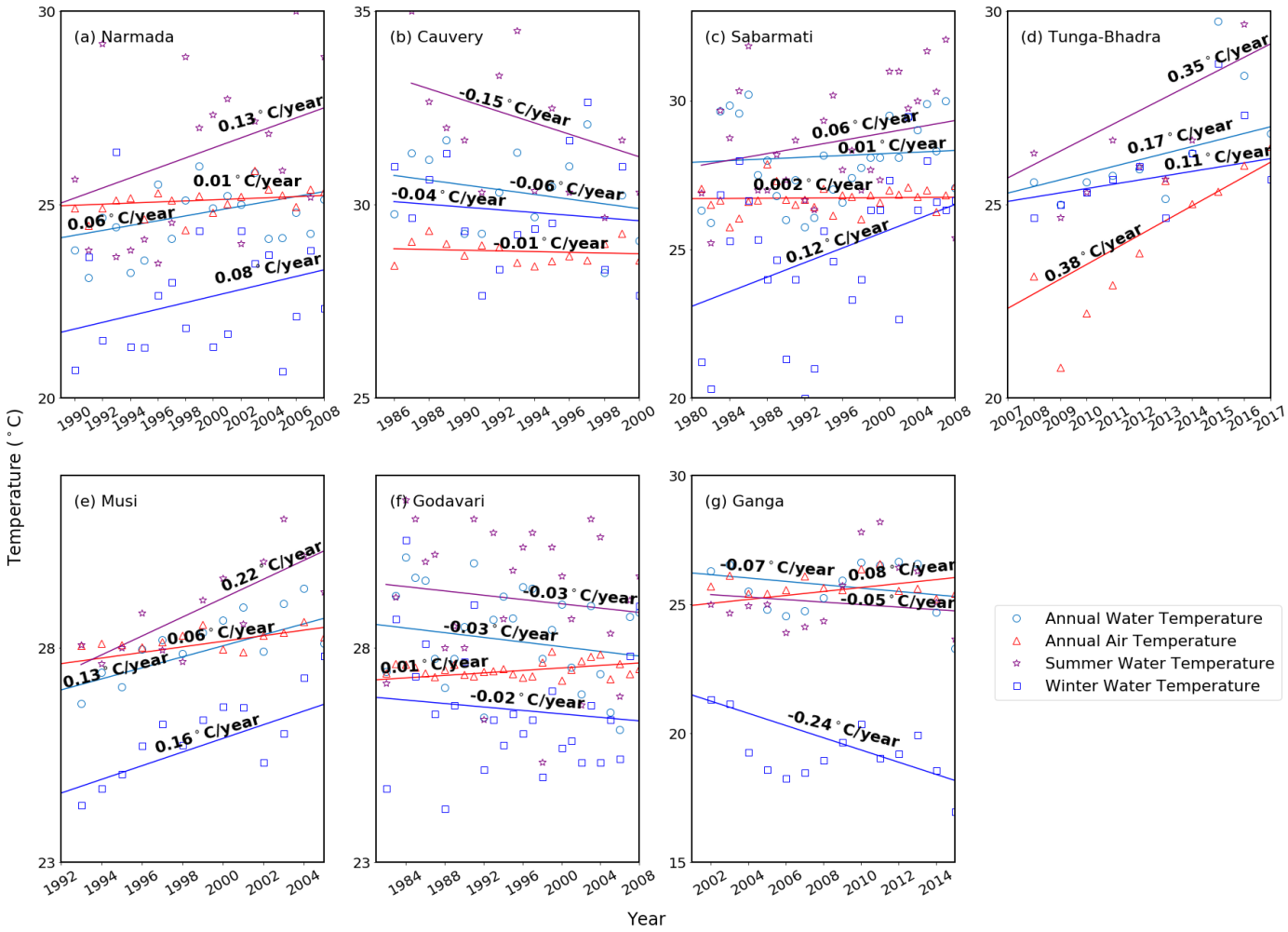


Figure 4.9. Seasonal, temporal variations of the mean annual air temperature (red), water temperature (light blue), summer water temperature (purple), and winter water temperature (blue) of the seven catchment stations (a) Narmada (b) Cauvery (c) Sabarmati (d) Tunga-Bhadra (e) Musi (f) Godavari (g) Ganga. Linear regressions of the time series are represented by trend lines, and the slope parameters are trend estimations.

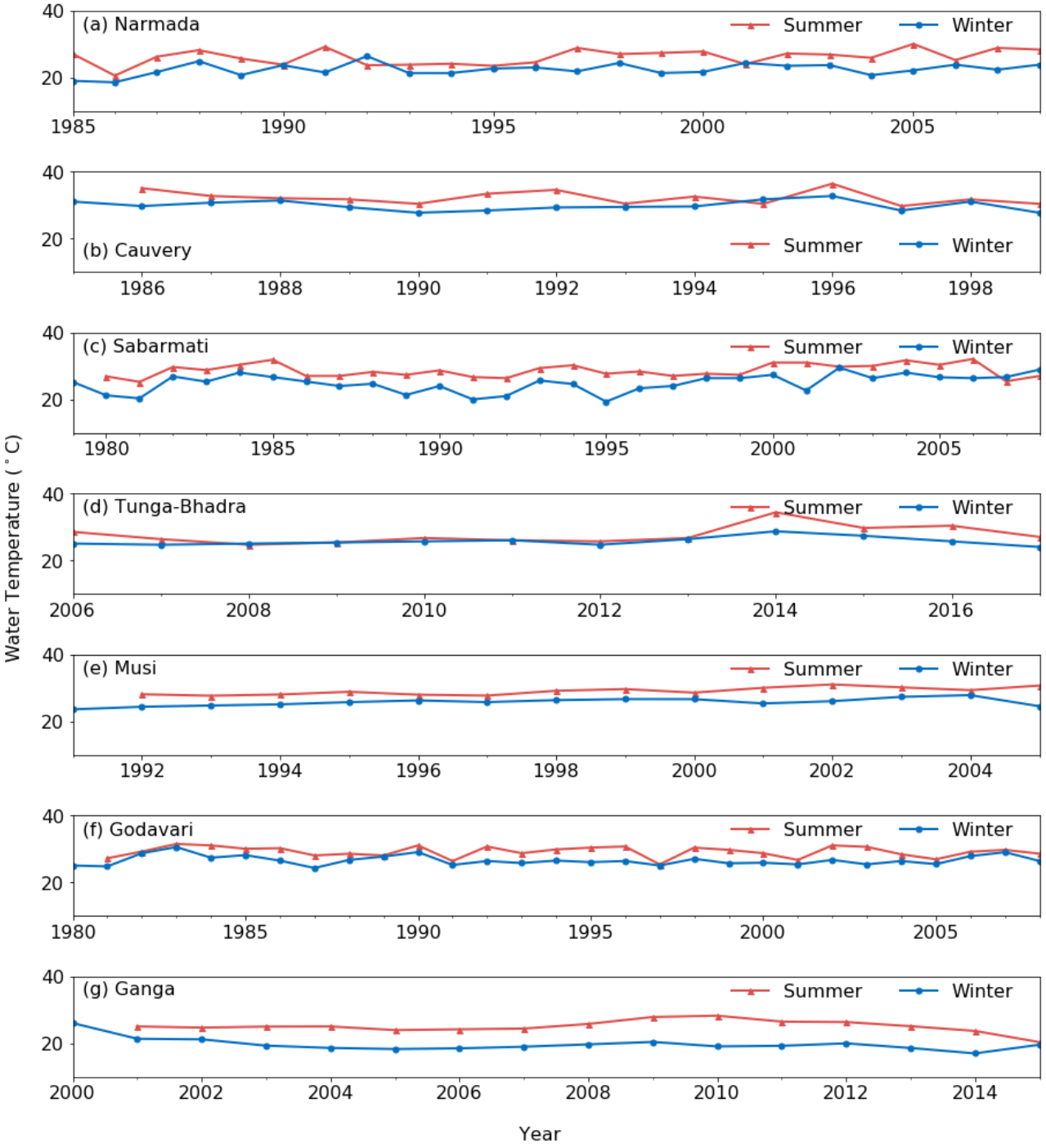


Figure 4.10. Temporal trends in summer and winter river water temperatures for catchments are located in India. The values are averages across seasonal months per point.

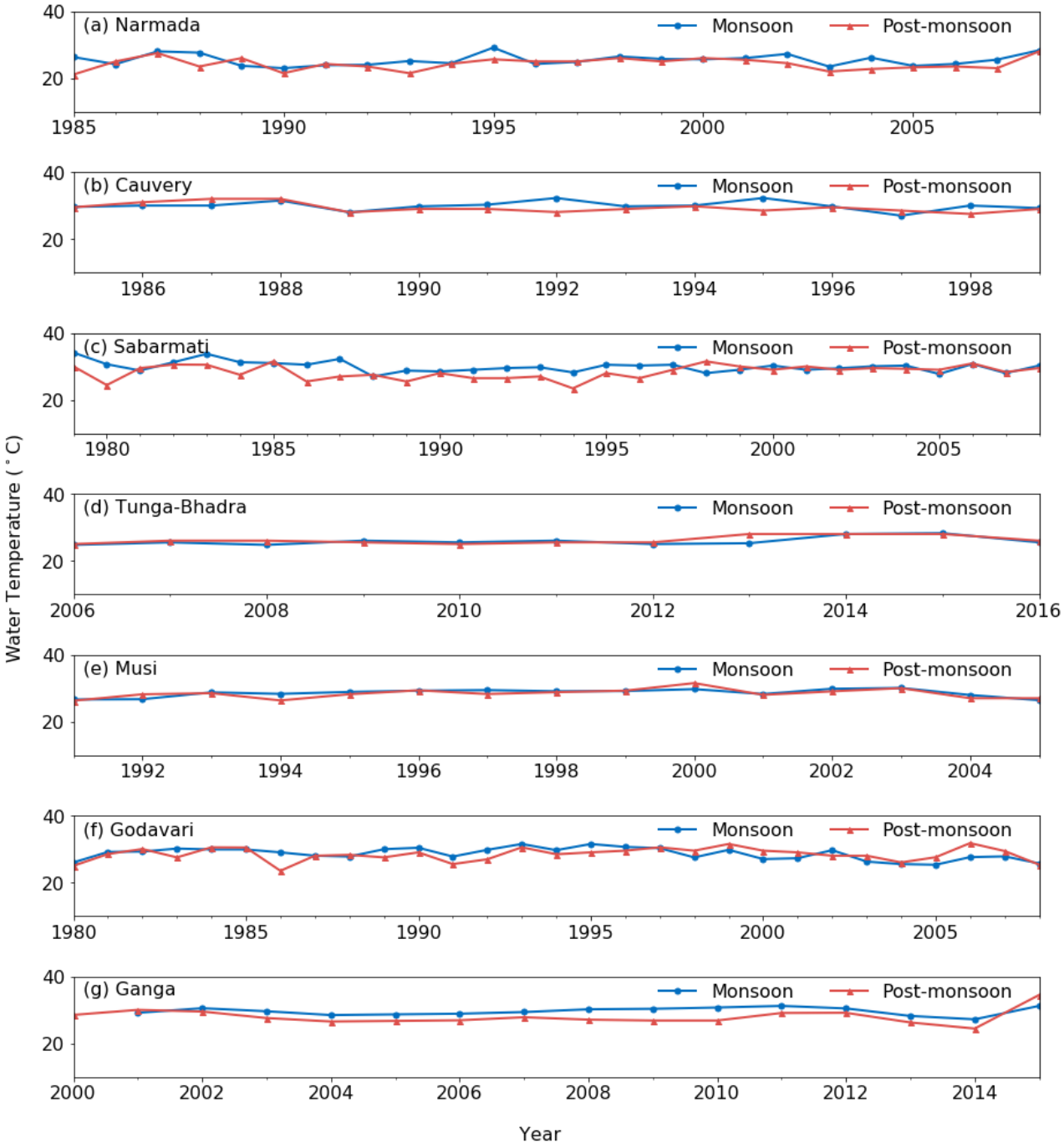


Figure 4.11. Temporal trends in monsoon and post-monsoon river water temperature for catchments are located in India. The values are averages across seasonal months per point.

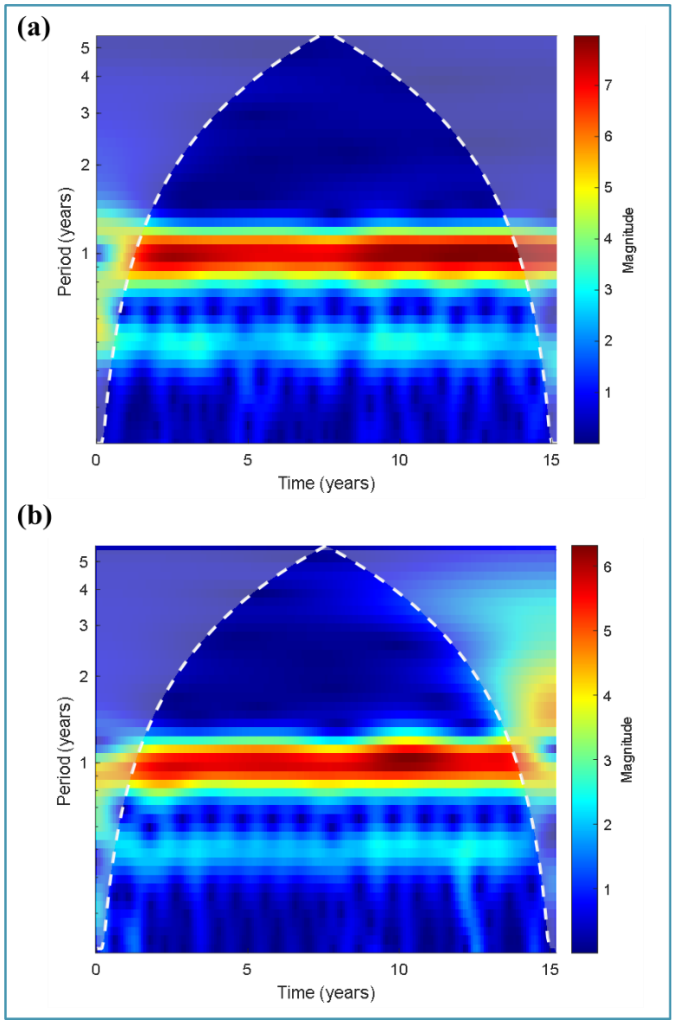


Figure 4.12. (a) Continuous wavelet power spectra of historical (2000-2015) Air temperature and (b) Continuous wavelet power spectra of the historical (2000-2015) water temperature show the periodicity of the Ganga catchment station; the blue color demonstrates the lower power spectra and the red color the higher, and the dotted line is the cone of influence.

4.4.2. Deep Learning Model Performance

The LSTM, WT-LSTM, kNN-LSTM, and 3-parameter version air2stream approaches were assessed using statistical measures (R^2, KGE, NSE, RSR, RMSE, and MAE) and visual comparisons. The experiment outcomes revealed a matching of observed against predicted values, indicating that hybrid LSTM techniques have robust generalization capacity. The DL model's performance results were generated for monthly data for all seven catchments using AT and lag variables as inputs, as provided in Table 4.4 and Figure 4.13. The performance of

the proposed methods was assessed by comparing them with those obtained from algorithms based on only AT as an input variable (Table 4.3). Results confirm that the developed models could predict RWT more accurately than AT as an input variable by utilizing AT and lag variables as input. Results also revealed that the RWT predictions are nearly synchronous and equivalent to observed time series (Figure 4.13). However, the air2stream model generated unsatisfactory results, and statistical measures can be found in Table 4.4.

The relationship between monthly RWT and AT at seven catchments is relatively strongly correlated for all models except the air2stream model (NSE values). The RMSE metrics for all the catchments vary from 1.199 to 3.294 for all the models for monthly data (Table 4.4). The WT-LSTM hybrid models used in RWT prediction indicated a high resemblance in all the seven catchments (Table 4.4, Figure 4.14). The predicted results were superior to those obtained from the standalone LSTM models, including the air2stream, as evidently shown in Figure 4.13 and Figure 4.14. Comparing the WT-LSTM and the standalone LSTM model, in terms of individual time series matching of observed against predicted values, it can be concluded that the combining of WT and LSTM produced improved results than the traditional LSTM model for RWT prediction. The NSE values for all the catchments range from 0.329 to 0.920 for the WT-LSTM model (Table 4.4) for monthly data. The NSE value for the Ganga catchment is obtained as 0.920 for the WT-LSTM model, which is reasonable compared with earlier standalone LSTM models by Stajkowski et al. (2020) (NSE: 0.913) and Qiu et al. (2021) (NSE: 0.74 – 0.99 °C). However, Stajkowski *et al.* (2020) used AT values as input for hourly data in their analysis, Qiu *et al.* (2021) used AT and discharge as input for daily data in RWT predictions, and the current study is dedicated to monthly timescales.

The simulated samples from k-NN bootstrap resampling algorithm given as input to the kNN-LSTM hybrid model to predict the RWT for all the seven catchments of India. The results were superior to those obtained from the LSTM, WT-LSTM models, including the air2stream, as evidently shown in Table 4.4, Figure 4.13, and Figure 4.14. In this study, the k-NN bootstrap resampling algorithm is experimented with discrete wavelet components, i.e., decomposed time series are used to simulate the values from the k-NN bootstrap resampling algorithm instead of the original time series data. From the testing results (Table 4.5), it was observed that the WT-kNN-LSTM was giving not good performant results like the kNN-

LSTM model. Comparing the kNN-LSTM hybrid model with the WT-LSTM model, LSTM model, and air2stream model, in terms of individual time series matching of observed against predicted values (Figure 4.13), according to the results, the kNN-LSTM model outperformed three other models (WT-LSTM, standalone LSTM, and air2stream) to predict the RWT. In this study, the kNN-LSTM model is tested using various monthly data points (Table 4.6) to see how the model performed for Musi and Ganga catchment stations with data limitations. It was observed that the kNN-LSTM was still producing good results with fewer monthly data time series values.

Based on RSR, KGE, and NSE performance values (Table 4.4 and Figure 4.14), the WT-LSTM model is the best performant model for Sabarmati and Tunga-Bhadra catchments (Table 4.4). These results are superior to those obtained from the standalone LSTM, air2stream model (Table 4.4 and Figure 4.14). The calibration and validation metrics of the air2stream model are shown in Table 4.7, and it observed that the air2stream model performed better in the calibration phase, however in the validation phase, its performance slightly decreased. In general, the performances of the air2stream model on a monthly scale were not satisfactory (Figure 4.14, Table 4.7). The performance of the LSTM model (NSE= 0.132 – 0.886, KGE = 0.131 – 0.818, RMSE=1.849 – 2.950, RSR = 0.336 – 0.932, and MAE = 1.215 – 2.467) was much superior to that of the air2stream. Overall, the kNN-LSTM model statistical metrics are reasonably within the range for all the catchment locations providing confidence that the developed model performs effectively.

Table 4.3. Overview of deep learning models performances for seven catchments with only air temperature as an input variable. The shown values all refer to the test time.

Catchment	Model	NSE	KGE	RMSE	RSR	MAE
Narmada	LSTM	0.232	0.323	2.597	0.876	1.866
	WT-LSTM	0.243	0.336	2.580	0.870	1.844
	kNN-LSTM	0.289	0.334	2.697	0.842	2.004
Cauvery	LSTM	0.095	0.018	3.016	0.951	2.597
	WT-LSTM	0.099	0.012	3.008	0.948	2.606
	kNN-LSTM	0.253	0.246	2.526	0.864	1.748
Sabarmati	LSTM	0.248	0.358	2.488	0.867	1.652
	WT-LSTM	0.266	0.367	2.457	0.856	1.627
	kNN-LSTM	0.344	0.434	2.835	0.809	2.084
Tunga-Bhadra	LSTM	0.298	0.551	2.251	0.837	1.745
	WT-LSTM	0.302	0.580	2.245	0.835	1.710
	kNN-LSTM	0.236	0.242	1.699	0.873	1.103
Musi	LSTM	0.256	0.332	2.139	0.862	1.616
	WT-LSTM	0.255	0.319	2.142	0.863	1.600
	kNN-LSTM	0.501	0.565	1.425	0.706	1.021
Godavari	LSTM	0.207	0.368	1.784	0.890	1.369
	WT-LSTM	0.183	0.356	1.810	0.903	1.405
	kNN-LSTM	0.350	0.450	1.887	0.806	1.367
Ganga	LSTM	0.811	0.757	2.393	0.435	1.841
	WT-LSTM	0.812	0.772	2.379	0.432	1.827
	kNN-LSTM	0.896	0.910	1.458	0.322	1.047

Table 4.4. Overview of deep learning models performances for seven catchments based on air temperature (AT[t]), including time-lag effects of air and water temperatures (AT[t-1], RWT[t-1]) as input variables. The air2stream performances are based on air temperature (AT[t]) as an input variable. The values displayed all referred to the testing period.

Catchment	Model	NSE	KGE	RMSE	RSR	MAE
Narmada	LSTM	0.473	0.542	2.150	0.725	1.548
	WT-LSTM	0.601	0.675	1.730	0.631	1.218
	kNN-LSTM	0.728	0.715	1.547	0.522	1.198
	air2stream	0.324	0.359	2.523	0.812	1.925
Cauvery	LSTM	0.132	0.131	2.950	0.932	2.467
	WT-LSTM	0.329	0.386	2.328	0.819	1.860
	kNN-LSTM	0.446	0.378	2.361	0.744	1.872
	air2stream	0.010	0.001	3.294	0.991	2.679
Sabarmati	LSTM	0.271	0.388	2.449	0.853	1.676
	WT-LSTM	0.630	0.616	1.429	0.608	1.063
	kNN-LSTM	0.579	0.668	1.861	0.648	1.300
	air2stream	0.072	0.335	2.601	0.958	1.946
Tunga-Bhadra	LSTM	0.501	0.636	1.899	0.706	1.579
	WT-LSTM	0.586	0.790	1.474	0.643	1.126
	kNN-LSTM	0.552	0.609	1.797	0.668	1.522
	air2stream	0.064	0.090	2.109	0.952	1.791
Musi	LSTM	0.434	0.524	1.865	0.751	1.215
	WT-LSTM	0.545	0.629	1.503	0.674	0.965
	kNN-LSTM	0.735	0.701	1.277	0.514	0.802
	air2stream	0.227	0.339	2.162	0.866	1.575
Godavari	LSTM	0.311	0.488	1.662	0.829	1.278
	WT-LSTM	0.563	0.676	1.199	0.660	0.937
	kNN-LSTM	0.600	0.643	1.266	0.631	1.038
	air2stream	0.107	0.325	2.432	1.183	1.943
Ganga	LSTM	0.886	0.818	1.849	0.336	1.431
	WT-LSTM	0.920	0.828	1.537	0.281	1.148
	kNN-LSTM	0.920	0.868	1.557	0.283	1.201
	air2stream	0.715	0.741	2.941	0.528	2.208

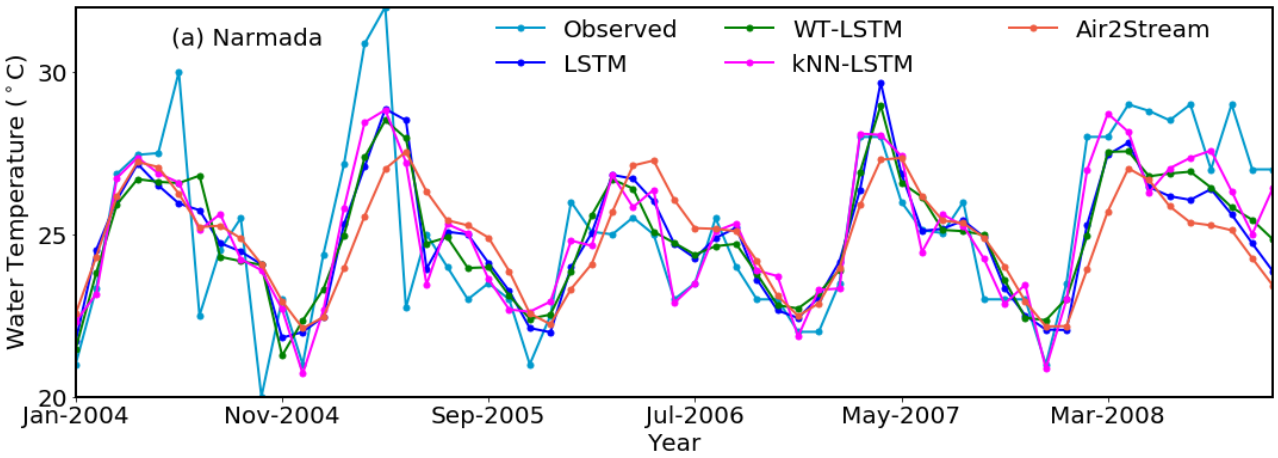
Table 4.5. Overview of deep learning models performances for seven catchments based on air temperature (AT[t]), including time-lag effects of air and water temperatures (AT[t-1], RWT[t-1]) as input variables. The air2stream performances are based on air temperature (AT[t]) as an input variable. The values displayed all referred to the testing period.

