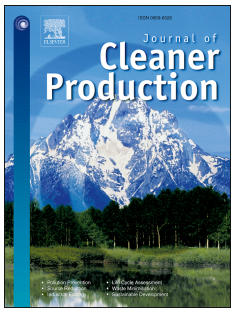
Journal Pre-proof



Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach

Mohammad Rezaie-Balf, Nasrin Fathollahzadeh Attar, Ardashir Mohammadzadeh, Muhammad Ary Murti, Ali Najah Ahmed, Chow Ming Fai, Narjes Nabipour, Sina Alaghmand, Ahmed El-Shafie

PII: S0959-6526(20)32623-8

DOI: https://doi.org/10.1016/j.jclepro.2020.122576

Reference: JCLP 122576

To appear in: *Journal of Cleaner Production*

Received Date: 30 November 2019

Revised Date: 12 May 2020

Accepted Date: 27 May 2020

Please cite this article as: Rezaie-Balf M, Attar NF, Mohammadzadeh A, Murti MA, Ahmed AN, Fai CM, Nabipour N, Alaghmand S, El-Shafie A, Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach, *Journal of Cleaner Production*, https://doi.org/10.1016/j.jclepro.2020.122576.

Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach

Mohammad Rezaie-Balf[1], Nasrin Fathollahzadeh Attar[2], Ardashir Mohammadzadeh[3], Muhammad Ary Murti[4], Ali Najah Ahmed[5], Chow Ming Fai[6], Narjes Nabipour[7], Sina Alaghmand[8], Ahmed El-Shafie[9,10]

[1]*Department of Water Engineering, Graduate University of Advanced Technology, Kerman, Iran. Email: moe.rezaie69@gmail.com*
[2]*Water Engineering Department, Urmia University, Urmia, Iran. Email: n.fatolahzadeh@urmia.ac.ir*
[3]*Department of electrical engineering, Faculty of Engineering University of Bonab, Bonab, Iran. Email:a.mzadeh@bonabu.ac.ir*
[4]*Research Center for IoT, Telkom University, Bandung 40257, Indonesia. Email: Arymurti@telkomuniversity.ac.id*
[5]*Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Selangor, Malaysia. Email: Mahfoodh@uniten.edu.my*
[6]*Institute of Sustainable Energy (ISE), Universiti Tenaga Nasional, Kajang 43000, Selangor, Malaysia. Email:* Chowmf@uniten.edu.my
[7]*Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam. Email: narjesnabipour@duytan.edu.vn.*
[8]*Department of Civil Engineering, Monash University, 23 College Walk, Clayton, VIC 3800, Australia. Email: Sina.Alaghmand@monash.edu*
[9]*Department of Civil Engineering, University of Malaya, Kuala Lumpur, 50603, Malaysia. Email: elshafie@um.edu.my*
[10]*National Water Center, United Arab Emirates University, Al Ain, United Arab Emirates*

Corresponding Author: Mohammad Rezaie-Balf

Email: moe.rezaie69@gmail.com

Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach

Mohammad Rezaie-Balf[1], Nasrin Fathollahzadeh Attar[2], Ardashir Mohammadzadeh[3], Muhammad Ary Murti[4], Ali Najah Ahmed[5], Chow Ming Fai[6], Narjes Nabipour[7], Sina Alaghmand[8], Ahmed El-Shafie[9,10]

[1]*Department of Water Engineering, Graduate University of Advanced Technology, Kerman, Iran. Email: moe.rezaie69@gmail.com*
[2]*Water Engineering Department, Urmia University, Urmia, Iran.*
Email: n.fatolahzadeh@urmia.ac.ir
[3]*Department of electrical engineering, Faculty of Engineering University of Bonab, Bonab, Iran.*
Email:a.mzadeh@bonabu.ac.ir
[4]*Research Center for IoT, Telkom University, Bandung 40257, Indonesia.*
Email: Arymurti@telkomuniversity.ac.id
[5]*Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Selangor, Malaysia. Email: Mahfoodh@uniten.edu.my*
[6]*Institute of Sustainable Energy (ISE), Universiti Tenaga Nasional, Kajang 43000, Selangor, Malaysia. Email:* Chowmf@uniten.edu.my
[7]*Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam. Email: narjesnabipour@duytan.edu.vn.*
[8]*Department of Civil Engineering, Monash University, 23 College Walk, Clayton, VIC 3800, Australia. Email: Sina.Alaghmand@monash.edu*
[9]*Department of Civil Engineering, University of Malaya, Kuala Lumpur, 50603, Malaysia.*
Email: elshafie@um.edu.my
[10]*National Water Center, United Arab Emirates University, Al Ain, United Arab Emirates*