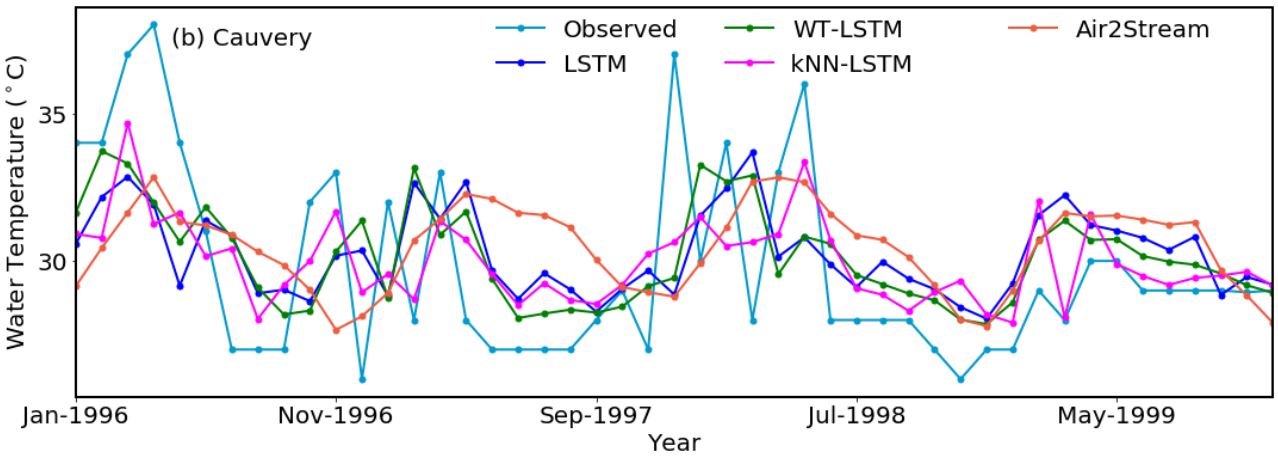
Catchment	Model	NSE	KGE	RMSE	RSR	MAE
Narmada	LSTM	0.473	0.542	2.150	0.725	1.548
	WT-LSTM	0.601	0.675	1.730	0.631	1.218
	WT-kNN-LSTM	0.555	0.591	1.975	0.666	1.531
	kNN-LSTM	0.728	0.715	1.547	0.522	1.198
	air2stream	0.324	0.359	2.523	0.812	1.925
Cauvery	LSTM	0.132	0.131	2.950	0.932	2.467
	WT-LSTM	0.329	0.386	2.328	0.819	1.860
	WT-kNN-LSTM	0.360	0.400	2.526	0.799	1.904
	kNN-LSTM	0.446	0.378	2.361	0.744	1.872
	air2stream	0.010	0.001	3.294	0.991	2.679
Sabarmati	LSTM	0.271	0.388	2.449	0.853	1.676
	WT-LSTM	0.630	0.616	1.429	0.608	1.063
	WT-kNN-LSTM	0.412	0.625	2.200	0.766	1.633
	kNN-LSTM	0.579	0.668	1.861	0.648	1.300
	air2stream	0.072	0.335	2.601	0.958	1.946
Tunga-Bhadra	LSTM	0.501	0.636	1.899	0.706	1.579
	WT-LSTM	0.586	0.790	1.474	0.643	1.126
	WT-kNN-LSTM	0.532	0.589	1.851	0.683	1.673
	kNN-LSTM	0.552	0.609	1.797	0.668	1.522
	air2stream	0.064	0.090	2.109	0.952	1.791
Musi	LSTM	0.434	0.524	1.865	0.751	1.215
	WT-LSTM	0.545	0.629	1.503	0.674	0.965
	WT-kNN-LSTM	0.683	0.737	1.396	0.562	0.920
	kNN-LSTM	0.735	0.701	1.277	0.514	0.802
	air2stream	0.227	0.339	2.162	0.866	1.575

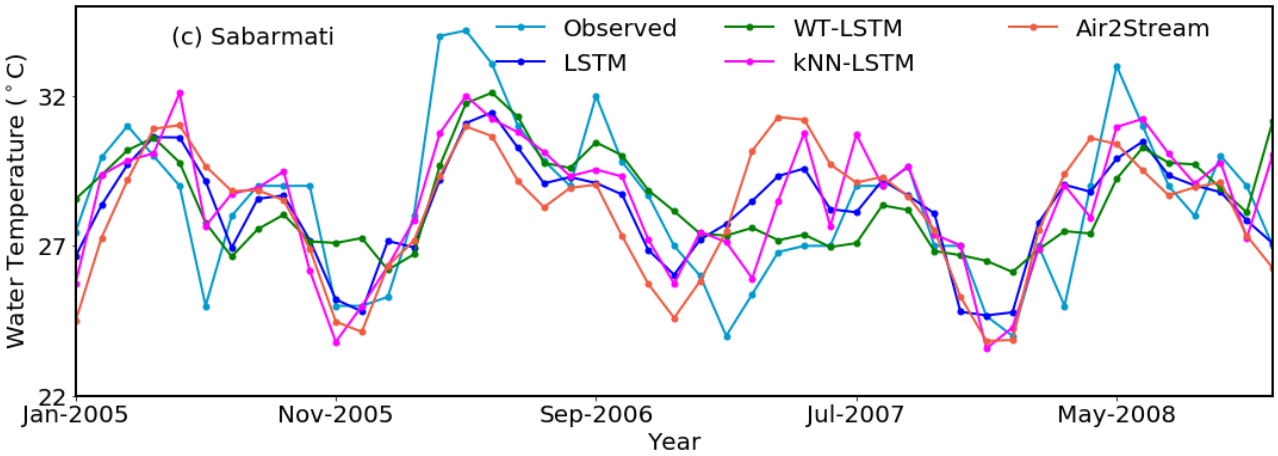
	LSTM	0.311	0.488	1.662	0.829	1.278
	WT-LSTM	0.563	0.676	1.199	0.660	0.937
Godavari	WT-kNN-LSTM	0.338	0.454	1.631	0.814	1.233
	kNN-LSTM	0.600	0.643	1.266	0.631	1.038
	air2stream	0.107	0.325	2.432	1.183	1.943
	LSTM	0.886	0.818	1.849	0.336	1.431
	WT-LSTM	0.920	0.828	1.537	0.281	1.148
Ganga	WT-kNN-LSTM	0.895	0.842	1.466	0.322	1.053
	kNN-LSTM	0.920	0.868	1.557	0.283	1.201
	air2stream	0.715	0.741	2.941	0.528	2.208

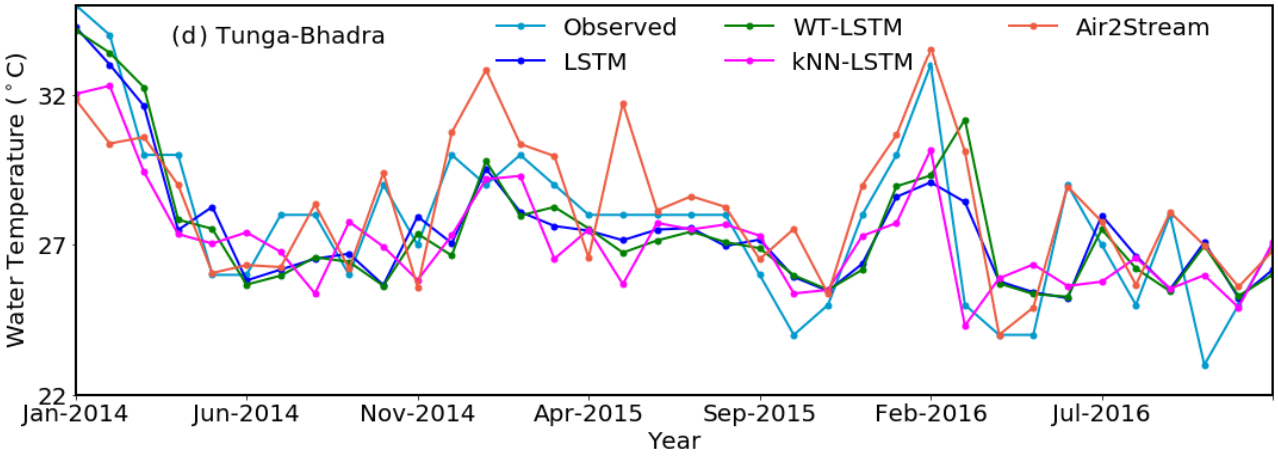
Table 4.6. Overview of the kNN-LSTM model performance with different monthly data points (values in parenthesis indicate the number of monthly data points used for simulations) for Musi and Ganga catchment stations based on air temperature (AT[t]), including time-lag effects of air and water temperatures (AT[t-1], RWT[t-1]) as input variables. The values displayed all referred to the testing period.

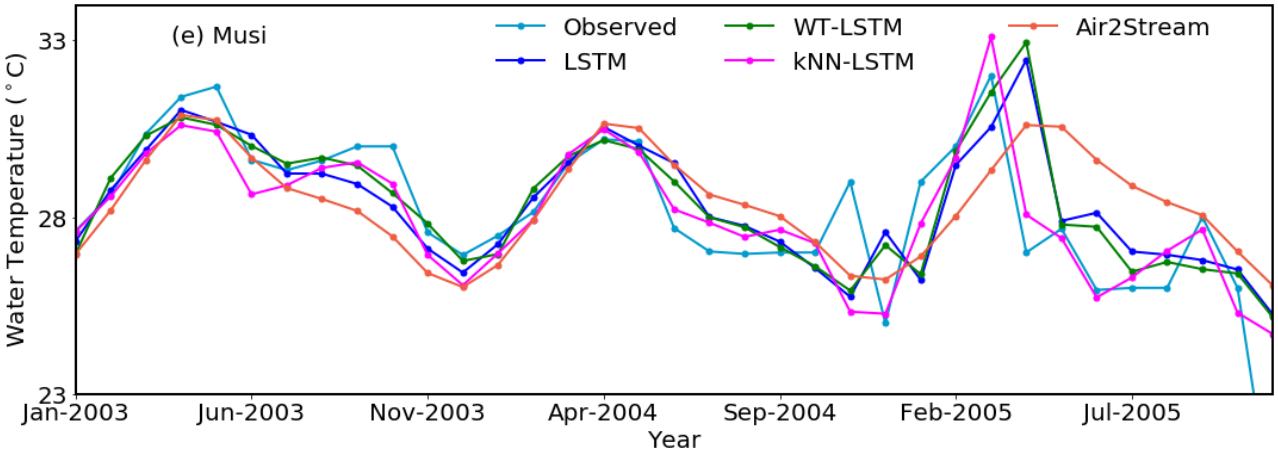
Catchment	**Model**	**NSE**	**KGE**	**RMSE**	**RSR**	**MAE**
	kNN-LSTM (176)	0.735	0.701	1.277	0.514	0.802
Musi	kNN-LSTM (152)	0.742	0.806	0.919	0.507	0.648
	kNN-LSTM (140)	0.737	0.835	0.944	0.512	0.659
	kNN-LSTM (180)	0.920	0.868	1.557	0.283	1.201
Ganga	kNN-LSTM (156)	0.940	0.974	1.004	0.225	0.726
	kNN-LSTM (144)	0.937	0.963	1.117	0.249	0.772

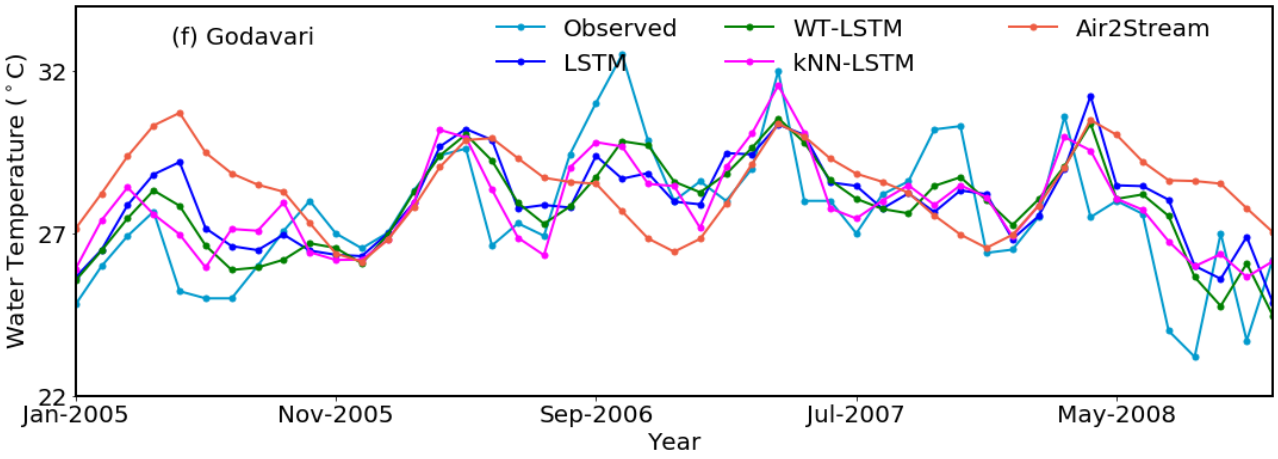

(a) Narmada
Observed
LSTM
WT-LSTM
kNN-LSTM
Air2Stream
Water Temperature (° C)
30
25
20
Jan-2004
Nov-2004
Sep-2005
Jul-2006
May-2007
Mar-2008
Year


(b) Cauvery
Observed
LSTM
WT-LSTM
kNN-LSTM
Air2Stream
Water Temperature (° C)
35
30
Jan-1996
Nov-1996
Sep-1997
Jul-1998
May-1999
Year


(c) Sabarmati
Observed
LSTM
WT-LSTM
kNN-LSTM
Air2Stream
Water Temperature (° C)
32
27
22
Jan-2005
Nov-2005
Sep-2006
Jul-2007
May-2008
Year


(d) Tunga-Bhadra
Observed
LSTM
WT-LSTM
kNN-LSTM
Air2Stream
Water Temperature (° C)
32
27
22
Jan-2014
Jun-2014
Nov-2014
Apr-2015
Sep-2015
Feb-2016
Jul-2016
Year


(e) Musi
Observed
LSTM
WT-LSTM
kNN-LSTM
Air2Stream
Water Temperature (° C)
33
28
23
Jan-2003
Jun-2003
Nov-2003
Apr-2004
Sep-2004
Feb-2005
Jul-2005
Year


(f) Godavari
Observed
LSTM
WT-LSTM
kNN-LSTM
Air2Stream
Water Temperature (° C)
32
27
22
Jan-2005
Nov-2005
Sep-2006
Jul-2007
May-2008
Year

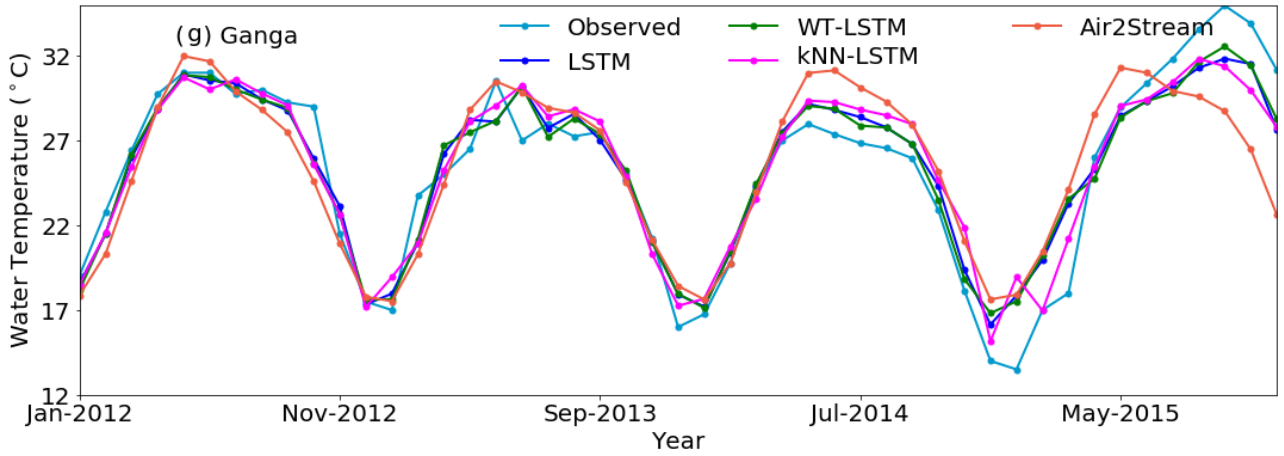


Figure 4.13. Comparison of the monthly observed values and LSTM, WT-LSTM, kNN-LSTM, and air2stream models predicted values of river water temperature (°C) for the seven catchments (a) Narmada (b) Cauvery (c) Sabarmati (d) Tunga-Bhadra (e) Musi (f) Godavari (g) Ganga during the testing phase.

Table 4.7. Overview of the air2stream reference model performances for each catchment. The shown values all refer to calibration and validation period.

Catchment	**Calibration**					**Validation**				
	NSE	**KGE**	**RMSE**	**RSR**	**MAE**	**NSE**	**KGE**	**RMSE**	**RSR**	**MAE**
Narmada	0.433	0.576	2.197	0.776	1.536	0.324	0.359	2.523	0.812	1.925
Cauvery	0.093	0.112	2.036	0.851	2.296	0.010	0.001	3.294	0.991	2.679
Sabarmati	0.428	0.654	2.088	0.757	1.612	0.072	0.335	2.601	0.958	1.946
Tunga-Bhadra	0.099	0.230	1.653	0.918	1.087	0.064	0.090	2.109	0.952	1.791
Musi	0.547	0.674	1.352	0.671	1.028	0.227	0.339	2.162	0.866	1.575
Godavari	0.244	0.453	2.108	0.867	1.609	0.107	0.325	2.432	1.183	1.943
Ganga	0.887	0.893	1.507	0.335	1.181	0.715	0.741	2.941	0.528	2.208

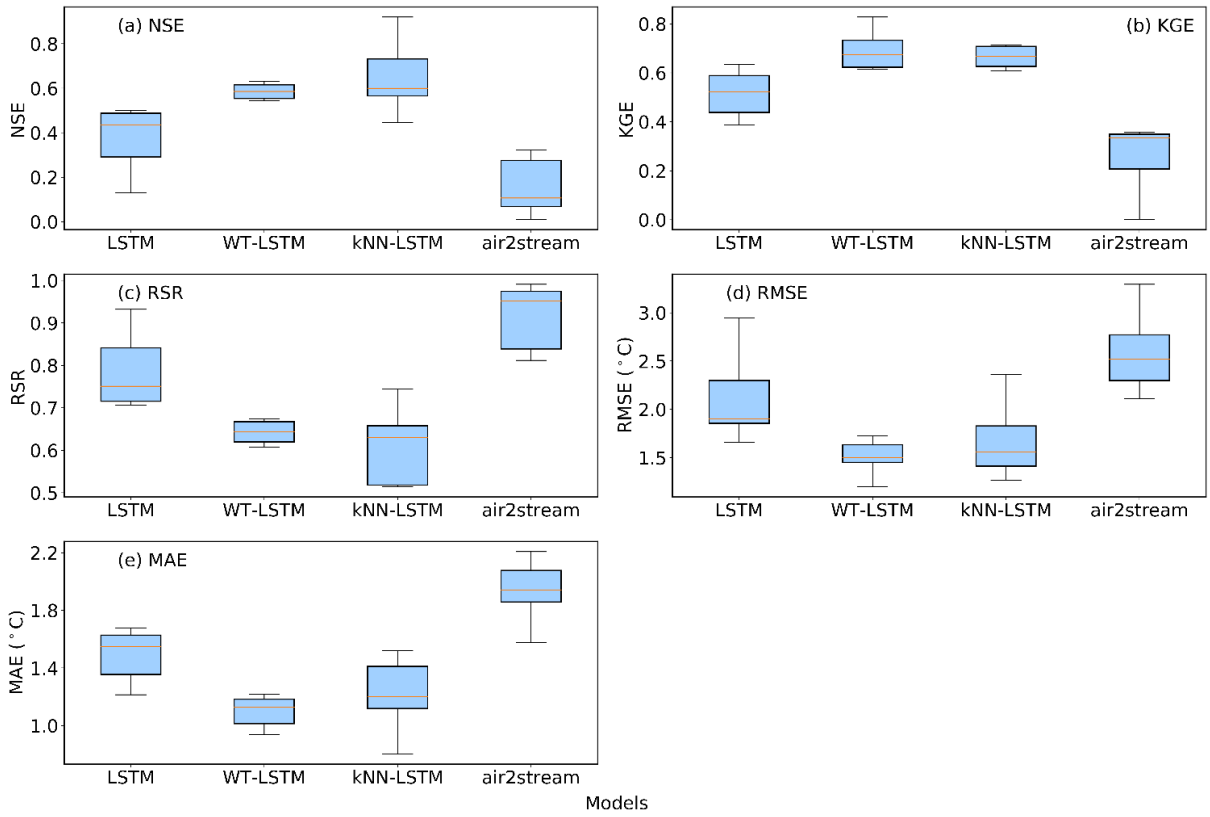


Figure 4.14. Boxplots of the NSE, KGE, RSR, RMSE, and MAE were based on seven catchments for the LSTM, WT-LSTM, kNN-LSTM, and air2stream models during the testing period.

Seasonality Aspects of Prediction:

The efficacy of the kNN-LSTM model with the seasonal month's data (summer, monsoon, and post-monsoon, and winter) (Laizé et al. 2017; Zhu et al. 2019a) experiments for seven catchments of India and experimental evidence produced that the kNN-LSTM model has performed poorly with the monsoon, post-monsoon, and winter seasonal aspects and shows good efficacy in predicting variability in summer RWTs for most of the catchment locations (Table 4.8). It can be noticed that R^2, KGE, RMSE, RSR, NSE and MAE values were poor for the monsoon, post-monsoon and winter for seven catchments when compared to the summer season (NSE = 0.183 – 0.801, KGE=0.221 – 0.876, RMSE=0.819 – 2.997, RSR=0.379 - 0.903, and MAE=0.605 – 2.515). These seasonal predictions followed the historical data trend (Table 4.2), i.e., most of the catchments had positively correlated in the summer months. This study used the Autoregressive Integrated Moving Average Model

(ARIMA) model to measure the RWT performance under seasonal variations for seven catchments based on the previous lag value to validate the seasonality. First, the RWT data's autocorrelation and partial autocorrelation functions (ACF and PACF) were examined to account for the time-lag information. These functions suggest that the one-month lag is significant in the observed RWT record. In an ARIMA model, the first component was the autoregressive (AR) term, the second component was the integration (I) term, which was responsible for making the data stationary, and the third component was the moving average (MA) term of the forecast errors. A standard notation is used in ARIMA (p, d, q), where the parameters are substituted with integer values. The parameters of the ARIMA model are defined as follows:

- p: The number of lag observations included in the model, also called the lag order.
- d: The number of times the raw observations are differenced, also called the degree of difference.
- q: The size of the moving average window, also called the order of moving average.

The present study used order as ARIMA (1,1,0). The ARIMA model programming was done in Python.

From the ARIMA model validation results (Table 4.9), it is observed that the ARIMA model performed poorly for all seasons for seven catchments of India. Overall, results from Table 4.8 revealed that the kNN-LSTM yielded improved results in the summer season for most of the catchment sites than the other three seasons. It can be noted that accurate summer water temperature predictions are more relevant for Indian river systems due to extremely low water quality levels under minimum stream flows because of non-monsoon seasons (Rehana and Dhanya 2018).

Climate Change Signals:

Figure 4.15 displays the box plots projected RWT (◦C) values that show a range of increases under the RCP 8.5 scenario compared to RCP 4.5 for 2021-2050 and 2071–2100. Figure 4.16 displays the box plots of the ensemble mean RCP 4.5 and 8.5 experiments projected RWT (◦C) values for the period of 2021-2050 and 2071–2100 with respect to historical values for seven catchments of India. According to Figures 4.15, and 4.16, RWT increases for the periods 2021-2050 and 2071–2100 due to an increase in AT (Table 4.10). Projected mean RWT changes for the periods 2071–2100 relative to mean observed values were calculated

using RCP 4.5 and 8.5 output data (Table 4.11) and observed that results vary between the different catchments. The magnitude of RWT increases is higher for Narmada, Musi, and Ganga catchment sites and variations in mean water temperatures (dTw_{mean}) for 2071–2100 relative to the observed mean values noted 4.7, 4.0, and 4.6 ◦C, respectively (Table 4.11). For Narmada, Musi, and Ganga catchments, the RWT rises were discovered in high magnitudes (95^{th} percentile in monthly distribution; dTw_{95}) for 2071–2100 relative to observed high (95^{th} percentile) values noted as 4.0, 3.4, and 4.3 ◦C, respectively. Moderate dTw_{mean} for 2071–2100 are projected for catchments in the southern parts of India relative to the observed mean, noted as 0.8 ◦C for Cauvery and 1.8 ◦C for Godavari (Table 4.11). Overall, results indicated that water temperatures over Indian catchments would likely to rise by more than 3.0 °C (0.8 – 4.7 °C) for 2071-2100 (Table 4.11).

Table 4.8. Overview of kNN-LSTM models performance under seasonal variations for seven catchments based on air temperature, including time-lag effects as an input variable. The shown values all refer to the testing phase.

Catchment	Season	NSE	KGE	RMSE	RSR	MAE
Narmada	Summer	0.677	0.753	1.736	0.567	1.349
	Monsoon	0.456	0.521	1.616	0.737	1.245
	Post-monsoon	0.537	0.397	1.615	0.680	1.242
	Winter	0.547	0.578	1.208	0.672	1.002
Cauvery	Summer	0.183	0.221	2.997	0.903	2.515
	Monsoon	0.446	0.469	1.915	0.744	1.608
	Post-monsoon	0.316	0.399	1.234	0.826	0.956
	Winter	0.394	0.306	2.752	0.778	2.219
Sabarmati	Summer	0.675	0.608	1.805	0.569	1.473
	Monsoon	0.359	0.529	1.634	0.800	1.169
	Post-monsoon	0.012	0.582	1.532	0.978	1.150
	Winter	0.587	0.504	2.372	0.641	1.421
Tunga-Bhadra	Summer	0.339	0.839	2.099	0.812	1.840
	Monsoon	0.215	0.197	1.848	0.885	1.658
	Post-monsoon	0.039	0.017	1.298	0.855	1.237
	Winter	0.083	0.055	1.864	0.940	1.416

Musi	Summer	0.855	0.696	0.819	0.379	0.605
	Monsoon	0.864	0.744	0.668	0.369	0.569
	Post-monsoon	0.825	0.870	0.634	0.411	0.577
	Winter	0.355	0.293	2.327	0.802	1.542
Godavari	Summer	0.609	0.733	1.020	0.633	0.843
	Monsoon	0.238	0.531	1.446	0.877	1.112
	Post-monsoon	0.576	0.554	1.671	0.650	1.440
	Winter	0.541	0.564	1.065	0.677	0.881
Ganga	Summer	0.801	0.876	1.642	0.445	1.311
	Monsoon	0.400	0.407	1.546	0.774	1.275
	Post-monsoon	0.628	0.596	2.392	0.610	1.804
	Winter	0.765	0.660	2.235	0.484	1.815

Table 4.9. Overview of Autoregressive Integrated Moving Average Model (ARIMA) model performance under seasonal variations for seven catchments based on the previous one-month lag value. The shown values all refer to the testing phase.

Catchment	**Season**	**NSE**	**KGE**	**RMSE**	**RSR**	**MAE**
Narmada	Summer	-0.821	0.368	4.121	1.346	2.964
	Monsoon	-0.732	0.301	2.884	1.315	1.712
	Post-monsoon	0. 190	0.487	2.136	0.899	1.626
	Winter	-0.018	0.347	2.126	1.009	1.674
Cauvery	Summer	0.370	0.566	2.631	0.791	2.394
	Monsoon	-0.509	0.518	3.162	1.225	2.261
	Post-monsoon	-0.624	-0.159	1.902	1.274	1.057
	Winter	-0.189	0.126	3.711	1.090	2.589
Sabarmati	Summer	0.185	0.637	2.859	0.902	2.403
	Monsoon	-0.191	0.492	2.229	1.091	1.818
	Post-monsoon	-0.402	-0.092	1.542	1.184	1.204
	Winter	0.251	0.272	3.063	0.865	2.137
Tunga-Bhadra	Summer	0.308	0.667	2.200	0.831	1.939
	Monsoon	-1.422	0.282	3.247	1.556	2.275
	Post-monsoon	-0.508	-0.072	1.357	1.228	0.978

	Winter	-0.791	0.019	2.397	1.338	2.144
Musi	Summer	-1.473	-0.164	3.385	1.572	2.191
	Monsoon	0.638	0.847	1.088	0.600	0.670
	Post-monsoon	0.420	0.698	1.163	0.761	0.744
	Winter	-0.374	-0.099	3.206	1.172	2.409
Godavari	Summer	-0.205	0.543	1.755	1.097	1.487
	Monsoon	-0.179	0.530	1.800	1.086	1.345
	Post-monsoon	0.364	0.719	2.048	0.797	1.844
	Winter	0.082	0.575	1.638	0.958	1.217
Ganga	Summer	0.009	0.508	3.672	0.995	2.914
	Monsoon	0.028	0.566	1.969	0.985	1.372
	Post-monsoon	0.842	0.859	1.558	0.395	1.231
	Winter	0.521	0.664	3.059	0.691	2.554

Table 4.10. Projected changes in the mean air temperatures for 2071–2100 with mean of Representative Concentration Pathway (RCP) 4.5 and 8.5 experiments for seven catchments relative to historical values.

Catchment	**AT (historical) (°C)**	**AT (2071-2100) (°C)**	**Changes in AT (°C)**
Narmada	25.09	29.56	4.5
Cauvery	28.81	31.20	2.4
Sabarmati	26.72	30.85	4.1
Tunga-Bhadra	24.24	30.03	5.8
Musi	28.13	31.71	3.6
Godavari	27.48	30.65	3.1
Ganga	25.71	29.76	4.1

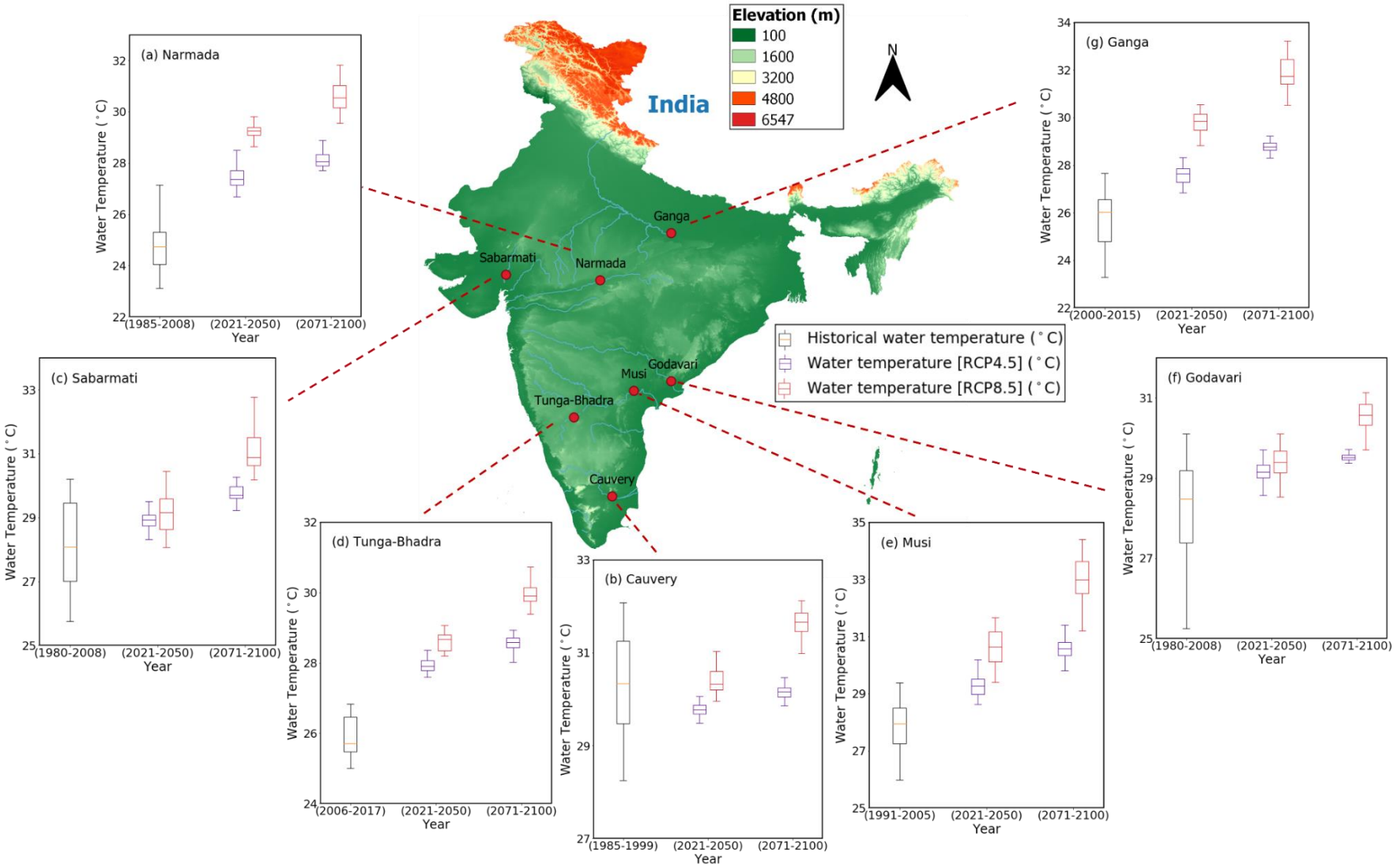


Figure 4.15. Boxplots represent the Representative Concentration Pathway (RCP) 4.5 and 8.5 experiments projected river water temperature (°C) values for the periods 2021-2050 and 2071–2100 with respect to historical values for seven Indian catchments.

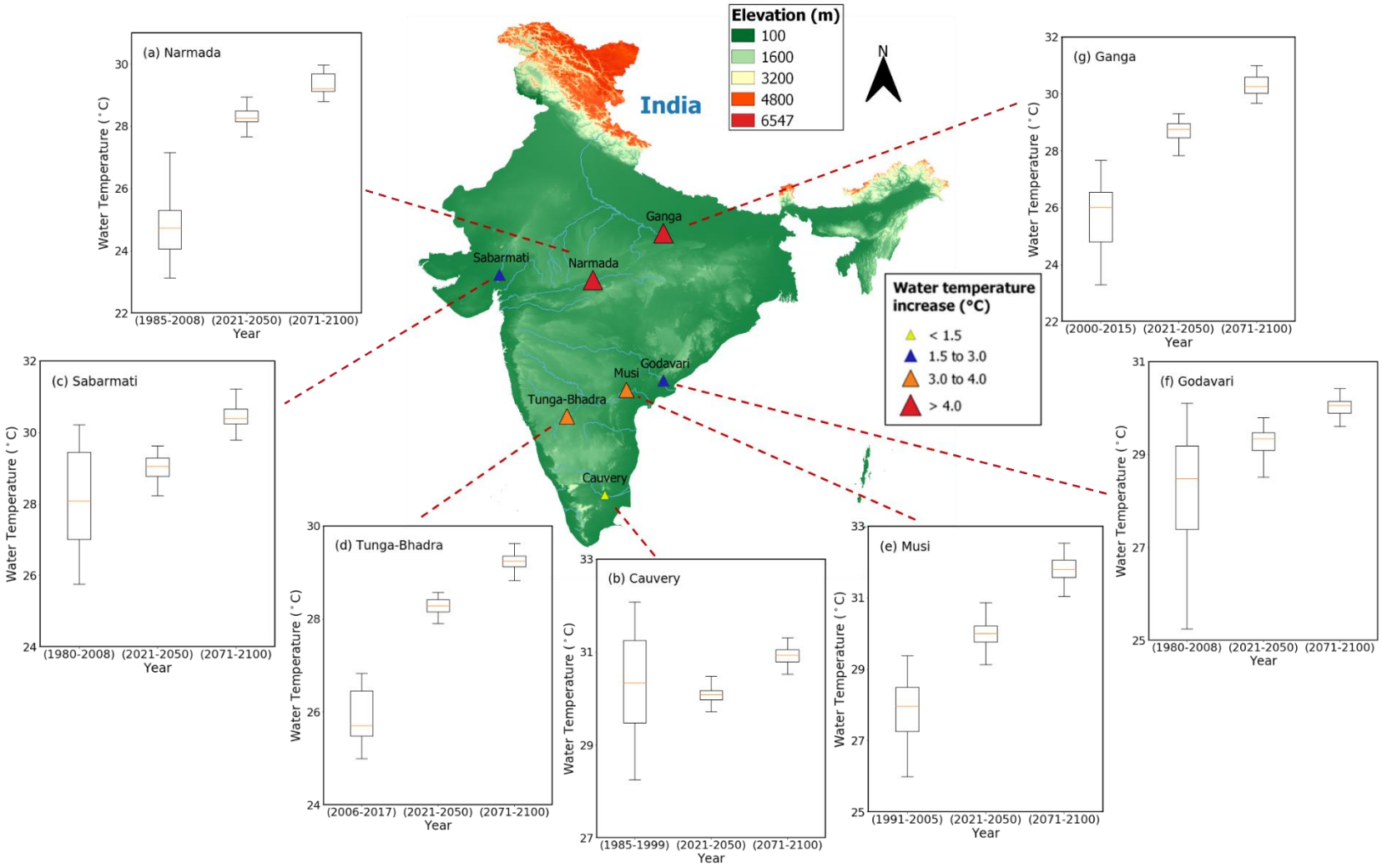


Figure 4.16. Boxplots represent the ensemble mean of Representative Concentration Pathway (RCP) 4.5 and 8.5 experiments projected river water temperature (°C) values for the period 2021-2050, and 2071–2100 with respect to historical values for seven Indian catchments. Triangle sizes represent the magnitude of the river water temperature increase (°C) for 2071-2100.

Table 4.11. Projected changes in the annual mean (dTw_{mean}) and high (95^{th} percentile; dTw_{95}) river water temperatures for 2071–2100 with Representative Concentration Pathway (RCP) 8.5 experiments for seven catchments relative to observed values. Values in parenthesis indicate RCP 4.5 experiments.

Catchment	**Tw_{mean} (Observed mean) (°C)**	**Tw_{mean} (2071-2100) (°C)**	**dTw_{mean} (°C)**	**dTwmean (°C) (Mean of RCP4.5 & RCP8.5)**	**Tw_{95} (Observed high) (°C)**	**Tw_{95} (2071-2100) (°C)**	**dTw_{95} (°C)**	**dTw_{95} (°C) (Mean of RCP4.5 & RCP8.5)**
Narmada	24.7	30.6 (28.2)	5.9 (3.5)	4.7	26.3	31.6 (28.9)	5.3 (2.6)	4.0
Cauvery	30.3	31.7 (30.4)	1.4 (0.1)	0.8	31.4	32.1 (31.5)	0.7 (0.1)	0.4
Sabarmati	28.2	31.1 (29.8)	2.9 (1.6)	2.3	30.1	32.2 (30.2)	2.1 (0.1)	1.1
Tunga-Bhadra	26.3	29.9 (28.6)	3.6 (2.3)	3.0	28.9	30.6 (29.0)	1.7 (0.1)	0.9
Musi	27.9	33.1 (30.6)	5.2 (2.7)	4.0	29.2	34.1 (31.1)	4.9 (1.9)	3.4
Godavari	28.2	30.5 (29.5)	2.3 (1.3)	1.8	29.8	31.1 (29.9)	1.3 (0.1)	0.7
Ganga	25.7	31.8 (28.8)	6.1 (3.1)	4.6	26.9	32.9 (29.4)	6.0 (2.5)	4.3

4.4.3. Granger Causality Results

The present chapter used lag-1 time series data of AT as input to fit the LSTM models. It was confirmed that predicting the RWT from changes in AT yields better results when done based on lag-1 data (Table 4.4). To validate these results, this study used the GCA to measure the causal linkages (Table 4.12). The Granger causality test is performed on one of the water quality indicators (RWT), which is considered the effects, and the meteorological variables (AT) are considered as possible causes. The null hypothesis is defined as follows: there is no Granger causality between the cause and effect. Thus, lower p-values correspond to stronger causality and vice versa. The results unambiguously show uni-directional causal linkages between AT and RWT for all the catchments (i.e., AT causes RWT), and bi-directional linkage was noted for the Ganga catchment (i.e., both AT causes RWT, and vice versa also) (Table 4.12, values that are significant at 95% confidence level are in boldface). Specifically, it is confirmed that the former, especially AT, is the main causal driver of rising temperatures. Table 4.12 shows p-values for the Granger causality test when RWT is considered the effect. If a given p-value is less than significance level (0.05), for example, take the value 0.00 in (Narmada: row 2, column 1), the null hypothesis can be rejected (i.e., AT does not Granger-cause RWT is accepted if and only if no lagged values of AT are retained in the regression equation) and conclude that AT Granger causes RWT. Likewise, the 0.69 in (Narmada: row 1, column 2) refers to the insignificant relationship of RWT causes AT. For AT and RWT, the Granger causality relationship is significant for basins. After analyzing the scenarios, it is concluded that the strongest relationships (i.e., smallest p-values, AT causes RWT) are observed for all basins.

Table 4.12. The Granger causality test p-values among the air temperature and water temperature. The values that are significant at 95% confidence level are in boldface (lower p-values correspond to stronger causality and vice versa).

Catchment	Dependent variables	Source of causality	
		Air temperature	**Water temperature**
Narmada	Air temperature	1.00	0.69
	Water temperature	**0.00**	1.00
Cauvery	Air temperature	1.00	0.27
	Water temperature	**0.07**	1.00
Sabarmati	Air temperature	1.00	0.28
	Water temperature	**0.00**	1.00
Tunga-Bhadra	Air temperature	1.00	0.16
	Water temperature	**0.01**	1.00
Musi	Air temperature	1.00	0.47
	Water temperature	**0.00**	1.00
Godavari	Air temperature	1.00	0.75
	Water temperature	**0.00**	1.00
Ganga	Air temperature	1.00	**0.00**
	Water temperature	**0.00**	1.00

4.5. Discussion

This chapter presents new intuitions on the assessment of RWQ variables prediction and the impact of climate change on water quality. As RWT is a primary variable that influence water quality, in this chapter RWT variable was selected for prediction based on AT, including time-lag effects for seven different catchment sites across India in different physiographic settings using LSTM, WT-LSTM, kNN-LSTM, and air2stream models, demonstrates the improved forecasting accuracy with hybrid kNN-LSTM model. The monthly NSE scores ranged between 0.132–0.920, KGE scores ranged between 0.131–0.868, RSR scores ranged between 0.281–0.744, and RMSE scores were ≤ 3 °C during the testing periods, revealing high model reliability. The predicted RWT variability matched observations well; thereby, the DL model's output can be trusted. All the developed model

statistical metrics covered the range of model reliability described in the literature. The RMSE scores for all the catchments ranged between 1.199 - 3.294 °C pertaining to all DL models (Table 4.4) for monthly data, which are reasonable in comparison to earlier models of the Spatio-temporal approach by Jackson et al. (2018) (1.570 °C); Bayesian regression approach by Sohrabi et al. (2017) (1.250 °C); random forest (RF), ANN, RNN by Feigl et al. (2021) (0.422 – 0.815 °C); Wavelets-ANN by Graf et al. (2019) (0.981-1.434 °C); LSTM by Stajkowski et al. (2020) (0.755 °C); LSTM by Qiu et al. (2021) (0.500 – 2.700 °C); and ANN by Temizyurek and Dadaser-Celik (2018) (2.100 – 2.640 °C); River Assessment for Forecasting Temperature (RAFT) model by Pike et al. (2013) (0.500 °C); and NorWeST Summer Stream Temperature model by Isaak et al. (2017) (1.100 °C). The MAE values for all the catchments range from 0.802 to 2.467 °C pertaining to all DL models (Table 4.4) for monthly data, which are reasonable in comparison to earlier models of RF, ANN, RNN by Feigl et al. (2021) (0.329 – 0.675 °C); Wavelets-ANN by Graf et al. (2019) (0.781-1.286 °C); and LSTM by Qiu et al. (2021) (0.39 – 2.15 °C). The KGE values for Narmada (0.715), Tunga-Bhadra (0.790), Musi (0.701), and Ganga (0.868) catchments pertaining to all DL models for monthly data (Table 4.4), which are in comparison to the earlier model of LSTM by Stajkowski et al. (2020) (0.923). The NSE values for Narmada (0.728), Musi (0.735), and Ganga (0.920) catchments pertaining to all DL models for monthly data (Table 4.4), which are sensible in comparison with the earlier model of LSTM by Qiu et al. (2021) (0.74 - 0.99). The superiority of LSTM in RWT prediction, as demonstrated in this work, was found to agree with Feigl et al. (2021), Qiu et al. (2021), and Stajkowski et al. (2020). The input and output of LSTM are considered as a two-time-series sequence, which is why LSTM outperforms the other ML models (Qiu et al. 2021). However, it can be noted that the study was conducted by Feigl et al. (2021) used AT, runoff, precipitation, and global radiation values as input in the RWT prediction for daily data, and the study by Qiu *et al.* (2021) used daily AT and discharge as input in RWT prediction. As demonstrated in this work, the superiority of hybrid LSTM models in RWT prediction was found to agree with the analysis of Stajkowski *et al.* (2020) where the author has adopted a hybrid approach i.e., genetic-algorithm (GA)-optimized LSTM technique (GA-LSTM) to improve the model performance. However, it can be noted that the study by Stajkowski *et al.* (2020) used AT values as input in RWT prediction for hourly data. Besides the similar studies that have been done for the

prediction of RWT (Stajkowski et al. 2020; Feigl et al. 2021; Qiu et al. 2021), this study showed that the recent advent of so-called hybrid models, which entails combining WT, k-NN bootstrap resampling algorithm with DL methods could be applied for RWT prediction under data noisy and limited data scenarios. During this study, the WT-LSTM and kNN-LSTM model's performance was better when modeling monthly data, demonstrating that coupling with WT and k-NN bootstrap resampling algorithm would yield more precise results and guaranteed results to outperform standalone models (Table 4.4, Figure 4.14).

The reason why the WT-LSTM combination performed better might be that the WT method offered useful decompositions of the original AT and RWT time series, and the transformed data improved the performance of the WT-LSTM model by analysing useful information on various decomposition levels. The reason why the kNN-LSTM combination performed better might be that the kNN method simulates extreme events beyond those observed in the short length of the historical record in the data-sparse regions. In order to extract temporal characteristic information from long time series data, the kNN-LSTM model is trained using a large amount of simulated data. This might be another factor in the superior performance of the kNN-LSTM model.

This chapter confirmed that the lag variables had a strong relationship with RWT and improved the model performance (Tables 4.3, and 4.4). This chapter also confirmed that the kNN-LSTM model yielded improved results in the summer season for most of the catchment sites than the other three seasons when compared to Autoregressive Model (AR) model. The 3-parameter version air2stream model delivered the lowest monthly performance for almost every case (Table 4.4, Figure 4.13). It can be noted that the performance of air2stream depends on the time scale and data points used in the calibration and validation. Most of the earlier studies successfully applied the air2stream in a wide range of hydrological studies over a range of catchments with dense data at daily time scale and generally had an improved performance compared to ML models (Piccolroaz et al. 2016; Yang and Peterson 2017; Piotrowski and Napiorkowski 2018, 2019; Zhu et al. 2019f; Tavares et al. 2020), whereas the present study emphasized on sparse and discontinued data at monthly time scale. These are some reasons why the 3-parameter version air2stream model performed poorly.

This study also showed that using decomposed time series data instead of original time series data to simulate the values from the k-NN bootstrap resampling algorithm was not

producing good performant results like the kNN-LSTM model (Table 4.5). As demonstrated in this work, the wavelet power spectrum (Figure 4.12) displayed a consistent 1-year periodicity for the whole duration of the time series and it was found to agree with the analysis of Alcocer et al. (2022) where the author performed the Continuous wavelet analysis to examine the variability of water quality variables and concluded that the water quality variables display a recurrent annual cycle. This steady behavior means that the system's complexity has not deteriorated so that the seasonality of this process has not perceptibly changed over time. In this study, the hybrid WT- LSTM and kNN-LSTM models were shown to be a solution for improved predictions of RWT for most of the catchments, i.e., Sabarmati, Tunga-Bhadra, Musi, Godavari, and Ganga. Overall, the kNN-LSTM model produced more accurate results for the prediction of RWT. Also observed that the Ganga catchment site (Pratappur) exhibits dampened warming trends of RWT comparative to other catchment sites, probably because of the influence of Himalayan mountains glacier-melt water that modulates downstream RWTs in this system (Islam et al. 2019). Such a hypothesis was found to agree with the investigation of Zhu et al. (2019), i.e., flow discharge with high altitude catchment sites influenced by cold water releases from snow melting or hydropower significantly impacts the RWT dynamics. It can also be noted that the Tunga-Bhadra and Musi rivers are the Southern Indian rain-fed River basins dominated by a semi-arid climate with hot summers and temperate winters with more pronounced summer RWT. The prediction of RWT can consider variables such as flow discharge, wind speed, radiation, etc. to analyze the variation in water temperatures in different rivers with varied climatological conditions (van Vliet et al. 2013; Cole et al. 2014; Feigl et al. 2021). Consideration of various hydroclimatic covariates in the RWT predictions can lead to variations in the RWT predictions for different rivers. For example, for snow-fed rivers, the flow discharge will be prominent covariate since snowmelt flow dominates RWT predictions, with more the snow-fed discharges less the RWT. Radiation and evaporation will be important covariates in the case of semi-arid climate rivers as the sun heats the water much faster as the water depth is lower due to frequent low flow events under evaporation (Arnell 1996).

It was also observed that inherent uncertainties from each of the DL models can accumulate and can affect the final performance measures. Such uncertainties can originate from various sources, starting from the noise, covariates considered, temporal discontinuity

present in the original water quality sampled observations to the model parameters, type of DL algorithm used to predict RWT, and non-stationarity assumptions related to forecasting model parameters (Beven 2016)**.** Furthermore, each DL model can predict unique RWT, leading to model uncertainty. To address such model uncertainties originating from various DL, the ensemble of DL models, stacking algorithms, etc. (Song et al. 2020; Piotrowski et al. 2021) can be adopted which can combine RWT predictions from various DL models. Furthermore, such ensemble/stacking algorithms allow the decision makers to choose the best possible prediction within a range of predictions (Rehana & Mujumdar, 2014).