Corresponding Author: Mohammad Rezaie-Balf

Email: moe.rezaie69@gmail.com

Abstract

Water quality has a crucial impact on human health; therefore, water quality index modeling is one of the challenging issues in the water sector. The accurate prediction of water quality index is an essential requisite for water quality management, human health, public consumption, and domestic uses. A comprehensive review as an initial attempt is conducted on existing solutions through data-driven models. In addition, the ensemble Kalman filter is found to be a suitable data assimilation method, which is successfully applied in hydrological variables modeling and other complexes, nonlinear, and chaotic problems. In this study, a new application of ensemble Kalman filter-artificial neural network is proposed to predict water quality index using physicochemical parameters for two commonly pollutant rivers, namely Klang and Langat, in Malaysia. As a further attempt, in order to improve the models' performance, a new preprocessing technique is adopted as the newly constructed assimilated model. The results confirm that ensemble hybrid based intrinsic time-scale decomposition has reduced root mean square error by 24 % for Klang and 34 % for Langat, respectively, compared with the intrinsic time-scale decomposition-conventional neural network model. Overall, the developed assimilated methodology shows the robustness of the proposed ensemble hybrid model in analyzing water quality index over monthly horizons that experts could evaluate the water quality of rivers more efficiently.

Keywords: Physicochemical Parameters, Water Quality Index, Data Assimilation, Ensemble Kalman Filter, Intrinsic Time-scale Decomposition.

Nomenclature

ALK= Alkalinity
AN= Ammoniacal-Nitrate
ANFIS=Adaptive Neuro-Fuzzy Inference System
ANN= Artificial Neural Network
ANOVA= One-Way Analysis of Variance
As=Arsenic
Atr= Atrazine
BOD= Biological Oxygen Demand
BTEX= Benzene–Toluene–Ethylbenzene–Xylenes
C=Coliform
Ca= Calcium
CA= Cluster Analysis
Cd= Cadmium
COD= Chemical Oxygen Demand
Cl= Chlorine
Cr= Chromium
Cu= Copper
DA= Data Assimilation
DO= Dissolved Oxygen
DoE= Department of Environment
DDMs= Data-Driven Models
DS= Dissolved Solids
DT=Decision Tree
EC= Electrical Conductivity
EnKF= Ensemble Kalman Filter
F= Fluorides
FC=Faecal Coliforms
Fe= Iron
FS=Fourier Series
FST= Faecal Streptococcus
GA=Genetic Algorithm
GD= Gradient Descent
HCA=Hierarchical Cluster Analysis
HCBD= HexaChlorButaDiene
Hg= Mercury
ITD= intrinsic time-scale decomposition
K=Potassium
KNN= K-Nearest Neighbor
LS-SVM=Least Square-Support Vector Machine
MAE= Mean Absolute Error
Mg= Magnesium
MLR=Multiple Linear Regression
MNNs=Multiple Neural Networks
MSA= Multivariate Statistical Analyses
Na= Natrium
NB = Naive Bayes
NH3=Ammonia
NH3-N= Ammoniacal Nitrogen
NH4= Ammonia
NH4-N=Ammonia-Nitrogen
Ni= Nickel
NO2= Nitrite
NO2-N=Nitrite Nitrogen
NO3=Nitrate

NO3-N= Nitrate Nitrogen	NSE= Nash-Sutcliffe Efficiency
NTU= Turbidity	OG= Oil and Grease
PAH= Polycyclic Aromatic Hydrocarbons	Pb= Plumbum
pH= Potential Hydrogen	PO4=Phosphate
PO4-P= Phosphate Phosphorous	PRC= Proper Rotation Components
PSO =Particle Swarm Optimization	RBC=Rule-Based Classifier
RBFN= Radial Basis Function Network	RMSE= Root Mean Square Error
RSD= Ratio of RMSE to Standard Deviation	Sa= Salmonellas
Sim= Simazine	SMLR= Stepwise Multiple Linear Regressions
SO4= Sulphates	SS= Suspended Solid
SVR= Support Vector Regression	T= Temperature
TA- CaCO3= Total Alkalinity of Calcium Carbonate	TC= Total Coliforms
TCB= TriChloroBenzenes	TDS= Total Dissolved Solid
TH=Total Hardness	TH- CaCO3=Total Hardness of Calcium Carbonate
TP= Total Phosphorus	TOC= Total Organic Carbon
TS= Total Solids	TSS= Total Suspended Solids
Twater= Water Temperature	U95= Uncertainty at 95 %
WQI= Water Quality Index	Zn= Zinc