Assessment of Projected Changes in Monthly River Water Temperature:

Forcing the kNN-LSTM hybrid model with monthly RCP 4.5 and 8.5 output provides for calculating probable monthly RWT changes over the whole probability range, rather than mean values. These projected change patterns are most consistent with earlier hydrological model studies (Döll & Zhang, 2010; Rehana, 2019; van Vliet et al., 2013). van Vliet *et al.* (2013) found the highest increase in mean RWT projected for river basins in Australia, Europe, Southeast Asia, South Africa, and the United States. Rehana (2019) evaluated a statistical downscaling model based on Canonical Correlation Analysis (CCA) for future RWT projections along the Tunga-Bhadra River, India, and found that the river's annual RWT increase from 2020–2040 to 2081–2100 is predicted to be 3.2 °C. In this study, overall, for all seven catchments, the increases in mean (95^{th} percentile) water temperature are 0.1–3.5 (0.1–2.6) ^{o}C for RCP 4.5, and 1.4–6.1 (0.7–6.0) ^{o}C for RCP 8.5 for plausible future (2071–2100) relative to observed mean (95th percentile) water temperatures (Table 4.11). This increase appears to be modest in comparison to projected rise of mean AT of 2.4-5.8 ^{o}C for the RCP 4.5 and 8.5 scenarios over seven Indian catchments (Table 4.10). Such projections are in convincing with the global mean AT of 3.0–4.9 ^{o}C for the chosen GCM experiments done by van Vliet *et al.* (2013) and the annual mean temperatures of 1.8 and 3.2 °C for the RCP 4.5 and 8.5 scenarios over Southeast Asia done by Raghavan *et al.* (2018). RCP 8.5, which is generally taken as the basis for worst-case climate change scenarios (i.e., high CO_2 concentrations with radiative forcing greater than 8.5 W m^{-2} by 2100).

Though the hybrid kNN-LSTM model performed well, further research is needed to improve it. Despite the effectiveness of the modeling frameworks, as demonstrated in work,

it has some limitations. Based on earlier research, rainfall is an important covariate in RWT predictions (Cole et al. 2014; Feigl et al. 2021). In this study, RWT was weakly correlated with precipitation for most of the catchment stations; hence precipitation has not been considered in the present study. Flow discharge, which is a smoothened force caused by rainfall, may play a vital role in RWT predictions (van Vliet et al. 2011; Zhu et al. 2019f; Feigl et al. 2021), especially in Indian rivers impacted by low flows during summer seasons. However, flow discharge was not examined in this work due to a lack of complete streamflow data. Therefore, the hybrid modeling framework in future research will be enhanced by integrating flow discharge as model input for rivers. RWT is directly influenced by multiple parameters, including streamflow (Toffolon and Piccolroaz 2015; Sohrabi et al. 2017; Islam et al. 2019), river geometry, groundwater inputs, slope, water depth, etc. (Gu and Li 2002), which are not considered in the present study.

Further enhancement of RWT prediction can be examined by using other types of WT coupled with DL algorithms. Lastly, though seven catchments with diverse geographical features were evaluated in the current work, they were all located in India. Therefore, presenting broad conclusions about the efficacy of hybrid WT-LSTM, kNN-LSTM models, as well as their superiority over standalone LSTM and air2stream models for any river around the world, is not possible. Despite this, the research offers vital intuitions about the historical and projected thermal states of seven Indian river catchment locations, which may be beneficial in creating future water management plans that may impact aquatic resources.

4.6. Chapter Summary

This chapter proposed a suite of LSTM models by coupling with WT and k-NN bootstrap resampling algorithms to predict the RWQ variables under data uncertainties for the Indian catchments. To demonstrate this framework, RWT is taken as water quality variable for prediction with the aid of current AT and lag variables as predictors at a monthly timescale. In this study, LSTM, WT-LSTM, and kNN-LSTM models were proposed to better predict water quality variables. The developed model's robustness was compared with the traditional air2stream model for RWQ variables prediction at seven river gauging stations located in Indian catchments characterized by different hydrological conditions. The impacts of climate change on RWQ variables were evaluated over Indian catchments using the kNN-

LSTM monthly model forced by an ensemble of RCP scenarios 4.5 and 8.5 down-scaled projections of AT data from NEX-GDDP. Also, validated whether one variable time series at time t-lag provides important information helping to predict values of another variable time series at time t by using the Granger Causality Analysis test. The results lead to the following conclusions:

1. When WT and k-NN bootstrap resampling algorithms were included, LSTM outperformed the conventional models; hence these hybrid models are the new promising frameworks for RWT prediction under data-sparse regions and may deserve further research in the area of water resources.
2. The hybrid kNN-LSTM models yielded better performance results for five catchment sites (i.e., Narmada, Cauvery, Musi, Godavari, and Ganga) out of seven catchment sites than LSTM, WT-LSTM, and air2stream forecasting models at a monthly scale. This study confirmed that WT-based models consistently outperformed standalone models and demonstrated that lag variables are significantly related to RWT and improved the model performance.
3. The widely used process-based model air2stream is used to validate the proposed ML models and to make the results comparable to previous studies. It was found that the 3-parameter version of air2stream mainly delivered the lowest performance compared to LSTM, WT-LSTM, and kNN-LSTM models for almost every basin.
4. Validated the time sequence of the causal linkages between the time series of data (i.e., whether cause precedes the effect relations for “water–air” and “air–water” directions of influence at a monthly scale). And observed that strongest “air–water” relationships (i.e., smallest p-values, AT causes RWT) for all basins.
5. Higher RWTs were predicted for the Ganga catchment, where climate change will decrease glacial ice in the Himalayan mountains, the source of the Ganga; this will result in even lower water levels in the river over time.
6. Also observed that the Ganga catchment site (Pratappur) exhibits dampened warming trends of RWT compared to other catchment sites. This is probably because of the influence of the Himalayan mountain's glacier-melt water, which modulates the downstream RWT in this system.

7. Overall, RWT over Indian catchments is likely to rise by more than 3.0 °C for 2071-2100.

In summary, in this chapter developed a suite of LSTM models by coupling with WT and k-NN bootstrap resampling to predict accurate RWQ variables under sparse and non-stationary data scenarios. Further assessed the climate change impacts on RWT by using RCP scenarios 4.5 and 8.5 scenarios from the NEX-GDDP dataset. The proposed methods, demonstrated methodologies, and frameworks presented in this chapter are generic and can be implementable for any river water quality variables. Further research is required to assess the impact of climate change on DO saturation levels with respect to RWT and streamflow under sparse river water quality data scenarios and were presented in the following chapter.

Chapter 5
IMPACT OF CLIMATE CHANGE ON SATURATED DISSOLVED OXYGEN OVER MAJOR INDIAN RIVER BASINS

5.1. Introduction

Studies on regional and global climate change, variability, and their impacts on water resources have received a lot of attention recently, but few studies concentrated on the expected changes in river water quality that will occur as a result of climate change. The relationship between the climate and freshwater systems is strong, and numerous climatic variables, including AT and RWT, precipitation, and the frequency of extreme events, have an impact on river water quality. The RWT and river flow, which are the main factors influencing RWQ, can be affected under the increase of AT and changes in rainfall variability, respectively. Climate change caused by anthropogenic greenhouse gases in the atmosphere directly impacts the quality of river water, which raises the possibility of the river ecosystem degrading in terms of decreased DO saturation levels under the decrease of stream flows and increase in RWT. Therefore, it is crucial to research how climate change will affect the thermal processes (e.g., RWT) and other self-purification capacity defining variables, such as the saturated DO of the river system. The literature related to the assessment of water quality impacts due to climate change is reviewed in Chapter 2. The review reveals that relatively very less attention has been given to assessing the impacts of regional and global climate change and variabilities on river water quality.

The present chapter aims to assess the climate change impacts on the thermal regimes of rivers in India and possible variability in DO saturation levels under RWT projections using the state-of-the-art GCM projections and hypothetical climate change scenarios. Saturation DO is generally considered as a desirable level of DO by the Pollution Control Boards (PCBs) in Waste Load Allocation Models (WLAM) for river water quality management (Mujumdar and Subbarao Vemula 2004). Therefore, the study of climate change impacts on saturation DO levels under climate change can provide prominent insights for defining/altering the quality standards under climate change. Climate change has been demonstrated to have an impact on the relationship between RWT and DO concentrations in tropical rivers (Danladi Bello et al. 2017). Tropical rivers receive more solar radiation and have higher RWTs (Taniwaki et al. 2017). For example, Indian tropical river systems

experience the highest RWTs during low flow periods of non-monsoon and summer months (Rehana & Dhanya, 2018a; Santy et al., 2020). Seasonality plays a vital role in the Indian river systems as maintaining flows in the summer season is a challenge leading to water quality deteriorations.

Recently, several studies have assessed the impact of climate change on RWQ variables using process-based models (e.g., CEQUEAU model, QUAL2K model, SWAT model, Air2Stream model etc.) (Ficklin et al., 2013; Islam et al., 2019; Khorsandi et al., 2023; Rehana & Mujumdar, 2011; van Vliet et al., 2011), regression based models (Santy et al. 2020), and ML models (Rehana, 2019; Zhu, Nyarko, et al., 2019). A better RWT model, instead of the regression models or process-based models used, is likely to give more accurate results (Santy et al. 2020). However, there hasn't been much research done on how climate change affects DO saturation levels with respective to RWT under sparse data, and the role of streamflow in ML models has seldom been investigated (Rabi et al. 2015; Zhu et al. 2019d). To this end, the assessment of DO saturation rates with respect to RWT is of much relevance for Indian river systems due to minimum flows and higher temperatures during non-monsoon seasons.

The present chapter aims to predict the RWQ variables and further quantify the climate change impact on water quality for major Indian catchments under data limitations. As the hybrid kNN-LSTM model based on AT performed well to predict RWT in Chapter 4, this chapter studied the pertinence of the k-NN algorithm under sparse data scenarios to predict the RWT by including streamflow time-series data in addition to AT, and time lag was adopted. The study compared this extended hybrid kNN-LSTM monthly model with standalone LSTM, a modified nonlinear regression model proposed by van Vliet et al. (2011), and an 8-parameter version air2stream model (Toffolon and Piccolroaz 2015; Piccolroaz et al. 2016). Air2stream is a hybrid model for predicting RWT, with AT and streamflow data used as model inputs. Air2stream was widely applied in multiple studies over a range of catchments (Toffolon and Piccolroaz 2015; Piotrowski and Napiorkowski 2019; Islam et al. 2019; Feigl et al. 2021; Yang et al. 2022; Shrestha and Pesklevits 2023). Furthermore, the study evaluated the effect of climate change on DO saturation levels with respect to RWT and streamflow using the kNN-LSTM model forced with climate change scenarios.

In summary, the objectives of this chapter are to (i) assess the combined effects of streamflow and AT in ML models for prediction of RWT variables under sparse data scenarios, (ii) compare the performance results of the kNN-LSTM model with standalone LSTM, a modified nonlinear regression model proposed by van Vliet et al. (2011), and 8-parameter version air2stream in the prediction of RWT variables when applied on major river systems of India, (iii) to calculate the impacts of climate change on riverine thermal processes and possible variability in DO saturation levels with respect to RWT by using the kNN-LSTM model addressing sparse spatiotemporal RWT data forced with both RCP 8.5 scenario dataset output downscaled from NEX-GDDP dataset projections, and nine hypothetical climate change scenarios.

5.2. Model Development

This chapter extended the kNN-LSTM model, which performed well to predict RWT in Chapter 4 by including streamflow time-series data in addition to AT. The study compared the extended hybrid kNN-LSTM monthly model with standalone LSTM, a modified nonlinear regression model proposed by van Vliet et al. (2011) (see Sect. 5.2.2 for further information about the nonlinear regression model), and 8-parameter version air2stream model discussed in section 4.2.4, Chapter 4 (Toffolon and Piccolroaz 2015; Piccolroaz et al. 2016). Detailed information on the k-NN algorithm, LSTM model, and air2stream model may be found in Chapter 4. Air temperature (AT[t]), streamflow (Q[t]), and time-lag effects of streamflow, air and water temperatures (Q[t-1], AT[t-1], RWT[t-1]) are used as input variables in the prediction of RWT for ML models. In the case of the modified nonlinear regression model, the 8-parameter air2stream model, air temperature (AT[t]), and streamflow (Q[t]) are used as input variables in the prediction of RWT. For future DO saturation levels projections with respect to RWT, this study used (i) RCP scenarios 8.5 downscaled projections of AT data were fed into the kNN-LSTM monthly prediction model which developed in Chapter 4 based on AT (Figure 5.1), and (ii) used the nine hypothetical scenarios of changes in AT and streamflow, which were fed into kNN-LSTM monthly prediction model (Figure 5.1).

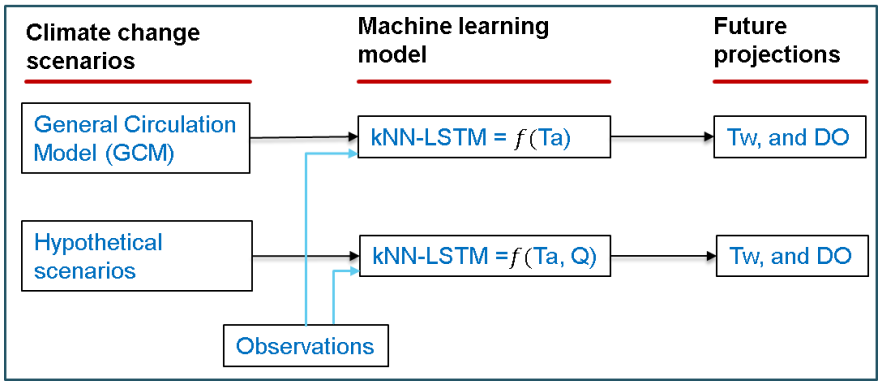


Figure 5.1. Schematic representation of the ML modeling framework with selected GCMs (i.e., NEXGDDP (RCP 8.5 scenarios)), and nine hypothetical climate change scenarios (Table 5.5) with observed dataset. The observed data was used to train the kNN-LSTM models. The climate change scenarios data was used to force the ML based modeling framework (kNN-LSTM), resulting in monthly simulations of water temperature (Tw) and dissolved oxygen (DO) saturation levels under future climate.

5.2.1. Oxygen Saturation

Waters with concentrations below saturation are called "deficit" whereas those with concentrations exceeding saturation are called "supersaturated". As a result, the oxygen saturation concentration serves as the baseline for any endeavor to measure oxygen-based water quality by determining the oxygen concentration of unpolluted water (Chapra et al. 2021). The saturated DO concentration depends on the temperature, salinity of water, and oxygen partial pressure. Saturated DO concentration is influenced by these elements, as indicated by (Rice et al., 2017)

$$o_s = \omega_k \,.\, \omega_s \,.\, e^{\ln o_{sf}(T)} \tag{5.1}$$

where o_s = saturated DO concentration (mgO_2/L), ω_k, ω_s = elevation above sea level (dimensionless), and salinity (dimensionless) respectively, and o_{sf} = the saturated DO concentration of sea-level freshwater (mgO_2/L). The following are the individual impacts of temperature, salinity, and elevation.

Temperature, *T* (◦C): The saturated oxygen of fresh water at sea level is estimated by evaluating the exponent of the exponential function of Equation (5.1) with (Rice et al., 2017)

$$\ln o_{sf}(T) = -139.34411 + \frac{1.575701 \times 10^5}{T_{abs}} - \frac{6.642308 \times 10^7}{T_{abs}^2} + \frac{1.243800 \times 10^{10}}{T_{abs}^3} - \frac{8.621949 \times 10^{11}}{T_{abs}^4} \quad (5.2)$$

where T_{abs} = absolute temperature in kelvin.

Salinity, S (ppt): The oxygen saturation of seawater is calculated by multiplying the sea-level freshwater saturation by (Rice et al., 2017)

$$\omega_s = e^{-S\left(1.7674\times10^{-2} + \frac{10.754}{T_{abs}} - \frac{2140.7}{T_{abs}^2}\right)} \quad (5.3)$$

Elevation, k (km): The influence of atmospheric pressure on gas saturation at elevation is based on the standard atmosphere as described by the cubic polynomial (Rice et al., 2017)

$$\omega_k = 1 - 0.11988\,k + 6.10834 \times 10^{-3}k^2 - 1.60747 \times 10^{-4}k^3 \quad (5.4)$$

Additional insight of DO can be obtained by computing the rate of change of saturation by differentiating Equation (5.1) with respect to temperature. Although functions like Equation (5.1) can sometimes be differentiated analytically, the results are cumbersome and typically provide no insight. Numerical differentiation provides an alternative means to obtain the same results with the centered divided difference (Chapra and Clough 2021)

$$h'(x) = \frac{h(x+\lambda) - h(x-\lambda)}{2\lambda} \quad (5.5)$$

where x = the value of the independent variable, $h'(x)$ = the function's first derivative with respect to x evaluated at x, and λ = a very small perturbation of x. For the present case, with $x = T$ and $h(x) = o_s(T)$, the result is $do_s(T)/dT$ with units of (mgO_2/L)/^oC.

5.2.2. Nonlinear Regression Model

To compare the ML model results, the modified regression model developed by van Vliet et al. (2011) is used in this study, which was based on the approach of Mohseni et al. (1998), who developed a nonlinear regression model representing the S-shaped function between AT and RWT to calculate stream temperature for monitoring stations in the United States. To generate RWT, the below equation is proposed by Mohseni et al. (1998):

$$T_w = \mu + \frac{\alpha - \mu}{1 + e^{\gamma(\beta - T_a)}} \tag{5.6}$$

$$\boldsymbol{\gamma} = \frac{\mathbf{4}\, tan\, \boldsymbol{\theta}}{\boldsymbol{\alpha - \mu}} \tag{5.7}$$

Where $\boldsymbol{\alpha}$ is the upper bound of RWT (^oC), μ is the lower bound of RWT (^oC), γ is the measure of the slope at the inflection point of the function ($^oC^{-1}$), β is the AT at the inflections point (^oC), $tan\theta$ is the slope at the inflection point (-). Figure 5.2 shows the meaning of parameters in Equation 5.6 (Mohseni et al. 1998).

Modifications to the above regression model have been made to include streamflow as a variable in addition to AT to include the effects of changes in river flow conditions on RWT and to apply the model on a monthly time step. Hence, the modified nonlinear regression model used in our study is:

$$T_w = \mu + \frac{\alpha - \mu}{1 + e^{\gamma(\beta - T_a)}} + \frac{\eta}{Q} + \varepsilon \tag{5.8}$$

Where η is the fitting parameter $^oC\ m^3\ s^{-1}$); Q is streamflow ($m^3\ s^{-1}$); ε is the error term (^oC).

5.3. Study Area and Data

For this study, seven majorly polluted catchments of India (CPCB 2015; National River Conservation Directorate (NRCD) 2018) were selected to analyze climate change impacts on DO with respect to RWT with various physiographic features as discussed in Section 4.3.1, chapter 4. In this study, Equation 5.1 is used to simulate the saturated DO concentration which depends on the water temperature, salinity of water, and elevation. Further, to examine the combined effects of streamflow and AT in the prediction of RWQ variables and subsequent future DO saturation levels, three polluted catchments of India (Tunga-Bhadra, Musi, and Ganga) where the continuous streamflow data is available were selected.

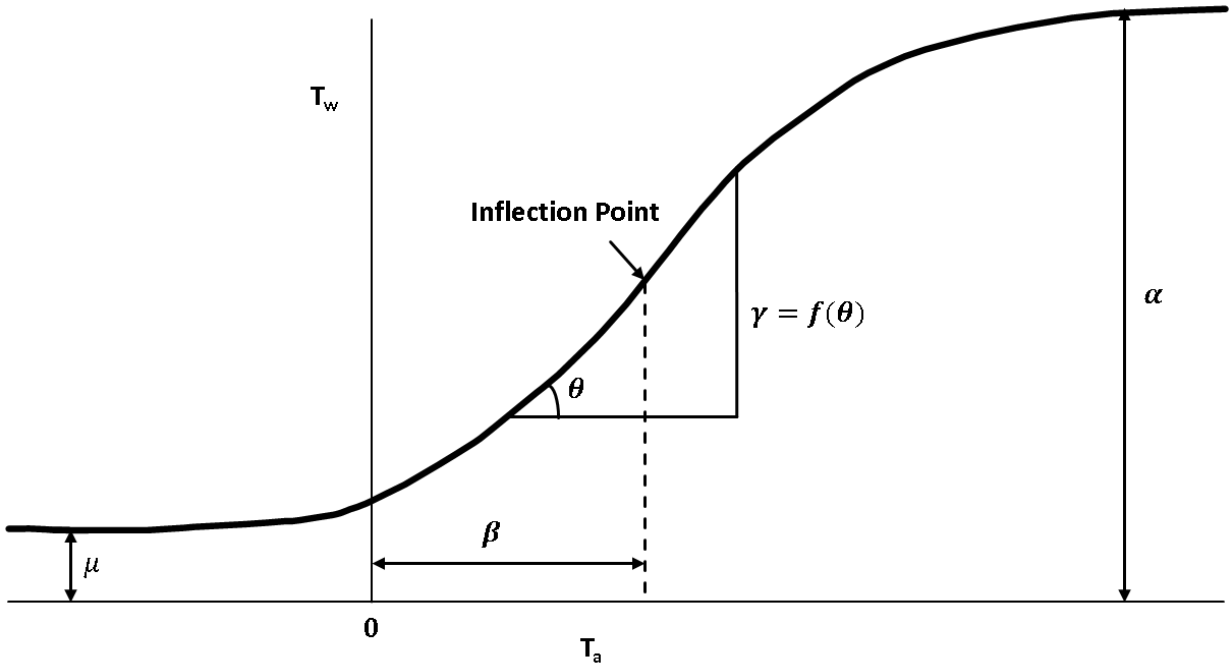


Figure 5.2. Schematic representation of the logistic function parameters (Equation 5.6). α is the upper bound of T_w (^oC), μ is the lower bound of T_w (^oC), γ is the measure of the slope at the inflection point of the function ($^oC^{-1}$), β is the T_a at the inflections point (^oC) (Mohseni et al. 1998).

5.4. Results

To examine the variability of annually averaged AT, RWT, and DO changes, the study calculated the linear trends using the observed data for seven catchments of India (Figure 5.3). The AT and RWT increased and observed DO has decreased during the studied period for all catchments except Cauvery, Godavari, and Ganga catchments (Figure 5.3). The RWT rising rates are lower than those of AT in general. Air temperature has shown a rising trend except for Cauvery (-0.01 ◦C/year) catchment, and the rising rates range from 0.002 to 0.380 ◦C/year. RWT shows a rising trend except for Cauvery (-0.06 ◦C/year), Godavari (-0.03 ◦C/year), and Ganga (-0.07 ◦C/year) catchments, and the rising rates vary between 0.01 and 0.17 ◦C/year.

DO shows a decreasing trend except for Cauvery (0.01 (mgO_2/L)/year), Godavari (0.004 (mgO_2/L)/year), and Ganga (0.01 (mgO_2/L)/year) catchments (where there is a significant decreasing trend of AT and RWT has been noted), and the decreasing rates vary

between -0.01 and -0.003 (mgO_2/L)/year. Such DO decrease patterns have been explored in several locations throughout the world. The DO, for instance, has been a seasonal DO variation, low ($DO < 10$ mgO_2/L) and high ($DO > 14$ mgO_2/L) over Clackamas River near Oregon City, OR, USA (Khani and Rajaee 2017), and rising RWTs in the Delaware River, the USA by 2 ^oC to peak summer levels of 30 ^oC, based on saturation, DO levels will decline by about 0.2 mgO_2/L (Kauffman 2018). Generally, RWT and AT are directly correlated, but RWT and DO are inversely correlated (Ficklin et al. 2013; Khani and Rajaee 2017). However, for the Godavari and Ganga catchment, the water temperature has shown decreasing trend (-0.03 ◦C/year and -0.07 ◦C/year respectively) with an increasing trend of AT (0.01 ◦C/year and 0.08 ◦C/year, respectively), which specifies that the temporal shifts of RWT may not be explained AT alone. RWT is directly influenced by multiple parameters, including streamflow (Sohrabi et al. 2017), river geometry, groundwater inputs, slope, water depth, etc. (Gu and Li 2002). Time series of monthly DO concentration (mgO_2/L) and river water temperature (^oC) for the period 2001-2015 at Ganga catchment is shown in Figure 5.4a, and monthly mean DO concentration (mgO_2/L) and river water temperature (^oC) based on 14 years average at Ganga catchment for the period 2001-2015 is shown in Figure 5.4b.

Model performances:

To study the role of streamflow and AT in ML models for prediction of RWT under sparse data scenarios, this chapter extended the kNN-LSTM model to predict RWT by integrating streamflow in addition to AT as model input for three majorly polluted river locations in India. The catchment monthly average streamflow for the Tunga-Bhadra, Musi, and Ganga catchments is 82.51 m^3/sec, 16.70 m^3/sec, and 949.68 m^3/sec, respectively. Spearman's correlation coefficients (SCC) for three catchments were estimated to examine the statistical dependency between AT, Q, and RWT variables. According to the metrics from Table 5.1, the RWT was positively correlated with AT and negatively correlated with Q for all three catchments on the annual scale. Overall, it was observed that RWT was weakly correlated with Q but strongly correlated with AT. To show the variability of AT, Q with RWT, the monthly data for three catchments has been compared, as shown in Figure 5.5.

The kNN-LSTM, LSTM, modified nonlinear regression model and 8-parameter version air2stream approaches were assessed using statistical measures (NSE, KGE, RSR,

RMSE, and MAE). Generated the kNN-LSTM, LSTM, modified nonlinear regression model, and air2stream model's performance results for monthly data for all three catchments using AT and Q variables as inputs, as provided in Table 5.2. For kNN-LSTM and LSTM models, included the time-lag effects of streamflow, air, and water temperatures (Q[t-1], AT[t-1], RWT[t-1]) as additional input variables. Results confirm that the developed models could predict RWT more accurately with Q as an input variable. Results also revealed that the kNN-LSTM model could predict RWT more accurately than the LSTM model by utilizing Q, AT, and lag variables as input. However, the air2stream model generated unsatisfactory results, and statistical measures can be found in Table 5.2.

The simulated samples from the k-NN bootstrap resampling algorithm (described in Chapter 4) are given as input to the kNN-LSTM hybrid model to predict the RWT for all three catchments of India. The relationship between monthly RWT, Q, and AT is relatively strongly correlated for the kNN-LSTM model (NSE values). The results were superior to those obtained from the LSTM, modified nonlinear regression model, including the air2stream model. The RMSE metrics for all the catchments vary from 1.00 to 1.74 for the kNN-LSTM model for monthly data (Table 5.2). The NSE value for the Ganga catchment is obtained as 0.94 for the kNN-LSTM model, which is reasonable compared with kNN-LSTM results generated in Chapter 4 (NSE: 0.446 – 0.920 °C). Based on RSR, KGE, and NSE performance values (Table 5.2), the LSTM and modified nonlinear regression model results are superior to those obtained from the air2stream model (Table 5.2). The performance of the LSTM model (NSE = 0.27 – 0.85, KGE = 0.38 – 0.81, RMSE=1.95 – 2.13, RSR = 0.39 – 0.78, and MAE = 1.35 – 1.70) was much superior to that of the modified nonlinear regression model (NSE = 0.25 – 0.60, KGE = 0.29 – 0.67, RMSE=1.57 – 3.08, RSR = 0.64 – 0.86, and MAE = 1.19 – 2.41) and the air2stream model (NSE = 0.04 – 0.70, KGE = 0.08 – 0.73, RMSE=2.12 – 3.03, RSR = 0.54 – 0.96, and MAE = 1.62 – 2.24). In general, the performance of the air2stream model on a monthly scale was not satisfactory. Overall, the extended kNN-LSTM model statistical metrics are reasonably within the range for all three catchment locations, and which agrees with the kNN-LSTM model results generated in Chapter 4 studies, providing confidence that the developed model performs effectively under data uncertainties scenarios. The following analyses concentrate on how RWT affects the DO saturation levels under both RCP experiments and hypothetical scenarios.

Oxygen Saturation and Oxygen Concentration under RCP experiments:

Figure 5.6 displays the box plots of RCP 8.5 experiments air temperature (°C) values; projected RWT (°C) and DO (mgO_2/L) values of historical, 2021-2050, and 2071–2100 for seven catchments of India. According to Figure 5.6, due to the increase of AT, the saturated DO concentrations are decreased mainly due to the increases of RWT for the periods 2021-2050, 2071-2100. Table 5.3 listed the rate of change of DO saturation levels under minimum, maximum, and average river water temperatures $do_s(T)/dT$ ((mgO_2/L)/°C) for observed and projected (2071-2100) for seven Indian catchments. Projected mean RWT changes for the periods 2071–2100 relative to mean observed values were calculated using RCP 8.5 output data and observed that results vary between the different catchments. The magnitude of DO decrease with respect to average RWT increase is higher for Narmada, Musi, and Ganga catchments, and variations in the rate of change of oxygen saturation for 2071–2100 relative to the historical values were noted as a drop of about 0.024, 0.018, and 0.025 (mgO_2/L)/°C, respectively (Table 5.3). Moderate DO decreases with respect to mean RWT for 2071–2100 are projected for catchments in the southern parts of India relative to the historical values, noted as a drop of about 0.005 (mgO_2/L)/°C for Cauvery and 0.009 (mgO_2/L)/°C for Godavari (Table 5.3). The magnitude of DO decrease with respect to minimum RWT increase is higher for Narmada, Sabarmati, Godavari, and Ganga catchments, and with respect to maximum RWT increase is higher for Narmada, Musi, and Ganga catchments (Table 5.3). Overall, results indicated that DO with respect to RWT over Indian catchments would likely drop by more than 0.02 (mgO_2/L)/°C for 2071-2100 (Table 5.3). Figure 5.7a shows the rate of change of DO saturation levels under mean river water temperature $do_s(T)/dT$ ((mgO_2/L)/°C) for observed and projected (2071-2100) data for seven Indian catchments. The vertical dotted lines indicate the mean of historical (Tw_{hist} °C) and projected (2071-2100) (Tw_{proj} °C) water temperatures. As depicted in Figure 5.7a, projected (2071-2100) (Tw_{proj} °C) water temperatures increase is higher for Narmada, Tunga-Bhadra, Musi, and Ganga catchments compared to historical (Tw_{hist} °C), which leads to a higher drop in the rate of change of oxygen saturation for these catchments. For Cauvery catchment, projected (2071-2100) (Tw_{proj} °C) water temperatures increase is low compared to historical (Tw_{hist} °C), which leads to a minimal drop of the rate of change of oxygen saturation.

The specification of a DO water-quality standard, o_{wq} (mgO_2/L), is used to evaluate oxygen assimilative capacity. Figure 5.7b shows the DO concentration (mgO_2/L) scale with respect to the observed (blue color) and projected (2071-2100) (maroon color) minimum, maximum and average water temperature (°C) levels of seven Indian catchments. From Figure 5.7b, observed that 10.3, 6.6, and 7.9 mgO_2/L, and 9.1, 6.3, and 7.3 mgO_2/L DO concentrations with respect to historical and projected (2071-2100) minimum, maximum and average water temperatures (°C) respectively of seven Indian catchments. DO concentration (mgO_2/L) scale scores dropped from 7.9 to 7.3 mgO_2/L respective to the observed and projected (2071-2100) mean RWT levels of seven catchments (Figure 5.7b). Table 5.4 listed the DO concentrations and DO decrease percentage with respect to monthly average summer and winter RWTs for historical and projected (2071-2100) with RCP 8.5 experiments for seven Indian catchments. The summer RWT increase for Tunga-Bhadra, Sabarmati, Musi, and Ganga basins are predicted as 3.1, 3.8, 5.8, 7.3 ^oC, respectively, with a more pronounced increase of 7.8 ^oC for the Narmada River for 2071-2100. The magnitude of DO concentrations decreases with respect to summer RWT increases is higher for Narmada, Musi, and Ganga catchment sites, and the percentage of DO decreases for 2071–2100 relative to the historical values noted 12.4, 9.3, and 11.9 %, respectively (Table 5.4). The low DO concentrations decrease observed for Cauvery and Godavari, and the percentage of DO decrease noted as 1.0 and 3.3 %, respectively (Table 5.4). Overall, the summer displayed larger percent decreases in DO compared to the winter season, and the largest DO decreases were found in the Narmada catchment (Table 5.4).

Oxygen Saturation and Oxygen Concentration under hypothetical climate change scenarios:

For investigating the potential changes in RWQ in the future, nine hypothetical climate change scenarios are considered. T2Q0, T2Q10, T2Q20, T3Q0, T3Q10, T3Q20, T4Q0, T4Q10, and T4Q20 are the climate change scenarios considered for this study (Table 5.5) (Santy et al. 2020), where the number followed by 'T' indicates the °C rise in AT and the number followed by 'Q' indicates the percentage reduction in the hydrological variable, streamflow. These scenarios are based on the projected rise of mean AT of 2.4-5.8 ^oC for the RCP 4.5 and 8.5 scenarios over seven Indian catchments (Table 4.10, Chapter 4). Such

projections are convincing with the global mean AT of 3.0–4.9 ^{o}C for the chosen GCM experiments done by van Vliet *et al.* (2013), the annual mean temperatures of 1.8 and 3.2 °C for the RCP 4.5 and 8.5 scenarios over Southeast Asia done by Raghavan *et al.* (2018), and AT of 0.0-2.0 ^{o}C, and streamflow of 0.0-20.0% reduction (i.e., T0FLOW10, T0FLOW20, T1FLOW0, T1FLOW10, T1FLOW20, T2FLOW0, T2FLOW10 and T2FLOW20 climate change scenarios) considered by Santy et al. (2020). In this study, based on the literature and evidence of changes, these nine hypothetical climate change scenarios are considered.

The chosen climate change scenarios offer a range of possibilities for future temperature increases and streamflow reductions, allowing us to assess the potential impacts on RWQ under various levels of severity. The scenarios range from T2 (°C), which represents a moderate increase, to T4 (°C), which signifies a more significant rise in AT. Recent climate change reports suggest that the global average AT has already risen by roughly 1°C compared to historical levels (IPCC 2023). The scenarios explore potential future increases beyond this current trend. All scenarios include a "Q" value, indicating a percentage reduction in streamflow. This reflects concerns about potential changes in precipitation patterns due to climate change. The scenarios range from Q0 (no reduction) to Q20 (20% reduction). Recent observations in some regions already show signs of altered precipitation patterns, with both increased flooding and drought events (e.g., Chennai floods). These scenarios explore potential future changes in streamflow severity.

Table 5.6 shows that the RWT (◦C) values increase, and the percentage of DO decreases with nine representative hypothetical climate change scenarios for three catchments with respect to historical values. The magnitude of RWT increases is higher for Musi and Ganga catchment sites and variations in RWT relative to the observed mean values noted 5.00, 4.30 ◦C, respectively (Table 5.6). Moderate variations in RWT are projected relative to the observed mean, noted as 3.00 ◦C for Tunga-Bhadra (Table 5.6).

Table 5.6 listed the DO decrease percentage with respect to monthly RWTs for historical and projected with nine hypothetical scenarios for three Indian catchments. According to Table 5.6, due to the increase of AT, and RWT, the DO percentages decreased for all the climate change scenarios. The magnitude of DO concentrations decreases with respect to RWT increases is higher for Musi and Ganga catchment sites, and the percentage of DO decreases for selected climate change scenarios relative to the historical values noted

9.07 and 13.22 %, respectively (Table 5.6). The low DO concentrations decrease observed for Tunga-Bhadra, and the percentage of DO decrease was noted as 4.64 % (Table 5.6). It is also found that RWT increase and DO decrease is not much influenced when the streamflow is reduced for all three catchments (Table 5.6). It may be noted that for scenarios of increased AT and reduced streamflow, DO is low for all three catchments. And DO is the lowest for scenario T4Q0. Therefore, scenario T4Q0 can be considered as the critical climate change scenario for all three catchments. Overall, the largest DO decreases were found in the Ganga catchment (Table 5.6).

Table 5.7 listed the rate of change of DO saturation levels under a average river water temperatures $do_s(T)/dT$ (mgO_2/L)/°C) for observed and future climate change scenarios (T4Q0, T4Q10, T4Q20) for three Indian catchments. The magnitude of DO decrease with respect to average RWT increase is higher for Musi, and Ganga catchments, and variations in the rate of change of oxygen saturation for T4Q0 scenario relative to the historical values were noted as a drop of about 0.013, 0.018, and 0.019 (mgO_2/L)/°C, respectively (Table 5.7). Overall, results indicated that DO with respect to RWT over three Indian catchments would likely drop by more than 0.016 (mgO_2/L)/°C for future climate change scenarios (T4Q0, T4Q10, T4Q20) (Table 5.7).

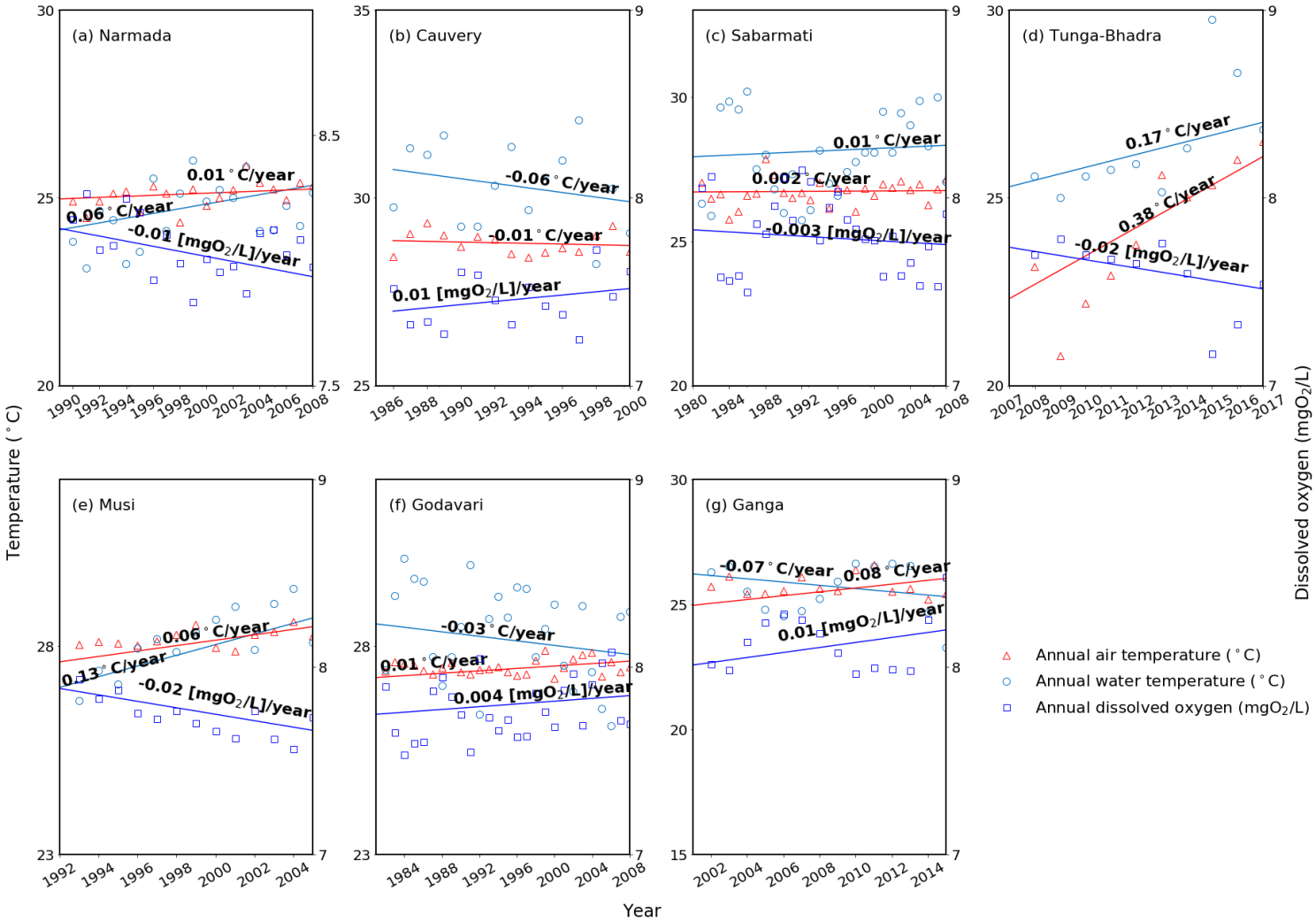


Figure 5.3. Seasonal, temporal variations of the mean annual air temperature (red), water temperature (light blue), and dissolved oxygen (blue) of the seven catchment stations (a) Narmada (b) Cauvery (c) Sabarmati (d) Tunga-Bhadra (e) Musi (f) Godavari (g) Ganga. Linear regressions of the time series are represented by trend lines, and the slope parameters are trend estimations.

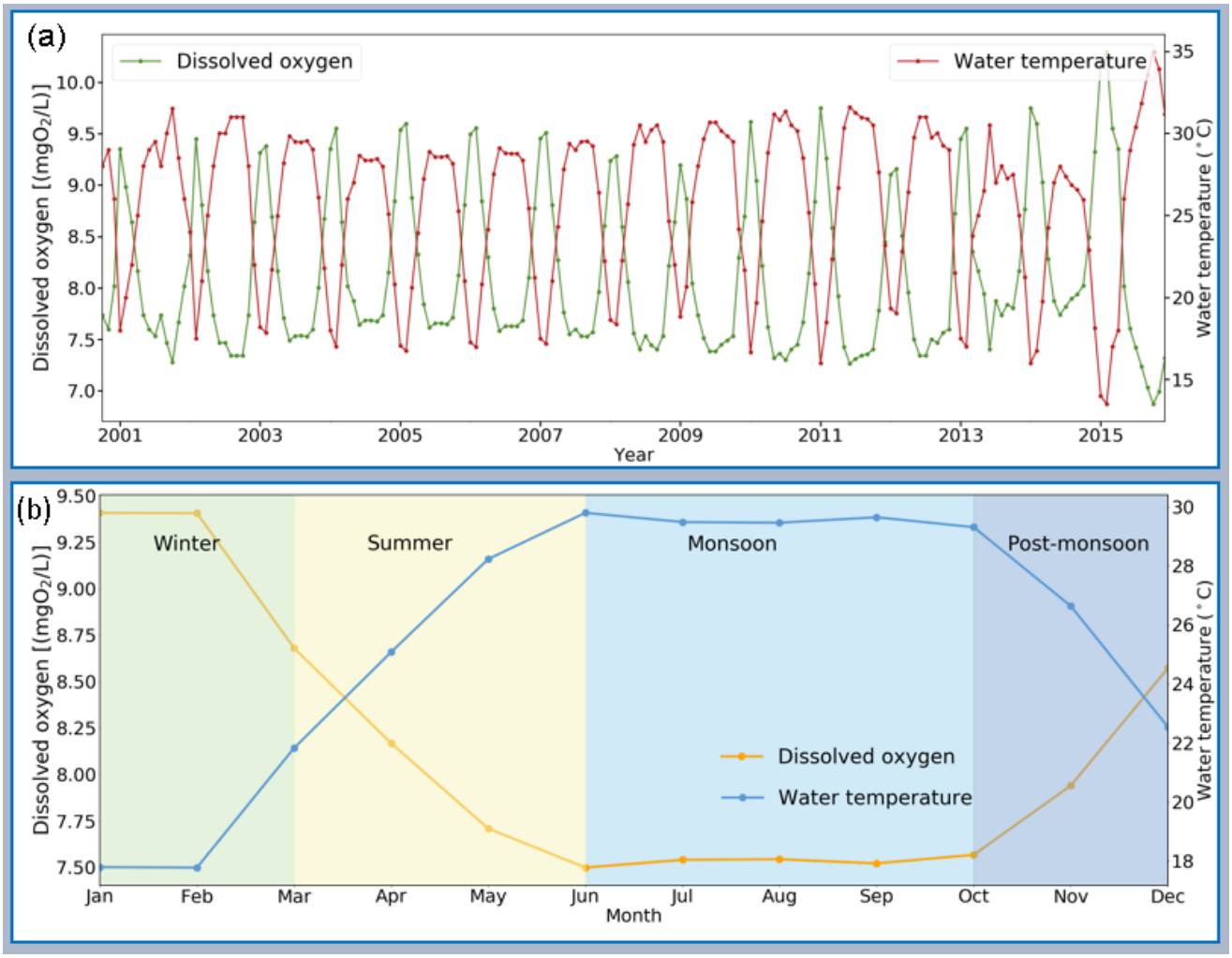


Figure 5.4. (a) time series of monthly dissolved oxygen concentration (mgO_2/L) and river water temperature (^oC) for the period 2001-2015 at Ganga catchment, and (b) monthly mean dissolved oxygen concentration (mgO_2/L) and river water temperature (^oC) based on 14 years average at Ganga catchment for the period 2001-2015.

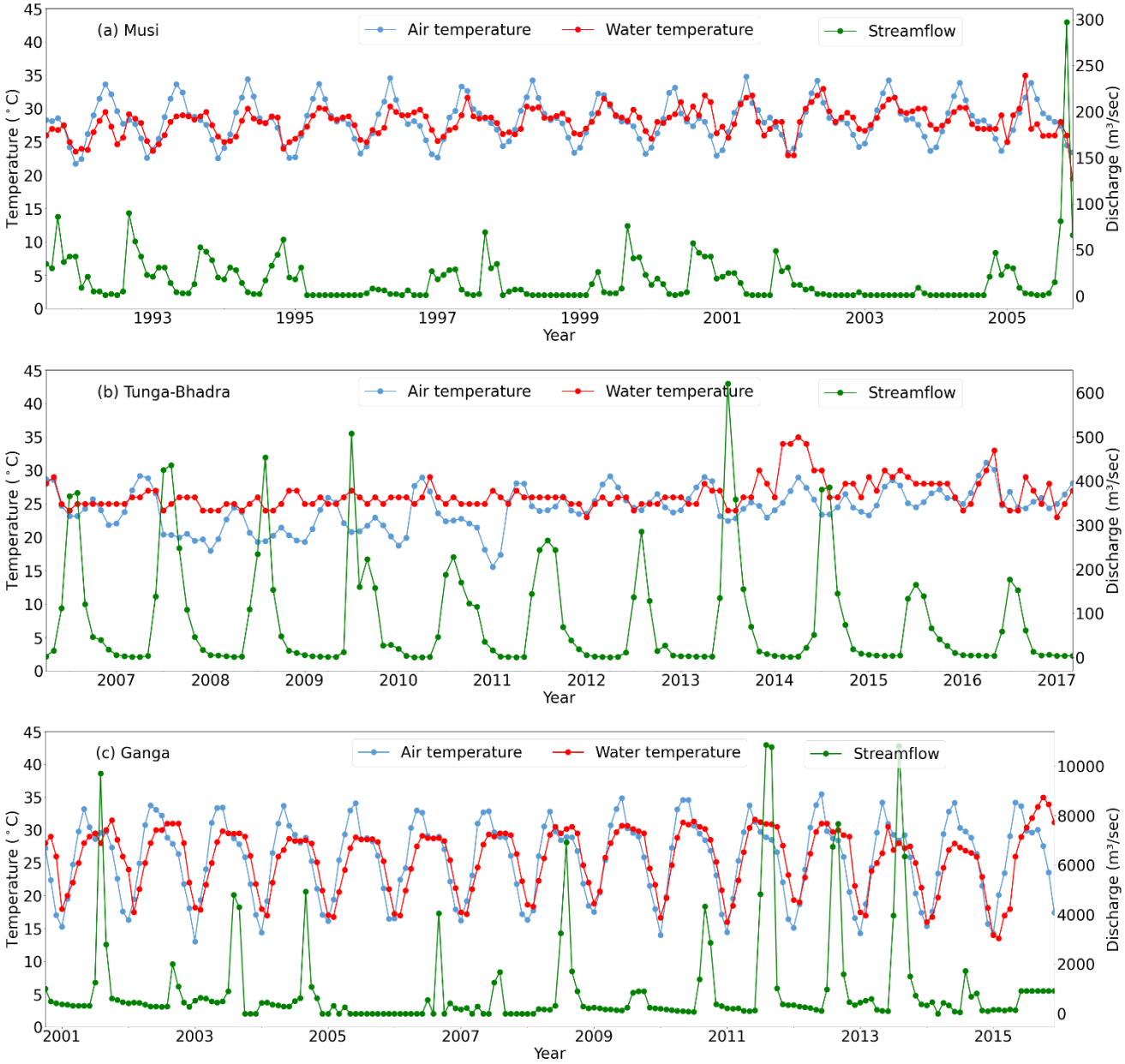


Figure 5.5. Time series plot of monthly air temperature **(ºC)** (blue), water temperature **(ºC)** (red), and streamflow (m^3/sec) (green) of the three catchment stations (a) Musi (b) Tunga-Bhadra (c) Ganga.

Table 5.1. Seasonal period Spearman's correlation coefficients between air and water temperature variables at different catchment areas. Values in parenthesis indicate correlation coefficients between streamflow and water temperature variables.

Catchment	Correlation Coefficient (RWT- AT) (RWT - Q)				
	Summer (March-May)	**Monsoon (June – September)**	**Post-monsoon (October – November)**	**Winter (December – February)**	**Annual**
Tunga-Bhadra	0.47(-0.58)	0.32(-0.30)	0.25(-0.21)	0.01(-0.07)	0.43(-0.15)
Musi	0.67(-0.44)	0.38(-0.08)	0.57(-0.34)	0.40(-0.38)	0.63(-0.28)
Ganga	0.87(-0.12)	0.19(-0.04)	0.66(-0.41)	0.06(-0.09)	0.66(-0.33)

Table 5.2. Overview of kNN-LSTM, LSTM, modified nonlinear regression model (van Vliet et al. 2011), and air2stream model performances based on streamflow (Q[t]), air temperature (AT[t]) as input variables for Tunga-Bhadra, Musi, and Ganga catchments. The values displayed all referred to the testing period. For LSTM and kNN-LSTM model, included the time-lag effects of streamflow, air, and water temperatures (Q[t-1], AT[t-1], RWT[t-1]) as additional input variables.