1. Introduction

Water is the crucial natural element for human survival and social development as well as the ecological (natural, biological, environmental) health (Li et al., 2009). Water is the fundamental element for industrial, agriculture, and biotransformation purposes regardless of drinking and personal hygiene. In the last few decades, water pollution has turned into a severe problem worldwide, particularly in developing countries. Water quality evaluation is, therefore, an essential issue since it directly influences people's lives, and requires further attention from decision-makers (Zhang and Li, 2019). For this purpose, the main characteristics of water, namely biological, physical, chemical, and radiological, are considered as the water quality (Liou

et al., 2004). This is the extent of the condition of water regarding the prerequisites of in any biotic animals and also to any human need. Low quality of surface water that is calculated by various standards such as the health of ecosystems, the safety of human, and drinking water is a crucial subject in the developing world, according to which threatens ecosystems and plants/animals life and human health (Sarkar et al., 2007). Rivers are the most accessible water resources and has been the primary water supply to human civilizations throughout history (Mohammadpour et al., 2016). Rivers among various sources of water supply have been utilized more frequently for human societies' development due to easy access (Ishikawa et al., 2019). The reason for utilizing rivers instead of other water resources like groundwater and seawater is that they might have some problems such as land subsidence (Motagh et al., 2017) and pollution transmission (El-Kowrany et al., 2016), respectively.

Many years ago, the Department of Environment (DoE) suggested the reception of WQI to evaluate and rank the degree of waterways contamination. From that point, the DoE recommended a methodology called (OP-WQI) which stands for Opinion Poll WQI for ascertaining the rank the level of water river of nearby waterways. The strategy that utilized for figuring the WQI in Malaysia includes extensive estimations, changes, devouring time, and exertion (Hameed et al., 2017). In this manner, suggesting an alternative approach, which is immediate and faster with high exactness of computing the WQI, is required. The advantage of water quality index modeling is to provide better management of rivers (Gurjar and Tare, 2019). For decades, precise prediction models of water quality parameters established by experts like (Ishikawa et al., 2019).

Artificial Neural Networks (ANNs) is one of the outstanding DDMs which have been successfully applied to address many prediction issues associated with the environment and

water resources such as stormwater prediction (Gaafar et al., 2019), wastewater modeling (Bagheri et al., 2015), heavy metal prediction (Nath et al., 2018), sediment transport modeling (Moeeni and Bonakdari, 2017), streamflow forecasting (Attar et al., 2020), water level forecasting (Nayak et al., 2006). Although ANN models increase the capacity of model functions by training the data sets, it has some disadvantages, including difficulties in assessing the proper network structure and finding the local optimum, slow convergence rate, and long training time (Chau, 2006). All prediction and measurement approaches have some errors related to them as models do not appropriately simulate the whole behavior of the real system (Attar et al., 2018).

Data Assimilation (DA) can be a useful technique for the generation of an accurate state estimation by fusing the data from these sources (Rezaie-Balf et al., 2019b). Predictive model parameters can be adjusted automatically through DA that is based on mathematic conceptions (Kashif Gill et al., 2007). The essential of DA is to evaluate errors in the model along with the observation data and to update model states by combining the model with observations (Abbaszadeh et al., 2017; Moradkhani et al., 2005).

Researchers have proposed various strategies for reducing input/output variables to overcome non-stationary time series in hydrological parameters (Zhang et al., 2018) These strategies are known as the pre-processing procedures for improving the original data to noise ratio (Rezaie-Balf et al., 2019b). Also, the time series variables can be changed into reasonable structures for further estimation (Dong et al., 2019). Intrinsic Time-scale Decomposition (ITD) is one of the time-frequency-energy analysis, which is utilized in this investigation to arrange multicomponent variables into a few Proper Rotation Components (PRCs) and change non-stationary signals into stationary ones (Martis et al., 2013). In other words, the nonparametric decomposition technique

has been influential for the dataset that inherently is nonstationary and nonlinear with minimal assumptions about data (Yu et al., 2017).

This study aims to provide an overview of available DDMs for WQI prediction. Several predictive models based on soft computing applications have been reviewed here in order to assess the literature. The core objective of the present research is to develop a new and accurate hybrid model for predicting WQI using physicochemical parameters in Klang and Langat Rivers, the two case studies in Malaysia. To the knowledge of the authors, there is no published study related to the application of the ANN learning machine and the Ensemble Kalman Filter (EnKF). The main contribution of the study is to address the erroneous noise reduction for both remarkable improvements in data quality, and prediction accuracy seems to have blurred the hydrology community on the effectiveness of reduction in nonlinear noise in WQI predicting. So then, ITD is firstly used in the present study to surmount the non-stationarity issues applying to decompose the original time series dataset regarding water quality parameters into several sub-sequences. Different models, therefore, are built for each sub-sequences according to its intrinsic features. Another purpose of this study is to estimate the robustness of the hybrid ITD-EnKF-ANN *vs.* other hybrid models such as GD-ANN, EnKF-ANN, and ITD-EnKF-ANN *viz* analytical calculation of performance with graphical plots and numerical metrics of modeled and observed WQI data.