Catchment	Model	NSE	KGE	RMSE	RSR	MAE
Tunga-Bhadra	kNN_LSTM	0.46	0.53	1.74	0.74	1.05
	LSTM	0.27	0.38	1.98	0.85	1.59
	van Vliet	0.25	0.29	1.89	0.86	1.34
	air2stream	0.04	0.08	2.12	0.96	1.89
Musi	kNN_LSTM	0.75	0.82	1.00	0.51	0.71
	LSTM	0.38	0.45	1.95	0.78	1.35
	van Vliet	0.45	0.53	1.57	0.74	1.19
	air2stream	0.21	0.32	2.19	0.88	1.62
Ganga	kNN_LSTM	0.94	0.96	1.15	0.25	0.79
	LSTM	0.85	0.81	2.13	0.39	1.70
	van Vliet	0.60	0.67	3.08	0.64	2.41
	air2stream	0.70	0.73	3.03	0.54	2.24

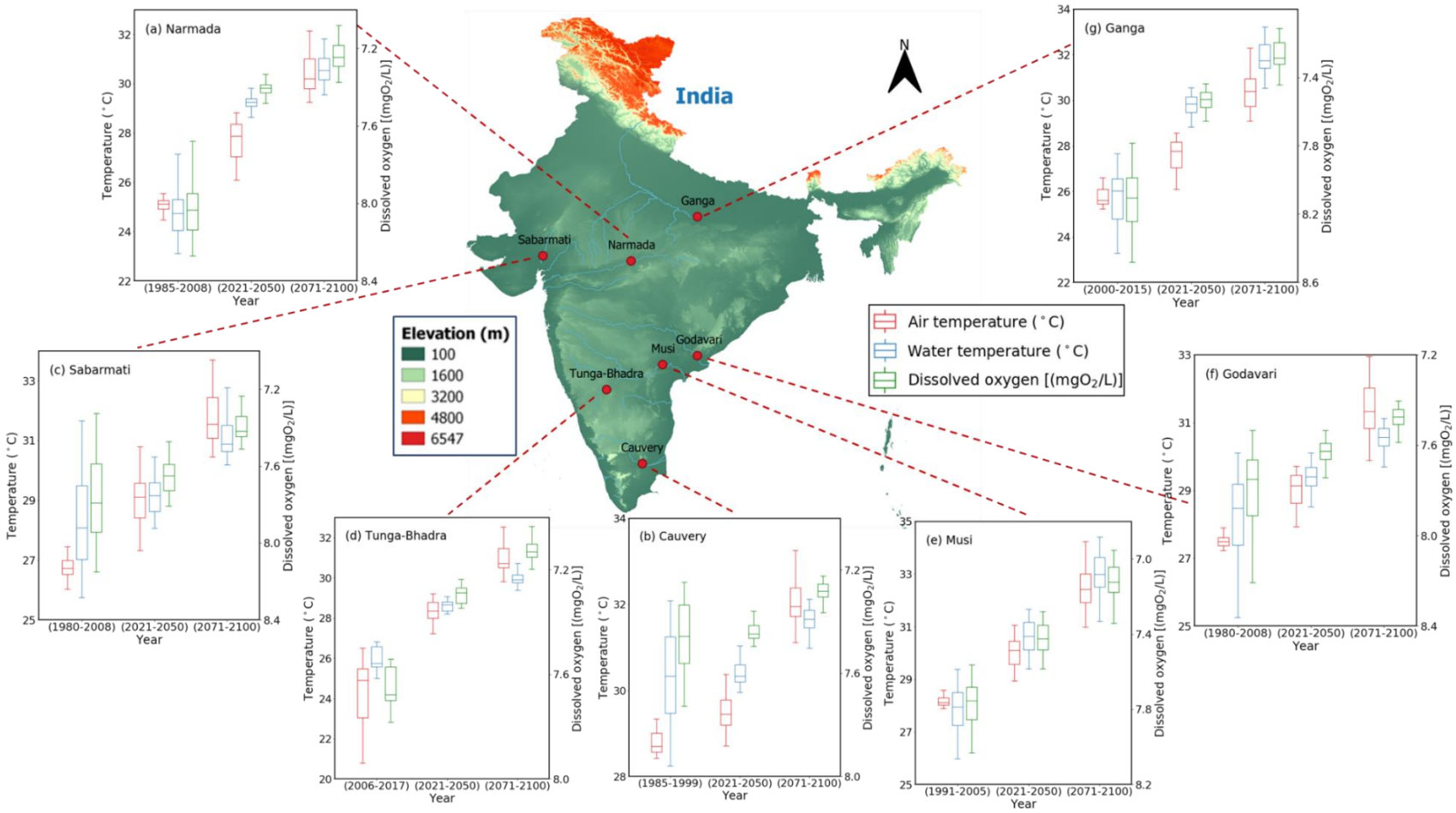


Figure 5.6. Boxplots represent the Representative Concentration Pathway (RCP) 8.5 experiments air temperature (°C) values; projected water temperature (°C) and dissolved oxygen (mgO_2/L) values of historical, 2021-2050, and 2071–2100 for seven catchments.

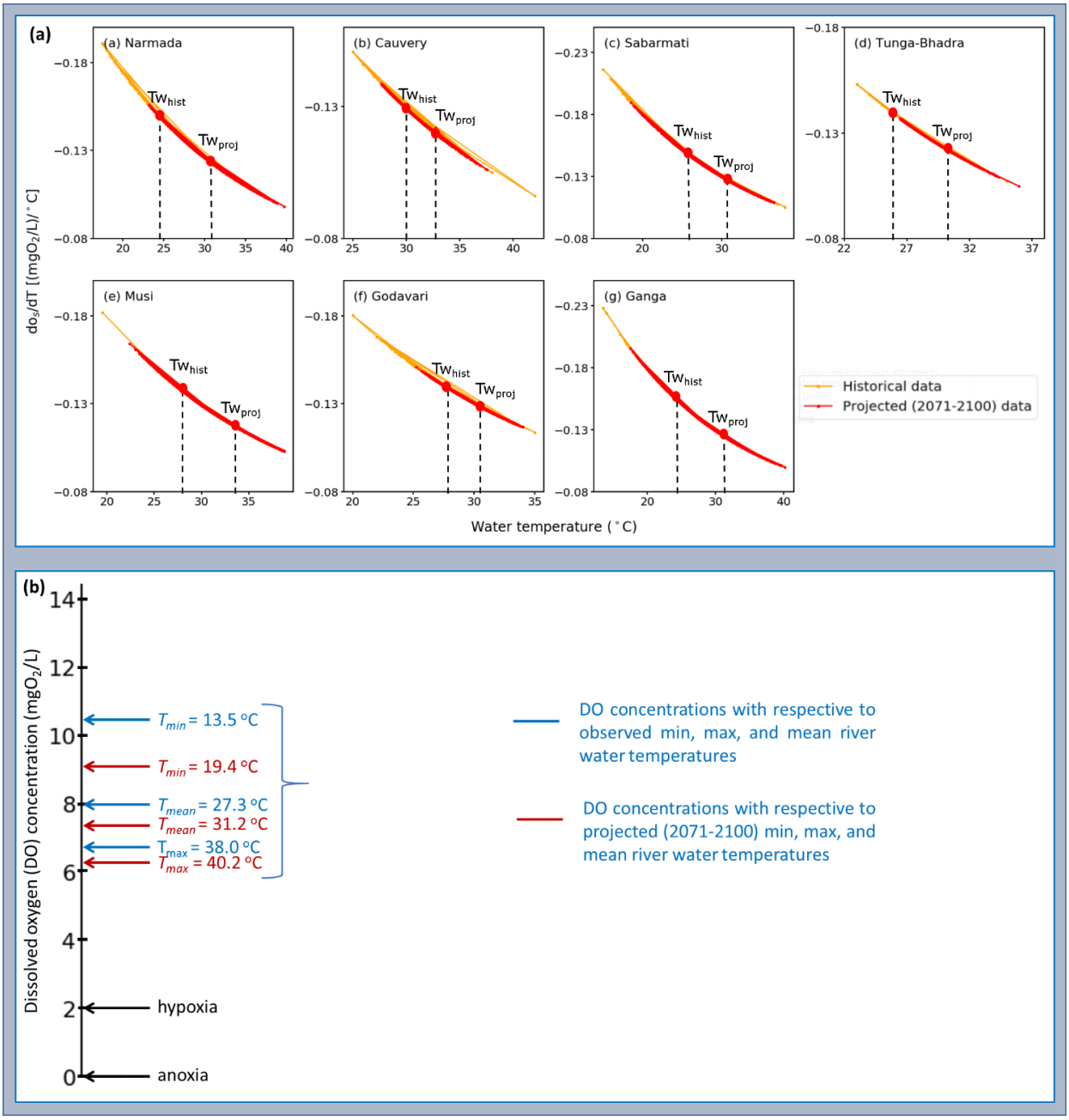


Figure 5.7. (a) The rate of change of oxygen saturation under mean river water temperature $do_s(T)/dT$ ((mgO_2/L)/°C) for historical and projected (2071-2100) data for seven Indian catchments. The vertical dotted lines indicate the mean of historical (Tw_{hist} °C) and projected (2071-2100) (Tw_{proj} °C) water temperatures, and (b) **t**he DO concentration (mgO_2/L) scale with respect to the observed (blue color) and projected (2071-2100) (maroon color) minimum, maximum and mean water temperature (°C) levels of seven Indian catchments.

Table 5.3. The rate of change of oxygen saturation levels under a minimum, maximum, and average river water temperatures (in parentheses) ($do_s(T)/dT$ ((mgO_2/L)/°C)) for historical and projected (2071-2100) at respective elevations for seven Indian catchments. Set the Salinity (S) value for seven river catchments to zero.

Catchment	**Elevation (k) in km**	**Historical data**			**Projected data (2071-2100)**			**$do_s(T)/dT$ variation**		
		$do_s(T)/dT$ (Tw_{min} (°C)) (1)	$do_s(T)/dT$ (Tw_{max} (°C)) (2)	$do_s(T)/dT$ (Tw_{mean} (°C)) (3)	$do_s(T)/dT$ ((Tw_{min} °C)) (4)	$do_s(T)/dT$ (Tw_{max} (°C)) (5)	$do_s(T)/dT$ (Tw_{mean} (°C)) (6)	(1) – (4)	(2) – (5)	(3) – (6)
Narmada	0.30	-0.191 (17.5)	-0.110 (35.0)	-0.148 (24.7)	-0.156 (23.2)	-0.098 (39.8)	-0.124 (30.6)	-0.035	-0.012	-0.024
Cauvery	0.08	-0.151 (25.0)	-0.103 (38.0)	-0.128 (30.4)	-0.139 (27.7)	-0.106 (37.5)	-0.123 (31.7)	-0.012	-0.003	-0.005
Sabarmati	0.05	-0.216 (15.0)	-0.105 (38.0)	-0.137 (28.1)	-0.160 (23.4)	-0.109 (36.6)	-0.126 (31.1)	-0.056	0.004	-0.011
Tunga-Bhadra	0.50	-0.153 (23.0)	-0.107 (35.0)	-0.137 (26.4)	-0.137 (26.5)	-0.105 (35.9)	-0.123 (30.0)	-0.016	-0.002	-0.014
Musi	0.09	-0.182 (19.5)	-0.113 (35.0)	-0.137 (27.9)	-0.164 (22.4)	-0.103 (38.6)	-0.119 (33.0)	-0.018	-0.010	-0.018
Godavari	0.02	-0.180 (20.0)	-0.114 (35.0)	-0.137 (28.2)	-0.148 (25.8)	-0.117 (34.0)	-0.128 (30.5)	-0.032	0.003	-0.009
Ganga	0.10	-0.228 (13.5)	-0.113 (34.9)	-0.148 (25.6)	-0.182 (19.4)	-0.100 (40.2)	-0.123 (31.8)	-0.046	-0.013	-0.025

Table 5.4. The DO concentrations and percentage of DO decrease with respect to monthly average summer and winter (in parentheses) water temperatures for historical and projected (2071-2100) with Representative Concentration Pathway (RCP) 8.5 experiments for seven Indian catchments.

Catchment	**Historical data**		**Projected data (2071-2100)**		**RWT (°C increase)**	**DO (%decrease)**
	Tw_{mean} (°C)	**DO (mgO_2/L)**	**Tw_{mean} (°C)**	**DO (mgO_2/L)**		
Narmada	26.14 (22.42)	7.80 (8.37)	33.90 (27.08)	6.83 (7.68)	7.76 (4.66)	12.44 (8.24)
Cauvery	32.19 (29.78)	7.22 (7.52)	32.68 (30.43)	7.15 (7.43)	0.49 (0.65)	0.97 (1.20)
Sabarmati	28.60 (24.85)	7.70 (8.24)	32.39 (27.13)	7.21 (7.90)	3.79 (2.28)	6.36 (4.13)
Tunga-Bhadra	27.60 (25.78)	7.43 (7.67)	30.66 (28.62)	7.04 (7.29)	3.06 (2.84)	5.25 (4.95)
Musi	29.03 (25.82)	7.60 (8.05)	34.80 (29.37)	6.89 (7.56)	5.77 (3.55)	9.34 (6.09)
Godavari	29.12 (26.59)	7.66 (8.01)	30.94 (27.88)	7.41 (7.82)	1.82 (1.29)	3.26 (2.37)
Ganga	25.04 (19.44)	8.15 (9.09)	32.34 (26.40)	7.18 (7.96)	7.30 (6.96)	11.90 (12.43)

Table 5.5. Climate change scenarios considered.

Scenario No.	Name	Description
1	T2Q0	Air temperature increase by 2°C with no change in streamflow
2	T2Q10	Air temperature increase by 2°C & streamflow reduce by 10%
3	T2Q20	Air temperature increase by 2°C & streamflow reduce by 20%
4	T3Q0	Air temperature increase by 3°C with no change in streamflow
5	T3Q10	Air temperature increase by 3°C & streamflow reduce by 10%
6	T3Q20	Air temperature increase by 3°C & streamflow reduce by 20%
7	T4Q0	Air temperature increase by 4°C with no change in streamflow
8	T4Q10	Air temperature increase by 4°C & streamflow reduce by 10%
9	T4Q20	Air temperature increase by 4°C & streamflow reduce by 20%

Table 5.6. River water temperatures (T_w) increase and the percentage of DO decreases with nine representative hypothetical climate change scenarios for three catchments. Values in parenthesis indicate the percentage of DO decrease.

Catchment	**Observed catchment mean**				**Projected T_w (°C) increase, and respective percentage DO saturation levels decrease for below Scenarios**								
	Ta (°C)	**Tw (°C)**	**Q (m^3/sec)**	**DO (mgO_2/L)**	T2Q0 (1)	T2Q10 (2)	T2Q20 (3)	T3Q0 (4)	T3Q10 (5)	T3Q20 (6)	T4Q0 (7)	T4Q10 (8)	T4Q20 (9)
Tunga-Bhadra	24.20	26.30	82.50	7.55	1.90 (2.78)	1.87 (2.78)	1.82 (2.65)	2.40 (3.58)	2.40 (3.58)	2.40 (3.58)	**3.00** (**4.64**)	2.97 (4.64)	2.93 (4.50)
Musi	28.10	28.00	16.70	7.82	2.41 (5.12)	2.41 (5.12)	2.41 (5.12)	3.73 (7.28)	3.73 (7.28)	3.73 (7.28)	**5.00** (**9.21**)	4.92 (9.07)	4.92 (9.07)
Ganga	25.70	25.60	949.68	8.62	2.74 (10.79)	2.72 (10.79)	2.69 (10.79)	3.54 (12.06)	3.51 (12.06)	3.48 (11.95)	**4.30** (**13.22**)	4.28 (13.22)	4.22 (13.11)

Table 5.7. The rate of change of oxygen saturation levels under an average river water temperatures (in parentheses) ($do_s(T)/dT$ ((mgO_2/L)/°C)) for historical and future climate change scenarios (T4Q0, T4Q10, T4Q20) at respective elevations for three Indian catchments. Set the Salinity (S) value for three river catchments to zero.

Catchment	**Elevation (k) in km**	**Historical data**	**Projected hypothetical scenario**			**$do_s(T)/dT$ variation**		
		$do_s(T)/dT$ (Tw_{mean} (°C)) (1)	$do_s(T)/dT$ (T4Q0) (2)	$do_s(T)/dT$ (T4Q10) (3)	$do_s(T)/dT$ (T4Q20) (4)	(1) – (2)	(1) – (3)	(1) – (4)
Tunga-Bhadra	0.50	-0.138 (26.3)	-0.125 (29.3)	-0.126 (29.27)	-0.126 (29.23)	**-0.013**	-0.012	-0.012
Musi	0.09	-0.137 (28.0)	-0.119 (33.0)	-0.119 (32.92)	-0.119 (32.92)	**-0.018**	-0.018	-0.018
Ganga	0.10	-0.148 (25.6)	-0.129 (29.9)	-0.129 (29.88)	-0.130 (29.82)	**-0.019**	-0.019	-0.018

5.5. Discussion

This chapter presents new intuitions on the assessment of climate change impacts on saturated DO concentrations with respect to RWT for seven different catchment sites across India in different physiographic settings. For this, using the monthly kNN-LSTM prediction model, which was developed in Chapter 4, based on AT, including time-lag effects, forced with GCM based RCP scenarios (Figure 5.1), demonstrates rising RWTs will reduce a river's assimilative capacity by affecting its oxygen metabolism, in addition to lowering saturation. This chapter also presents the combined effects of streamflow and AT in prediction of RWT using the kNN-LSTM model, LSTM model, a modified nonlinear regression model, and an 8-parameter version air2stream when applied to three major river systems of India. Further quantified the climate change impact DO saturation levels using the kNN-LSTM model forced with nine hypothetical climate change scenarios (Figure 5.1).

The monthly NSE scores ranged between 0.04–0.94, KGE scores ranged between 0.08–0.96, RSR scores ranged between 0.25–0.96, and RMSE scores were ≤ 3 °C during the testing periods, revealing high model reliability. All the developed model statistical metrics covered the range of model reliability described in the literature. The RMSE scores for all three catchments ranged between 1.00 - 2.13 °C pertaining to all DL models (Table 5.2) for monthly data, which are reasonable in comparison to DL models developed in Chapter 4 (1.199 - 3.294 °C). The NSE values for Musi (0.75), and Ganga (0.94) catchments pertaining to all DL models for monthly data (Table 5.2), which are sensible in comparison with models developed in Chapter 4 (Musi (0.735), and Ganga (0.920)) and the earlier model of LSTM by Qiu et al. (2021) (0.74 - 0.99). Overall, in this study, the kNN-LSTM model's performance was better when modeling monthly data by including streamflow as an additional input variable, demonstrating that coupling with the k-NN bootstrap resampling algorithm would yield more precise results, which agrees with the Chapter 4 studies based on AT and guaranteed results to outperform standalone DL models, nonlinear regression models and air2stream models (Table 5.2).

Assessment of Projected Changes in RWT and DO Saturation Levels under RCP scenarios:

The RWT increases up to 7 ^{o}C for summer, reaching close to 35 ^{o}C, decreases DO by 2%–12%, thus decreasing the saturation capacity for DO for 2071–2100. DO concentration (mgO_2/L) scale scores dropped from 7.9 to 7.3 mgO_2/L respective to the observed and projected (2071-2100) mean RWT levels of seven catchments. These scores reveal that DO concentration (mgO_2/L) values are dropping for projected years as RWTs rise. The RWT increases of up to7 ^{o}C for summer, demonstrated in this work, were found to agree with Chapra et al. (2021) (5 ^{o}C increments in summer RWTs in most of the world's rivers over the next 50 years). The DO concentration (mgO_2/L) scale scores, as demonstrated in this work, were found to agree with Du et al. (2019) (DO concentrations on the basin average scale will decrease by 0.72 mgO_2/L under RCP 8.5 scenario for 2061-2100) and Chapra et al. (2021) (DO oxygen concentrations are 9.0 and 6.8 mgO_2/L for freshwater temperatures 20 and 35 °C, respectively).

The percentage of DO decrease with respect to summer RWTs is higher for Narmada, Musi, and Ganga catchment sites for 2071–2100 relative to the historical values noted as 12.4, 9.3, and 11.9 %, respectively, probably because of the influence of disposal of untreated sewage and industrial wastewater along with due to increased reaction kinetics at a higher temperature under climate change scenarios (Table 5.4). In this study, overall, for all seven catchments, the decrease in DO is 8% for the plausible future (2071–2100) (Figure 5.7b and Table 5.4). These projected change patterns are most consistent with earlier hydrological model studies by Ficklin et al. (2013) (10% decreases in DO by 2100 at Sierra Nevada in California, USA) and by Du et al. (2019) (DO decrease on the basin average scale by 0.72 mgO_2/L under RCP 8.5 scenario for 2061-2100 in the Athabasca River Basin, Canada).

Assessment of Projected Changes in RWT and DO Saturation Levels under Hypothetical Scenarios:

This section presents the discussion about the assessment of climate change impacts on saturated DO concentrations with respect to RWT for three different catchment sites across India. For this, forcing the monthly kNN-LSTM prediction model with hypothetical climate

change scenarios demonstrates rising RWTs will reduce a river's assimilative capacity by affecting its oxygen metabolism, in addition to lowering saturation. The RWT increases up to 5 ^{o}C, and decreases DO by 4.64%–13.22%, thus decreasing the saturation capacity for DO for the selected climate change scenarios. The percentage of DO decrease with respect to RWTs is higher for Musi and Ganga catchment sites for the T4Q0 climate change scenario relative to the historical values noted as 9.21, and 13.22 %, respectively (Table 5.6). Based on the results from the nine hypothetical scenarios, it was observed that streamflow played a minor role in explaining the RWT, which agrees with the earlier studies (Zhu et al. 2019f; Drainas et al. 2023). In this study, overall, for all three catchments, the decrease in DO is 9% for the plausible climate change scenarios (T4Q0) (Table 5.6). These projected change patterns are most consistent with RCP results discussed in Section 5.4 (8% decrease in DO by 2100 for Indian catchments (Table 5.4)).

Overall, this study demonstrated how river oxygen levels would be influenced by rising RWT due to climate change using the kNN-LSTM model for the Indian riverine system under sparse data. The rising RWTs will reduce river assimilative capacity by affecting its oxygen metabolism, in addition to lowering saturation, and necessitates redefining/alterations of the river water quality standards under climate change.

Furthermore, the DO simulated by Equation (5.1) is the saturated oxygen concentration, which is the total amount of DO that can be dissolved within the streamflow volume, and thus, it can be expected that the DO concentrations presented in this study represent the ceiling of potential DO levels. Though the performant hybrid kNN-LSTM model is used in this chapter, further research is needed for better RWT model to estimate accurate DO saturation levels. This study uses highly idealized hypothetical scenarios for inferring the impacts of streamflow, AT effect on the water quality of the catchments considered. Projections of future climate change from GCMs will provide more realistic insight into the problem. This study set the Salinity (S) value for seven river catchments to zero because most rivers and streams had minimal salinity (Chapra et al. 2021). Overall, this research offers vital intuitions about the historical and projected RWT and DO states of major Indian river catchment locations, which may be beneficial in creating future water management plans that may impact aquatic resources.

5.6. Chapter Summary

This chapter demonstrates the climate change impacts on saturated DO concentrations with respect to RWT for the seven major polluted Indian catchments at a monthly timescale. First, the hybrid kNN-LSTM model was used, which was implemented in Chapter 4 to predict the RWT addressing sparse spatiotemporal RWT data. Further assessed the climate change impacts on DO concentrations with respect to RWT using a forced by an ensemble of RCP 8.5 scenario downscaled projections of AT data from the NEX-GDDP dataset. Further, demonstrates the climate change impacts on water quality indicators such as RWT, and saturated DO concentrations under nine highly hypothetical climate change scenarios of AT and streamflow for the three major polluted Indian catchments at a monthly timescale. For this, the hybrid kNN-LSTM model (Chapter 4) is extended by including streamflow as an additional input variable to predict the RWT. The results lead to the following conclusions:

1. The hybrid kNN-LSTM model outperforms the standalone LSTM model, nonlinear regression model, and air2stream model under limited data scenarios for the prediction of RWT when including discharge as feature variable in addition to AT, which agrees with the Chapter 4 studies.
2. An increase in AT will have an effect on RWTs, and saturated DO concentrations. The latter will trigger higher RWT and lower DO concentration. These changes appear especially significant for the summer seasons and include RWT increases of up to7 ^{o}C for summer, reaching close to 35 ^{o}C, decreases of DO by 2%–12%, thus decreasing the saturation capacity for DO.
3. The percentage of decrease of DO saturation levels with respect to summer RWTs is higher for Narmada, Musi, and Ganga catchment sites for 2071–2100 relative to the historical values noted as 12.4, 9.3, and 11.9 %, respectively.
4. DO concentration (mgO_2/L) scale scores dropped from 7.9 to 7.3 mgO_2/L respective to the observed and projected (2071-2100) mean RWT levels of seven catchments.
5. In hypothetical scenarios, RWT increases by 5 ^{o}C, and decreases DO by 4.8%–13.2%, thus decreasing the saturation capacity for DO. The percentage of decrease of DO saturation levels with respect to increase of RWTs is higher for Musi and Ganga catchment sites for the selected climate change scenarios relative to the historical values noted as 9.21, and 13.22 %, respectively.

6. Additionally, it was found that streamflow played a minimal role in ML models for RWT predictions for the selected catchments in India.
7. Overall, saturated DO concentration (mgO_2/L) levels are dropping by 8% under the rise of summer RWT by more than 4.3 °C for 2071-2100. That is, for every 1 ^oC RWT increase, there will be about a 2.3 % decrease in DO saturation level concentrations over Indian catchments under climate signals.
8. The study provides an assessment of the individual contribution of RWT rise on depletion of saturated DO levels, which is helpful for the policymakers and pollution control authorities for sustainable river water quality management in future climate change scenarios.

Overall, the present chapter demonstrated the climate change impacts on DO saturation levels with respective RWT under data uncertainties.

Chapter 6
CONCLUSIONS AND FUTURE DIRECTIONS

6.1. Conclusions

The research reported in this thesis contributes to developing methodologies for RWQ modeling. The data sparseness is the problem of sampling which leads to a lack of complete data. Data sparseness comes from the frequency of measurements, if the frequency is high no data sparseness. In this thesis na.interp() method in R's forecast library was utilized to interpolate the missing observations using the STL (Seasonal and Trend decomposition using Loess) decomposition (Hyndman et al. 2018). In the introduction to this thesis, several questions were posed regarding the ability of ML approaches to predict RWQ variables in scenarios with minimum input variables, a lack of long-time series data, non-stationary data, and moreover how climate change impacts RWQ variables. In order to tackle these questions, new methods were proposed and evaluated them for Indian river systems by considering the two most important water quality variables, i.e., RWT and DO saturation concentration levels. The summary of the research findings from this study is presented here.

Q1: *"How can sensitivity analysis reveal a deeper understanding of the underlying processes governing water quality in the river systems? How a sensitivity analysis can be coupled with ML approaches to select the best suitable and effective variables in predicting river water quality variables? How to assimilate theory-driven, understanding rich processes with data-driven approaches to improve the predicted values based on the measurement data?"*

The present study made efforts to identify the most sensitive AT variable (average, maximum and minimum) using Sobol' sensitivity analysis method described in Section 3.2.2. Further The present study demonstrates how new ML approaches, such as Ridge regression (RR), K-nearest neighbors (KNN), Random Forest (RF), and Support Vector Regression (SVR), can be coupled with Sobol' global sensitivity analysis (GSA) to predict accurate RWQ variables estimates with minimum data inputs. The results lead to the following conclusions:

1. The results indicated that the maximum AT was the most sensitive variable in the prediction of RWT among average, minimum, and maximum AT

for Tunga Bhadra River system.

2. The SVR has been noted as the most robust ML model to predict RWT at a monthly time scale compared to daily and seasonal.
3. The study revealed that hybrid ML models, i.e., EnKF data assimilation algorithm with ML approaches improves the predicted values based on the measurement data in RWQ modeling.

Overall, the study demonstrates how ML methods can be coupled with sensitivity analysis and DA techniques to generate accurate RWQ variables prediction under minimum data inputs (such as AT).

Q2: *"How to infer the relationships between river water quality indicators and hydroclimatic variables (e.g., Air Temperature (AT), streamflow)? How do different potentially influence lagged variables as additional predictive power features in river water quality modeling to improve the model performance under sparse, non-stationary data, and seasonality scenarios?"*

In water quality modeling, long time series data is required to extract time characteristic information. To address these issues, the k-NN bootstrap resampling algorithm was used for generating simulated time-series data from historical data. To overcome both processing of limited data, and non-stationary in river water quality data, in this thesis the kNN-LSTM, and WT-LSTM hybrid ML models were developed to predict RWT under data uncertainties (Section 4.2). The results lead to the following conclusions:

1. When WT and k-NN bootstrap resampling algorithms were included, LSTM outperformed the conventional models.
2. The hybrid kNN-LSTM model yielded better performance results for five catchment sites (i.e., Narmada, Cauvery, Musi, Godavari, and Ganga) out of seven catchment sites than LSTM, WT-LSTM, and air2stream prediction models at a monthly scale.
3. The widely used process-based model air2stream is used to validate the proposed ML models and to make the results comparable to previous studies. It was found that the 3-parameter version of air2stream mainly delivered the

lowest performance compared to LSTM, WT-LSTM, and kNN-LSTM models for almost every basin.

4. Performed the continuous wavelet analysis to examine the variability of water quality variables and concluded that the water quality variables showed stable 1-year periodicities for the whole duration of the time series and displayed a recurrent annual pattern.
5. In this study, the kNN-LSTM model was tested using various monthly data points to see how the model performed for Musi and Ganga catchment stations with data limitations. It was observed that the kNN-LSTM was still producing good results with fewer monthly data time series values.

Q3: *"How do the climate change variables (e.g., temperature, precipitation) impact key physical processes within a river system (e.g., biological activities, dilution), ultimately influencing river water quality variables?"*

In this thesis, quantified the climate change impacts on the thermal regimes of rivers in India and possible variability in RWQ variables, i.e., RWT and DO saturation levels by using the best performance hybrid kNN-LSTM model forced with an ensemble of RCP 4.5, and 8.5 scenarios, and nine hypothetical climate change scenarios. The results lead to the following conclusions:

1. The RWT increase for Tunga-Bhadra, Musi, Ganga, and Narmada basins are predicted as 3.0, 4.0, 4.6, and 4.7 ^{o}C, respectively for 2071-2100.
2. Overall, RWT over Indian catchments is likely to rise by more than 3.0 °C for 2071-2100.
3. In summer seasons, the RWT reaching close to 35 ^{o}C, decreases DO by 2%–12%, thus decreasing the saturation capacity for DO for 2071-2100.
4. The percentage of decrease of DO saturation levels with respect to summer RWTs is higher for Narmada, Musi, and Ganga catchment sites for 2071–2100 relative to the historical values noted as 12.4, 9.3, and 11.9 %, respectively.
5. DO concentration (mgO_2/L) scale scores dropped from 7.9 to 7.3 mgO_2/L respective to the observed and projected (2071-2100) mean RWT levels of

seven catchments.

6. In this study, the hybrid kNN-LSTM model (Chapter 4) is extended by including streamflow as feature variable in addition to AT to predict the RWT and compared the results with standalone LSTM, a modified nonlinear regression model, and 8-parameter version Air2Stream (Section 5.4). It was observed that kNN-LSTM was still producing good results, which agrees with the Chapter 4 studies.
7. In hypothetical climate change scenarios, RWT increases by 5 ^{o}C, and decreases DO by 4.8%–13.2%, thus lowering the DO's saturation capacity. The percentage of decrease of DO saturation levels with respect to increase of RWTs is higher for Musi and Ganga catchment sites for the selected climate change scenarios relative to the historical values noted as 9.21, and 13.22 %, respectively.
8. Additionally, it was found that streamflow played a minimal role in ML models for RWT predictions for the selected catchments in India.
9. Overall, for every 1 ^{o}C RWT increase, there will be about 2.3 % decrease in DO saturation level concentrations over Indian catchments under climate signals.

Overall, the thesis describes methodologies for prediction of RWQ variables under data uncertainties using the hybrid ML algorithms and further assessed the climate change impacts on RWQ variables. The proposed methods, demonstrated methodologies, frameworks are generic and can be implementable for any given river water quality variables for river water quality management.

6.2. Limitations and Future Work

The methods described in this thesis are generic, can be extended and further improved to handle different types of data and tasks. Also, the results of this thesis suggest a number of interesting research avenues for future work. Discussed some methodological improvements and potential new directions here.

1. The methodologies described in Chapter 3 and Chapter 4 are considered AT alone as the input variable, and in Chapter 5 is considered streamflow as feature variable in addition to AT to predict the water quality variables. However, other feature variables such as radiation, wind speed, etc., were not examined in this work due to a lack of complete data. Therefore, the hybrid modeling framework can be developed by integrating solar radiation, wind speed, etc., as model inputs which is potential future research.
2. The methodologies described in Chapter 4 are considered the kNN bootstrap resampling algorithm to simulate the water quality data by feeding the limited number of observational water quality data. As part of surrogate modeling, use process-based model component to simulate the water quality data and use these data points as training samples for the data-driven component forms another potential future research.
3. The methodologies described in Chapter 4 is considered the Daubechies wavelet of order 5 (DB5), was chosen to train the WT-LSTM model. Using different types of DWTs to train the WT-LSTM model and assess the model performances leads to potential future research.
4. It can be observed that inherent uncertainties from each of the DL models can accumulate and can affect the final performance measures. Furthermore, each DL model can predict unique water quality value, leading to model uncertainty. Furthermore, to address such model uncertainties originating from various DL, the ensemble of DL models, stacking algorithms, etc. (Song et al. 2020; Piotrowski et al. 2021) can be adopted as an extension to current methodology, which can combine water quality variable predictions from various DL models.
5. Lastly, throughout the thesis the proposed modeling framework is demonstrated with the RWT and DO saturation levels as the river water quality variables for seven majorly polluted catchments of India. However, the methodologies developed in this thesis are generic and are applicable for other water quality variables.

LIST OF PUBLICATIONS RESULTING FROM THE THESIS

Peer Reviewed Journal Publications

1. Rehana, S., Rajesh, M., 2023 Assessment of Impacts of Climate Change on Indian Riverine Thermal Regimes using Hybrid Deep Learning Methods. ***Water Resource Research*****, Impact factor (5.4),** https://doi.org/10.1029/2021WR031347/
2. Rajesh, M., Rehana, S., 2022. Impact of Climate Change on River Water Temperature and Dissolved Oxygen with Sparse River Water Quality Data: Indian Riverine Thermal Regimes, ***Nature Scientific Reports*****, Impact factor (4.6),** https://doi.org/10.1038/s41598-022-12996-7
3. Rajesh, M., Rehana, S., 2021. Prediction of river water temperature using machine learning algorithms: a tropical river system of India. ***Journal of Hydroinformatics,*** **Impact factor (2.7),** https://doi.org/10.2166/hydro.2021.121

Publication in International Conference Proceedings

- Rajesh, M., Rizwan, M. L., Jaswanth, N., Rehana, S., 2023. Modeling and Prediction of Feature Centric River Water Temperature using Machine Learning Algorithms in Data Scarce Regions. HYDRO 2023 International, 28th International Conference on Hydraulics, Water Resources, River and Coastal Engineering, 21-23 December 2023, NIT WARANGAL, INDIA.
- Rajesh, M., Rehana, S., 2021. Prediction of River Water Temperature Using the Coupling Support Vector Regression and Data Assimilation Technique – Tropical River System of India. International Conference, Asia Oceania Geosciences Society (AOGS) 2021 Virtual 18th Annual Meeting, August 01 - 06, 2021, Asia Geosciences Society. https://doi.org/10.1142/9789811260100_0021

OTHER PUBLICATIONS

Peer Reviewed Journal Publications

- Rajesh, M., Indranil, P., Ebrahim A., Shailesh K. S., Rehana, S., 2022. Short-range Reservoir Inflow Forecasting using Hydrological and Large-Scale Atmospheric Circulation Information, *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2022.128153
- Rajesh, M., Anishka, S., Satyam V. P., Arohi, S., Rehana, S., 2022. Improving Short-range Reservoir Inflow Forecasts with Machine Learning Model Combination, Water Resources Management. https://link.springer.com/article/10.1007/s11269-022-03356-1
- Rajesh, M., Abhishek, R. V., Jashwanth, K. S., Ghouse, B. S., Rehana, S., 2021. Prediction of Land Surface Temperature of Major Coastal Cities of India using Bidirectional LSTM Neural Networks, *Journal of Water and Climate Change*, 2021, https://doi.org/10.2166/wcc.2021.460
- Adeeba Ayaz, Maddu Rajesh, Shailesh Kumar Singh, Shaik Rehana. Estimation of reference evapotranspiration using machine learning models with limited data. *AIMS Geosciences*,2021,7(3):268-290. http://www.aimspress.com/article/doi/10.3934/geosci.2021016

Publication in International Conference Proceedings

- Rajesh, M., Indranil, P., Rehana, S., 2021. Reservoir Inflow Forecasting Based on Gradient Boosting Regressor Model - A Case Study of Bhadra Reservoir, India. International Conference, Asia Oceania Geosciences Society (AOGS) 2021 Virtual 18th Annual Meeting, August 01 - 06, 2021, Asia Geosciences Society. https://doi.org/10.1142/9789811260100_0022

Publication in National Conference Proceedings

- Rajesh, M., Krishnamohan, G., Rehana, S., and Dhanya, C.T., 2021. Impact Assessment of Environmental Flows using CORDEX Regional Climate Models: Case Study of Nagarjuna Sagar Dam, Krishna River, India, National Conference

on “Advanced Modelling and Innovations in Water Resources Engineering" (AMIWRE-2021), February 20 - 21, 2021, Department of Civil Engineering, NIT, Jamshedpur, Jharkhand, India. https://link.springer.com/chapter/10.1007/978-981-16-4629-4_14

Book Chapters:

- Rajesh, M., Rehana, S., and Dhanya, C.T., 2020. Environmental Flow Impacts on Water Quality of Peninsular River System: Tunga-Bhadra River, India. In: Mukherjee, A. (eds) Riverine Systems. Springer, Cham. https://doi.org/10.1007/978-3-030-87067-6_12

BIBLIOGRAPHY

Ahmadi-Nedushan B, St-Hilaire A, Ouarda TBMJ, et al (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

Alcocer J, Quiroz-Martínez B, Merino-Ibarra M, et al (2022) Using Wavelet Analysis to Examine Long-Term Variability of Phytoplankton Biomass in the Tropical, Saline Lake Alchichica, Mexico. Water 14:1346. https://doi.org/10.3390/w14091346

Almeida MC, Coelho PS (2022) Modeling river water temperature with limiting forcing data: air2stream v1.0.0, machine learning and multiple regression. Geoscientific Model Development Discussions 1–35. https://doi.org/10.5194/gmd-2022-206

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

Arnell N (1996) Global Warming, river flows and water resources

Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM (2021) River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. Journal of Environmental Chemical Engineering 9:104599. https://doi.org/10.1016/j.jece.2020.104599

Azad A, Karami H, Farzin S, et al (2019) Modeling river water quality parameters using modified adaptive neuro fuzzy inference system. Water Science and Engineering 12:45–54. https://doi.org/10.1016/j.wse.2018.11.001

Babovic V (2005) Data mining in hydrology - Invited Commentary. 1511–1515

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

Bates BC, Kundzewicz ZW, Wu S, Palutikof JP (2008) Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change

Bayram A, Uzlu E, Kankal M, Dede T (2015) Modeling stream dissolved oxygen concentration using teaching–learning based optimization algorithm. Environ Earth Sci 73:6565–6576. https://doi.org/10.1007/s12665-014-3876-3

Beersma J, Buishand T (2004) Joint probability of precipitation and discharge deficits in the Netherlands. Water Resour Res 40:. https://doi.org/10.1029/2004WR003265

Bengio Y, Courville A, Vincent P (2013) Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35:1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Beven K (2016) Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. null 61:1652–1665. https://doi.org/10.1080/02626667.2015.1031761

Bogan T, Mohseni O, Stefan HG (2003) Stream temperature-equilibrium temperature relationship. Water Resources Research 39:. https://doi.org/10.1029/2003WR002034

Boukabara S-A, Krasnopolsky V, Penny SG, et al (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

Brown GW (1972) Improved temperature prediction model for small streams. Research Report WRPI-16:

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:

Central Water Commission (2019) Effect of Time and Temperature on DO Levels in River Waters. http://cwc.gov.in/sites/default/files/effect-time-and-temperature-do-levels-river-water-2019.pdf:

Centre for Climate Change Research (CCCR) (2017) NEX-GDDP Data, Centre for Climate Change Research, Pune, India

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

Chapra SC (1998) Surface Water Quality Modelling. McGraw Hill Kogakusha Ltd New York

Chapra SC, Camacho LA, McBride GB (2021) Impact of Global Warming on Dissolved Oxygen and BOD Assimilative Capacity of the World's Rivers: Modeling Analysis. Water 13:2408. https://doi.org/10.3390/w13172408

Chapra SC, Clough DE (2021) Applied Numerical Methods with Python for Engineers and Scientists. WCB/McGraw-Hill: New York, NY, USA

Chaudhary S, Dhanya CT, Kumar A, Shaik R (2019) Water Quality–Based Environmental Flow under Plausible Temperature and Pollution Scenarios. Journal of Hydrologic Engineering 24:05019007. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001780

Chen D, Hu M, Guo Y, Dahlgren RA (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

Cho K, van Merriënboer B, Gulcehre C, et al (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1724–1734

Cibin R, Sudheer KP, 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 HL, 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

Cole JC, Maloney KO, Schmid M, McKenna JE (2014) Developing and testing temperature models for regulated systems: A case study on the Upper Delaware River. Journal of Hydrology 519:588–598. https://doi.org/10.1016/j.jhydrol.2014.07.058

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

Cox BA, Whitehead PG (2009) Impacts of climate change scenarios on dissolved oxygen in the River Thames, UK. Hydrology Research 40:138–152. https://doi.org/10.2166/nh.2009.096

CPCB (2020) Central Pollution Cotrol Board. Wikipedia, The Free Encyclopedia

CPCB (2019) Guidelines for water quality management. Delhi, India

CPCB (2015) Central Pollution Cotrol Board

Cunderlik JM, Simonovic SP (2005) Hydrological extremes in a southwestern Ontario river basin under future climate conditions/Extrêmes hydrologiques dans un basin versant du sud-ouest de l'Ontario sous conditions climatiques futures. Hydrological Sciences Journal 50:null-654. https://doi.org/10.1623/hysj.2005.50.4.631

Danladi Bello A-A, Hashim NB, Mohd Haniffah MR (2017) Predicting Impact of Climate Change on Water Temperature and Dissolved Oxygen in Tropical Rivers. Climate 5:58. https://doi.org/10.3390/cli5030058

Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory 36:961–1005. https://doi.org/10.1109/18.57199

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

Dibike Y, Velickov S, Solomatine D, Abbott M (2001) Model Induction With Support Vector Machines: Introduction and Applications. Journal of Computing in Civil Engineering - J COMPUT CIVIL ENG 15:. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208)

Dingman SL (1972) Equilibrium temperatures of water surfaces as related to air temperature and solar radiation. Water Resources Research 8:42–49. https://doi.org/10.1029/WR008i001p00042

Döll P, Zhang J (2010) Impact of climate change on freshwater ecosystems: a global-scale analysis of ecologically relevant river flow alterations. Hydrology and Earth System Sciences 14:783–799. https://doi.org/10.5194/hess-14-783-2010

Drainas K, Kaule L, Mohr S, et al (2023) Predicting stream water temperature with artificial neural networks based on open-access data. Hydrological Processes 37:e14991. https://doi.org/10.1002/hyp.14991

Du X, Shrestha NK, Ficklin DL, Wang J (2018) Incorporation of the equilibrium temperature approach in a Soil and Water Assessment Tool hydroclimatological stream temperature model. Hydrology and Earth System Sciences 22:2343–2357. https://doi.org/10.5194/hess-22-2343-2018

Du X, Shrestha NK, Wang J (2019) Assessing climate change impacts on stream temperature in the Athabasca River Basin using SWAT equilibrium temperature model and its potential impacts on stream ecosystem. Science of The Total Environment 650:1872–1881. https://doi.org/10.1016/j.scitotenv.2018.09.344

Dugan PR (1972) Pollution and Accelerated Eutrophication of Lakes. In: Dugan PR (ed) Biochemical Ecology of Water Pollution. Springer US, Boston, MA, pp 138–147

Dugdale SJ, Hannah DM, Malcolm IA (2017) River temperature modelling: A review of process-based approaches and future directions. Earth-Science Reviews 175:97–113. https://doi.org/10.1016/j.earscirev.2017.10.009

Ebrahimi H, Rajaee T (2017) Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. Global and Planetary Change 148:181–191. https://doi.org/10.1016/j.gloplacha.2016.11.014

Edinger JE, Duttweiler DW, Geyer JC (1968) The Response of Water Temperatures to Meteorological Conditions. Water Resources Research 4:1137–1143. https://doi.org/10.1029/WR004i005p01137

El-Jabi N, Caissie D, Turkkan N (2014) Water Quality Index Assessment under Climate Change. Journal of Water Resource and Protection 6:533–542. https://doi.org/10.4236/jwarp.2014.66052

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

Färber C, Lisniak D, Saile P, et al (2018) Water quality at the global scale: GEMStat database and information system. 20:15984

Federal Interagency Stream Restoration Working Group (FISRWG) (2014) Stream Corridor Restoration: Principles, Processes and Practices. ederal Interagency Stream Restoration Working Group (FISRWG) USDA, Washington, DC:

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

Ficklin DL, Luo Y, Stewart IT, Maurer EP (2012) Development and application of a hydroclimatological stream temperature model within the Soil and Water Assessment Tool. Water Resources Research 48:. https://doi.org/10.1029/2011WR011256

Ficklin DL, Stewart IT, Maurer EP (2013) Effects of climate change on stream temperature, dissolved oxygen, and sediment concentration in the Sierra Nevada in California. Water Resources Research 49:2765–2782. https://doi.org/10.1002/wrcr.20248

Gallice A, Schaefli B, Lehning M, et al (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 SJ, 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?

Graf R (2018) Analysis of Granger causality between daily and monthly temperatures of water and air, as illustrated with the example of Noteć River. 18:101–117. https://doi.org/10.15576/ASP.FC/2018.17.3.101

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

Granger CWJ (1969) Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 37:424–438. https://doi.org/10.2307/1912791

Greff K, Srivastava RK, Koutník J, et al (2017) LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems 28:2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Gu RR, 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, et al (2017) Water temperature increases in the river Rhine in response to climate change. Reg Environ Change 17:299–308. https://doi.org/10.1007/s10113-016-1006-3

Harvey R, Lye L, Khan A, Paterson R (2011) The Influence of Air Temperature on Water Temperature and the Concentration of Dissolved Oxygen in Newfoundland Rivers. Canadian Water Resources Journal / Revue canadienne des ressources hydriques 36:171–192. https://doi.org/10.4296/cwrj3602849

Heddam S, Kim S, Elbeltagi A, et al (2022) Predicting nitrate concentration in river using advanced

artificial intelligence techniques: extreme learning machines versus deep learning. pp 121–153

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

Hilborn R, Mangel M (1997) The Ecological Detective: Confronting Models With Data

Hirsch RM, Archfield SA, De Cicco LA (2015) A bootstrap method for estimating uncertainty of water quality trends. Environmental Modelling & Software 73:148–166. https://doi.org/10.1016/j.envsoft.2015.07.017

Hochreiter S (1998) The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. Int J Unc Fuzz Knowl Based Syst 06:107–116. https://doi.org/10.1142/S0218488598000094

Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Computation 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hochreiter S, Younger AS, Conwell PR (2001) Learning to Learn Using Gradient Descent. In: Dorffner G, Bischof H, Hornik K (eds) Artificial Neural Networks — ICANN 2001. Springer, Berlin, Heidelberg, pp 87–94

Hoerl AE, Kennard RW (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

Honorato AG da SM, Silva GBL da, Santos CAG (2018) Monthly streamflow forecasting using neuro-wavelet techniques and input analysis. Hydrological Sciences Journal 63:2060–2075. https://doi.org/10.1080/02626667.2018.1552788

Houghton JT, Meira Filho LG, Callander BA, et al (1996) Climate change 1995 : Cambridge University Press, for the Intergovernmental Panel on Climate Change,

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

Huang NE, Shen Z, Long SR, et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences 454:903–995. https://doi.org/10.1098/rspa.1998.0193

Hyndman RJ, Athanasopoulos G, Bergmeir C, et al (2018) forecast: Forecasting functions for time series and linear models. R package:

IMD (2021) Indian Meteorological Department, Ministry of Earth Sciences, Government of India

Intergovernmental Panel on Climate Change (2007) Climate Change2007: The Physical Science Basis. Contribution of Working Group I tothe Fourth Assessment Report of the Intergovernmental Panel on ClimateChange, Cambridge Univ. Press, Cambridge, U. K.:

Intergovernmental Panel on Climate Change (IPCC) (2001) The scientific basis: third assessment report of the intergovernmental panel on climate change.

Isaak DJ, Luce CH, Horan DL, et al (2020) Thermal Regimes of Perennial Rivers and Streams in the Western United States. JAWRA Journal of the American Water Resources Association 56:842–867. https://doi.org/10.1111/1752-1688.12864

Isaak DJ, Luce CH, Rieman BE, et al (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 DJ, Wenger SJ, Peterson EE, et al (2017) The NorWeST Summer Stream Temperature Model and Scenarios for the Western U.S.: A Crowd-Sourced Database and New Geospatial Tools Foster a User Community and Predict Broad Climate Warming of Rivers and Streams. Water Resources Research 53:9181–9205. https://doi.org/10.1002/2017WR020969

Isaak DJ, 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

Islam SU, Hay RW, Déry SJ, Booth BP (2019) Modelling the impacts of climate change on riverine thermal regimes in western Canada's largest Pacific watershed. Scientific Reports 9:11398. https://doi.org/10.1038/s41598-019-47804-2

Jackson FL, Fryer RJ, Hannah DM, et al (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

Jackson FL, Fryer RJ, Hannah DM, Malcolm IA (2017) Can spatial statistical river temperature models be transferred between catchments? Hydrology and Earth System Sciences 21:4727–4745. https://doi.org/10.5194/hess-21-4727-2017

Jiang T, Chen YD, Xu C, et al (2007) Comparison of hydrological impacts of climate change simulated by six hydrological models in the Dongjiang Basin, South China. Journal of Hydrology 336:316–333. https://doi.org/10.1016/j.jhydrol.2007.01.010

Johnson MF, Wilby RL, Toone JA (2014) Inferring air–water temperature relationships from river and catchment properties. Hydrological Processes 28:2912–2928. https://doi.org/10.1002/hyp.9842

Jones DR, Schonlau M, Welch WJ (1998) Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization 13:455–492. https://doi.org/10.1023/A:1008306431147

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

Kauffman GJ (2018) The Cost of Clean Water in the Delaware River Basin (USA). Water 10:95. https://doi.org/10.3390/w10020095

Kermorvant C, Liquet B, Litt G, et al (2021) Reconstructing Missing and Anomalous Data Collected from High-Frequency In-Situ Sensors in Fresh Waters. International Journal of Environmental Research and Public Health 18:12803. https://doi.org/10.3390/ijerph182312803

Khani S, Rajaee T (2017) Modeling of Dissolved Oxygen Concentration and Its Hysteresis Behavior in Rivers Using Wavelet Transform-Based Hybrid Models. CLEAN – Soil, Air, Water 45:. https://doi.org/10.1002/clen.201500395

Khorsandi M, St-Hilaire A, Arsenault R, et al (2023) Future flow and water temperature scenarios in an impounded drainage basin: implications for summer flow and temperature management downstream of the dam. Climatic Change 176:164. https://doi.org/10.1007/s10584-023-03634-w

Kirchgässner G, Wolters J, Hassler U (2013) Introduction to Modern Time Series Analysis

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 HR (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

Kratzert F, Klotz D, Brenner C, et al (2018) Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences 22:6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert F, Klotz D, Shalev G, et al (2019) Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences 23:5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Krider LA, Magner JA, Perry J, et al (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

Kushner HJ (1964) A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. Journal of Basic Engineering 86:97–106. https://doi.org/10.1115/1.3653121

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

Labat D (2005) Recent advances in wavelet analyses: Part 1. A review of concepts. Journal of Hydrology 314:275–288. https://doi.org/10.1016/j.jhydrol.2005.04.003

Laizé CLR, Bruna Meredith C, Dunbar MJ, Hannah DM (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

Lall U, Sharma A (1996) A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. Water Resources Research 32:679–693. https://doi.org/10.1029/95WR02966

Leander R, Buishand A, Aalders P, Wit MD (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:null-1103. https://doi.org/10.1623/hysj.2005.50.6.1089

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539

Lee K-H, Cho H-Y (2015) Projection of Climate-Induced Future Water Temperature for the Aquatic Environment. Journal of Environmental Engineering 141:06015004. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000974

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 M, Zhang Y, Wallace J, Campbell E (2020a) 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

Li W, Kiaghadi A, Dawson C (2020b) High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. Neural Comput & Applic. https://doi.org/10.1007/s00521-020-05010-6

Li X-L, Lü H, Horton R, et al (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

Lima AR, Cannon AJ, Hsieh WW (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

Liu D, Yu Z, L H (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

Loucks DP, van Beek E (2017a) Water Quality Modeling and Prediction. In: Loucks DP, van Beek E (eds) Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications. Springer International Publishing, Cham, pp 417–467

Loucks DP, van Beek E (2017b) Water Resources Planning and Management: An Overview. In: Loucks DP, van Beek E (eds) Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications. Springer International Publishing, Cham, pp 1–49

Lowney CL (2000) Stream temperature variation in regulated rivers: Evidence for a spatial pattern in daily minimum and maximum magnitudes. Water Resources Research 36:2947–2955. https://doi.org/10.1029/2000WR900142

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

Lundqvist J (2009) Water as a Human Resource. In: Encyclopedia of Inland Waters. Oxford: Academic Press., pp 31–42

Magnuson JJ, Crowder Rowder LB, Medvick PA (1979) Temperature as an Ecological Resource. American Zoologist 19:331–343. https://doi.org/10.1093/icb/19.1.331

Mahalanobis PC (1936) On the generalized distance in statistics. National Institute of Science of India

Malmaeus JM, Blenckner T, Markensten H, Persson I (2006) Lake phosphorus dynamics and climate warming: A mechanistic model approach. Ecological Modelling 190:1–14. https://doi.org/10.1016/j.ecolmodel.2005.03.017

Marinov I (1990) Engineering Hydrology. Technoka, Sofia

Maurer EP, Hidalgo HG (2008) Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrology and Earth System Sciences 12:551–563. https://doi.org/10.5194/hess-12-551-2008

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

Milly P, Betancourt J, Julio, et al (2008) Stationarity Is Dead: Whither Water Management? Science 319:573–574

Mockus J (1989) Global Optimization and the Bayesian Approach. In: Mockus J (ed) Bayesian Approach to Global Optimization: Theory and Applications. Springer Netherlands, Dordrecht, pp 1–3

Mohseni O, Erickson TR, Stefan HG (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

Mohseni O, Stefan HG, Erickson TR (1998) A nonlinear regression model for weekly stream temperatures. Water Resources Research 34:2685–2692. https://doi.org/10.1029/98WR01877

Montalvo C, García-Berrocal A (2015) Improving the in situ measurement of RTD response times through Discrete Wavelet Transform in NPP. Annals of Nuclear Energy 80:114–122. https://doi.org/10.1016/j.anucene.2015.02.004

Moriasi D, Arnold J, Van Liew M, et al (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 JC, Bales RC, Conklin MH (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 MGG (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

Mujumdar PP, Nagesh Kumar D (2012) Floods in a changing climate: hydrologic modeling

Mujumdar PP, Subbarao Vemula VR (2004) Fuzzy Waste Load Allocation Model: Simulation-Optimization Approach. Journal of Computing in Civil Engineering 18:120–131. https://doi.org/10.1061/(ASCE)0887-3801(2004)18:2(120)

Muluye GY (2012) Comparison of statistical methods for downscaling daily precipitation. Journal of Hydroinformatics 14:1006–1023. https://doi.org/10.2166/hydro.2012.197

Nagesh Kumar D, Srinivasa Raju K, Sathish T (2004) River Flow Forecasting using Recurrent Neural Networks. Water Resources Management 18:143–161. https://doi.org/10.1023/B:WARM.0000024727.94701.12

Nash JE, Sutcliffe JV (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

National River Conservation Directorate (NRCD) (2018) National River Conservation Directorate, Government of India. https://nrcd.nic.in/writereaddata/FileUpload/River_STRETCHES_Sept_2018.pdf

Nelson KC, Palmer MA (2007) Stream Temperature Surges Under Urbanization and Climate Change: Data, Models, and Responses1. JAWRA Journal of the American Water Resources Association 43:440–452. https://doi.org/10.1111/j.1752-1688.2007.00034.x

Neumann David W., Rajagopalan Balaji, 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

Nouraki A, Alavi M, Golabi M, Albaji M (2021) Prediction of water quality parameters using machine learning models: a case study of the Karun River, Iran. Environ Sci Pollut Res 28:57060–57072. https://doi.org/10.1007/s11356-021-14560-8

Nourani V, Alami MT, Aminfar MH (2009) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Engineering Applications of Artificial Intelligence 22:466–472. https://doi.org/10.1016/j.engappai.2008.09.003

Null SE, Viers JH, Deas ML, et al (2013) Stream temperature sensitivity to climate warming in California's Sierra Nevada: impacts to coldwater habitat. Climatic Change 116:149–170. https://doi.org/10.1007/s10584-012-0459-8

Olah C (2015) Understanding LSTM Networks

Orr HG, Simpson GL, Clers S des, et al (2015) Detecting changing river temperatures in England and Wales. Hydrological Processes 29:752–766. https://doi.org/10.1002/hyp.10181

Ouarda TBMJ, 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

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

Paily PP, Macagno EO, Kennedy JF (1974) Winter-Regime Thermal Response of Heated Streams. Journal of the Hydraulics Division 100:531–551. https://doi.org/10.1061/JYCEAJ.0003931

Pappenberger F, Beven KJ, 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, et al (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12

Penchev P (1972) General Hydrology. NI, Sofia

Piccolroaz S, Calamita E, Majone B, et al (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, et al (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 JM, Fang X, Stefan HG (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 AP, Napiorkowski JJ (2019) Simple modifications of the nonlinear regression stream temperature model for daily data. Journal of Hydrology 572:308–328. https://doi.org/10.1016/j.jhydrol.2019.02.035

Piotrowski AP, Napiorkowski JJ (2018) Performance of the air2stream model that relates air and stream water temperatures depends on the calibration method. Journal of Hydrology 561:395–412. https://doi.org/10.1016/j.jhydrol.2018.04.016

Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ, 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

Piotrowski AP, Osuch M, Napiorkowski JJ (2021) Influence of the choice of stream temperature model on the projections of water temperature in rivers. Journal of Hydrology 601:126629. https://doi.org/10.1016/j.jhydrol.2021.126629

Pohle I, Helliwell R, Aube C, et al (2019) Citizen science evidence from the past century shows that Scottish rivers are warming. Science of The Total Environment 659:53–65. https://doi.org/10.1016/j.scitotenv.2018.12.325

Qiu R, Wang Y, Rhoads B, et al (2021) River water temperature forecasting using a deep learning

method. Journal of Hydrology 595:126016. https://doi.org/10.1016/j.jhydrol.2021.126016

Qiu R, Wang Y, Wang D, et al (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

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

Raghavan SV, Hur J, Liong S-Y (2018) Evaluations of NASA NEX-GDDP data over Southeast Asia: present and future climates. Climatic Change 148:503–518. https://doi.org/10.1007/s10584-018-2213-3

Rahman ATMS, Hosono T, Quilty JM, et al (2020) Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms. Advances in Water Resources 141:103595. https://doi.org/10.1016/j.advwatres.2020.103595

Rajagopalan B, Erkyihun ST, Lall U, et al (2019) A Nonlinear Dynamical Systems-Based Modeling Approach for Stochastic Simulation of Streamflow and Understanding Predictability. Water Resources Research 55:6268–6284. https://doi.org/10.1029/2018WR023650

Rajagopalan B, Lall U (1999) A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water Resources Research 35:3089–3101. https://doi.org/10.1029/1999WR900028

Rajesh M, Pradhan I, Ahmadisharaf E, et al (2022) Short-range reservoir inflow forecasting using hydrological and large-scale atmospheric circulation information. Journal of Hydrology 612:128153. https://doi.org/10.1016/j.jhydrol.2022.128153

Rangeti I, Dzwairo B, Barratt GJ, et al (2015) Validity and Errors in Water Quality Data — A Review. IntechOpen

Raseman WJ, Rajagopalan B, Kasprzyk JR, Kleiber W (2020) Nearest neighbor time series bootstrap for generating influent water quality scenarios. Stoch Environ Res Risk Assess 34:23–31. https://doi.org/10.1007/s00477-019-01762-3

Rasouli K, Hsieh WW, Cannon AJ (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

Read JS, Jia X, Willard J, et al (2019) Process-Guided Deep Learning Predictions of Lake Water Temperature. Water Resources Research 55:9173–9190. https://doi.org/10.1029/2019WR024922

Rehana S (2019) River Water Temperature Modelling Under Climate Change Using Support Vector Regression. In: Singh SK, Dhanya CT (eds) Hydrology in a Changing World: Challenges in Modeling. Springer International Publishing, Cham, pp 171–183

Rehana S, Dhanya CT (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, 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

Rehana S, Mujumdar PP (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 PP (2014) Basin Scale Water Resources Systems Modeling Under Cascading Uncertainties. Water Resour Manage 28:3127–3142. https://doi.org/10.1007/s11269-014-0659-2

Rehana S, Mujumdar PP (2009) An imprecise fuzzy risk approach for water quality management of a river system. Journal of Environmental Management 90:3653–3664. https://doi.org/10.1016/j.jenvman.2009.07.007

Rice EW, Baird RB, Eaton AD, Eds. (2017) American Public Health Association (APHA). Standard Methods for the Examination of Water and Wastewater, 23rd ed APHA: Washington, DC, USA:

Rice KC, Jastram JD (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

Roushangar K, Nourani V, Alizadeh F (2018) A multiscale time-space approach to analyze and categorize the precipitation fluctuation based on the wavelet transform and information theory concept. Hydrology Research 49:724–743. https://doi.org/10.2166/nh.2018.143

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. In: Parallel Distributed Processing. MIT Press, pp 318–362

Sahoo GB, Schladow SG, Reuter JE (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, Annoni P, Azzini I, et al (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

Saltelli A, Ratto M, Andres T, et al (2008) Global Sensitivity Analysis. The Primer

Sang Y-F, Singh VP, Sun F, et al (2016) Wavelet-Based Hydrological Time Series Forecasting. Journal of Hydrologic Engineering 21:06016001. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001347

Sang Y-F, Singh VP, Wen J, Liu C (2015) Gradation of complexity and predictability of hydrological processes. Journal of Geophysical Research: Atmospheres 120:5334–5343. https://doi.org/10.1002/2014JD022844

Sang Y-F, Sun F, Singh VP, et al (2018) A discrete wavelet spectrum approach for identifying non-monotonic trends in hydroclimate data. Hydrology and Earth System Sciences 22:757–766. https://doi.org/10.5194/hess-22-757-2018

Sang Y-F, Wang D, Wu J-C, et al (2009) The relation between periods' identification and noises in hydrologic series data. Journal of Hydrology 368:165–177. https://doi.org/10.1016/j.jhydrol.2009.01.042

Sang Y-F, Wang D, Wu J-C, et al (2011) Wavelet-Based Analysis on the Complexity of Hydrologic Series Data under Multi-Temporal Scales. Entropy 13:195–210. https://doi.org/10.3390/e13010195

Santy S, Mujumdar P, Bala G (2020) Potential Impacts of Climate and Land Use Change on the Water Quality of Ganga River around the Industrialized Kanpur Region. Sci Rep 10:9107. https://doi.org/10.1038/s41598-020-66171-x

Scavuzzo JM, Trucco F, Espinosa M, et al (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

Seo Y, Kim S, Kisi O, Singh VP (2015) Daily water level forecasting using wavelet decomposition and artificial intelligence techniques. Journal of Hydrology 520:224–243. https://doi.org/10.1016/j.jhydrol.2014.11.050

Shaik R, Mujumdar 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

Sharif M, Burn DH (2007) Improved K -Nearest Neighbor Weather Generating Model. Journal of Hydrologic Engineering 12:42–51. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:1(42)

Sharma A, O'Neill R (2002) A nonparametric approach for representing interannual dependence in monthly streamflow sequences. Water Resources Research 38:5-1-5–10. https://doi.org/10.1029/2001WR000953

Shen C (2018) A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. Water Resources Research 54:8558–8593. https://doi.org/10.1029/2018WR022643

Shiklomanov LA (1993) World Freshwater Resources. In: Gleick, P.H., Ed., Water in Crisis: A Guide to World's Freshwater Resources. Oxford University Press, New York, 13-24

Shoaib M, Shamseldin AY, Khan S, et al (2019) Input Selection of Wavelet-Coupled Neural Network Models for Rainfall-Runoff Modelling. Water Resour Manage 33:955–973. https://doi.org/10.1007/s11269-018-2151-x

Shoaib M, Shamseldin AY, Melville BW (2014) Comparative study of different wavelet based neural network models for rainfall–runoff modeling. Journal of Hydrology 515:47–58. https://doi.org/10.1016/j.jhydrol.2014.04.055

Shrestha RR, Pesklevits JC (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

Shrestha RR, Pesklevits JC (2023) Reconstructed River Water Temperature Dataset for Western Canada 1980–2018. Data 8:48. https://doi.org/10.3390/data8030048

Shrestha RR, Prowse TD, Tso L (2019) Modelling historical variability of phosphorus and organic carbon fluxes to the Mackenzie River, Canada. Hydrology Research 50:1424–1439. https://doi.org/10.2166/nh.2019.161

Sinokrot BA, Stefan HG (1993) Stream temperature dynamics: Measurements and modeling. Water Resources Research 29:2299–2312. https://doi.org/10.1029/93WR00540

Sobol I (1990) On sensitivity estimation for nonlinear mathematical models. Keldysh Applied Mathematics Institute 1:112–118

Sobol' IM (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 MM, Benjankar R, Tonina D, et al (2017) Estimation of daily stream water temperatures with a Bayesian regression approach. Hydrological Processes 31:1719–1733. https://doi.org/10.1002/hyp.11139

Sojka M, Ptak M (2022) Possibilities of River Water Temperature Reconstruction Using Statistical Models in the Context of Long-Term Thermal Regime Changes Assessment. Applied Sciences 12:7503. https://doi.org/10.3390/app12157503

Song T, Ding W, Liu H, et al (2020) Uncertainty Quantification in Machine Learning Modeling for Multi-Step Time Series Forecasting: Example of Recurrent Neural Networks in Discharge Simulations. Water 12:912. https://doi.org/10.3390/w12030912

Souza FA, 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

Srinivasa Raju K, Nagesh Kumar D (2018) Introduction. In: Srinivasa Raju K, Nagesh Kumar D (eds) Impact of Climate Change on Water Resources : With Modeling Techniques and Case Studies. Springer, Singapore, pp 1–25

Stajkowski S, Kumar D, Samui P, et al (2020) Genetic-Algorithm-Optimized Sequential Model for Water Temperature Prediction. Sustainability 12:5374. https://doi.org/10.3390/su12135374

Stefan HG, Preud'homme EB (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 HG, Sinokrot BA (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

Svendsen MBS, Bushnell PG, Christensen E a. F, Steffensen JF (2016) Sources of variation in oxygen consumption of aquatic animals demonstrated by simulated constant oxygen consumption and respirometers of different sizes. Journal of Fish Biology 88:51–64. https://doi.org/10.1111/jfb.12851

Tabari H, Hosseinzadeh Talaee P (2015) Reconstruction of river water quality missing data using artificial neural networks. Water Quality Research Journal 50:326–335. https://doi.org/10.2166/wqrjc.2015.044

Tan R, Perkowski M (2015) Wavelet-Coupled Machine Learning Methods for Drought Forecast Utilizing Hybrid Meteorological and Remotely-Sensed Data. Proceedings of the International Conference on Data Mining (DMIN'15)

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

Taniwaki RH, Piggott JJ, Ferraz SFB, Matthaei CD (2017) Climate change and multiple stressors in small tropical streams. Hydrobiologia 793:41–53. https://doi.org/10.1007/s10750-016-2907-3

Tavares MH, Cunha AHF, Motta-Marques D, et al (2020) Derivation of consistent, continuous daily river temperature data series by combining remote sensing and water temperature models. Remote Sensing of Environment 241:111721. https://doi.org/10.1016/j.rse.2020.111721

Tehrany MS, Pradhan B, Jebur MN (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 Sci Technol 77:1724–1733. https://doi.org/10.2166/wst.2018.058

Tennant DL (1976) Instream Flow Regimens for Fish, Wildlife, Recreation and Related Environmental Resources. Fisheries 1:6–10. https://doi.org/10.1577/1548-8446(1976)001<0006:IFRFFW>2.0.CO;2

Thomann RV, Mueller JA (1987) Principles of surface water quality modeling and control. Harper & Row, New York

Thrasher B, Maurer EP, McKellar C, Duffy PB (2012) Technical Note: Bias correcting climate model simulated daily temperature extremes with quantile mapping. Hydrology and Earth System Sciences 16:3309–3314. https://doi.org/10.5194/hess-16-3309-2012

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

Tsachev T, Ivanov K, Pechinov D, Totev I (1982) Thermic pollution of the rivers in Bulgaria. BAS, Sofia

Ubah JI, Orakwe LC, Ogbu KN, et al (2021) Forecasting water quality parameters using artificial neural network for irrigation purposes. Sci Rep 11:24438. https://doi.org/10.1038/s41598-021-04062-5

United States Geological Survey (USGS) (2018a) Temperature and Water. Water Science School

United States Geological Survey (USGS) (2018b) Dissolved Oxygen and Water. Water Science School

van Vliet MTH, Franssen WHP, Yearsley JR, et al (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 Vliet MTH, Jones ER, Flörke M, et al (2021) Global water scarcity including surface water quality and expansions of clean water technologies. Environ Res Lett 16:024020. https://doi.org/10.1088/1748-9326/abbfc3

van Vliet MTH, Ludwig F, Zwolsman JJG, et al (2011) Global river temperatures and sensitivity to

atmospheric warming and changes in river flow. Water Resources Research 47:. https://doi.org/10.1029/2010WR009198

van Vliet MTH, Yearsley J, Franssen W, et al (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 MTH, Zwolsman JJG (2008) Impact of summer droughts on the water quality of the Meuse river. Journal of Hydrology 353:1–17. https://doi.org/10.1016/j.jhydrol.2008.01.001

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 SE, 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. MIT Press, Denver, Colorado, 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

Virro H, Amatulli G, Kmoch A, et al (2021) GRQA: Global River Water Quality Archive. Earth System Science Data Discussions 1–30. https://doi.org/10.5194/essd-2021-51

Wang D, Borthwick AG, He H, et al (2018) A hybrid wavelet de-noising and Rank-Set Pair Analysis approach for forecasting hydro-meteorological time series. Environmental Research 160:269–281. https://doi.org/10.1016/j.envres.2017.09.033

Wang L, Xu B, Zhang C, et al (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

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, 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

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 Y, Leung LR, McGREGOR JL, et al (2004) Regional Climate Modeling: Progress, Challenges, and Prospects. Journal of the Meteorological Society of Japan Ser II 82:1599–1628. https://doi.org/10.2151/jmsj.82.1599

Wang Y, Yuan Y, Pan Y, Fan Z (2020) Modeling Daily and Monthly Water Quality Indicators in a Canal

Using a Hybrid Wavelet-Based Support Vector Regression Structure. Water 12:1476. https://doi.org/10.3390/w12051476

Webb B (1992) Climate Change and the Thermal Regime of Rivers

Webb B w., Walling D e. (1993) Temporal variability in the impact of river regulation on thermal regime and some biological implications. Freshwater Biology 29:167–182. https://doi.org/10.1111/j.1365-2427.1993.tb00752.x

Webb BW, Clack PD, Walling DE (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

Webb BW, Nobilis F (2007) Long-term changes in river temperature and the influence of climatic and hydrological factors. Hydrological Sciences Journal 52:74–85. https://doi.org/10.1623/hysj.52.1.74

Wilby RL, Johnson MF (2020) Climate variability and implications for keeping rivers cool in England. Climate Risk Management 30:100259. https://doi.org/10.1016/j.crm.2020.100259

Wilby RL, Troni J, Biot Y, et al (2009) A review of climate risk information for adaptation and development planning. International Journal of Climatology 29:1193–1215. https://doi.org/10.1002/joc.1839

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Progress in Physical Geography: Earth and Environment 21:530–548. https://doi.org/10.1177/030913339702100403

WMO (1992) Simulated real-time intercomparison of hydrological models. WMO operational hydrology report (OHR), 38-WMO No779

WMO (1989) Calculation of monthly and annual 30-year standard normals. World Meteorological Organization Tech Doc 341 WCDP:10–11

Wood AW, Maurer EP, Kumar A, Lettenmaier DP (2002) Long-range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research: Atmospheres 107:ACL 6-1-ACL 6-15. https://doi.org/10.1029/2001JD000659

Xiang Z, Yan J, Demir I (2020) A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning. Water Resources Research 56:e2019WR025326. https://doi.org/10.1029/2019WR025326

Yang D, Peterson A (2017) River Water Temperature in Relation to Local Air Temperature in the Mackenzie and Yukon Basins. ARCTIC 70:47-58-47–58. https://doi.org/10.14430/arctic4627

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

Yang R, Wu S, Wu X, et al (2022) Quantifying the impacts of climate variation, damming, and flow regulation on river thermal dynamics: a case study of the Włocławek Reservoir in the Vistula River, Poland. Environmental Sciences Europe 34:3. https://doi.org/10.1186/s12302-021-00583-y

Yang T, Asanjan AA, Welles E, et al (2017) Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research 53:2786–2812. https://doi.org/10.1002/2017WR020482

Yates D, Gangopadhyay S, Rajagopalan B, Strzepek K (2003) A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resources Research 39:. https://doi.org/10.1029/2002WR001769

Yearsley JR (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, et al (2015) Hydrologic and Water Quality Models: Sensitivity. Trans ASABE 58:1721–1744. https://doi.org/10.13031/trans.58.10611

Zhang D, Peng Q, Lin J, et al (2019) Simulating Reservoir Operation Using a Recurrent Neural Network Algorithm. Water 11:865. https://doi.org/10.3390/w11040865

Zhu S, Bonacci O, Oskoruš D, et al (2019a) 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, Du X, Luo W (2019b) Incorporation of the simplified equilibrium temperature approach in a hydrodynamic and water quality model – CE-QUAL-W2. Water Supply 19:156–164. https://doi.org/10.2166/ws.2018.063

Zhu S, Hadzima-Nyarko M, Gao A, et al (2019c) Two hybrid data-driven models for modeling water-air temperature relationship in rivers. Environ Sci Pollut Res 26:12622–12630. https://doi.org/10.1007/s11356-019-04716-y

Zhu S, Heddam S, Nyarko EK, et al (2019d) Modeling daily water temperature for rivers: comparison between adaptive neuro-fuzzy inference systems and artificial neural networks models. Environ Sci Pollut Res 26:402–420. https://doi.org/10.1007/s11356-018-3650-2

Zhu S, Heddam S, Wu S, et al (2019e) Extreme learning machine-based prediction of daily water temperature for rivers. Environ Earth Sci 78:202. https://doi.org/10.1007/s12665-019-8202-7

Zhu S, Luo Y, Graf R, et al (2022) Reconstruction of long-term water temperature indicates significant warming in Polish rivers during 1966–2020. Journal of Hydrology: Regional Studies 44:101281. https://doi.org/10.1016/j.ejrh.2022.101281

Zhu S, Nyarko EK, Hadzima-Nyarko M, et al (2019f) 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, Nyarko EK, 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, Piotrowski AP (2020) River/stream water temperature forecasting using artificial intelligence models: a systematic review. Acta Geophys 68:1433–1442. https://doi.org/10.1007/s11600-020-00480-7