2. Literature review

WQI is a number that illustrates the sum of water quality parameters as a particular number and is useful for managers and decision-makers to assess the water quality in any specific site (Mijares et al., 2019). WQI is introduced in Germany in 1848 (Tasneem Abbasi, Shahid A., 2012), and Horton proposed the first WQI in 1965 (Robert K, 1965).WQI has ranges by its index

numbers, which shows how the water is clean, and it can be classified as excellent quality, good quality, poor quality, very poor, and unsuitable for drinking (Khalid et al., 2018). In general, water quality indexes are divided into six categories as follows: river WQI, drinking WQI, Groundwater WQI, sanitation WQI, irrigation WQI, and WQI in the wetland (Babaei et al., 2011). Table 1 provides a list of relevant studies on the application of DDMs in river WQI prediction. Also, the participant of physicochemical parameters on the prediction of WQI extracting from literature review between 2000 and 2019 are illustrated pH and DO with the 95.83 and 91.67, respectively, were the most influential parameters researchers considered for the studies (Figure 1).

Table 1. Application of DDM using WQI prediction-literature review from 2000 to 2019.

Authors	Year	Model	Time scale	Input Variables	Study Area	Journal
Khuan et. al (Khuan et al., 2002)	2002	ANNs	Annually	DO, BOD, COD, pH, AN, SS	Pahang and Selangor Rivers in Malaysia	Student Conference on Research and Development(IEEE)
Juahir et. al (Juahir et al., 2004)	2004	ANNs	Annually	DO, BOD, SS, AN, COD, pH	Langat River in Malaysia	Journal Kejuruteraan Awam
Ocampo-Duque et. al (Ocampo-Duque et al., 2006)	2006	ANFIS	Monthly	DO, pH, EC SS, BOD, TOC, TC, FC, Sa, FST, PO4, NO3, NH4, SO4, Cl, F, Atr, BTEX, Ni, Sim, TCB, Cr, HCBD, PAH, As, Pb, Hg	Ebro River in Spain	Environment International
Gazzaz et. al (Gazzaz et al., 2012)	2012	ANN	Monthly	T, EC, DS, pH, NTU, SS TS, NH3-N, DO, BOD, COD, Na, K, Ca, Mg, NO3-N, Cl, PO4-P, As, Zn, Fe, TC, C	Kinta River in Malaysia	Marine Pollution Bulletin
Amornsamankul et. al (Amornsamankul et al., n.d.)	2012	FS, GA	Monthly	pH, DO, TS, FC, BOD, SS, TP, T_{air}, T_{water}	Thailand	14th international conference on Automatic Control
Sinha et. al (Sinha and	2013	CA, ANNs	Monthly	pH, DO, FC, BOD, TC	The Hooghly River Basin of	Desalination and Water Treatment

(Saha), 2014)					West Bengal in India	
Mohammad pour et. al (Mohammadpour et al., 2016)	2015	SVM	Weekly	T, pH, DO, EC, SS, NO2, NO3, AN, BOD, COD, PO3	Wetland in the Universiti Sains in Malaysia	Environmental Science and Pollution Research
Sahoo et. al (Sahoo et al., 2015)	2015	ANFIS, PCA	Monsoon season	pH, DO, BOD, EC, NO3-N, TC, FC, COD, NH4-N, TA-CaCO3 TH-CaCO3	Brahmani River in India	Aquatic Procedia
Than et. al (Nguyen Hien Than et al., 2016)	2016	ANNs	Annually	T, Sunshine, Rainfall, Humidity, T_{water}, pH, DO, NTU, C, EC	The Dong Nai River in Vietnam	Journal of Environmental Science and Engineering
Babbar et. al (Babbar and Babbar, 2017)	2017	NB, DT, KNN, SVM, ANN, RBC	June 1995–1997	NTU, pH, DO, BOD, TDS, TH, Cl, NO3, SO4, TC	Yamuna River Basin in India	Environmental Earth Sciences
Ahmad et. al (Ahmad et al., 2017)	2017	MNNs	Weekly	DO, SS, pH, NH3-N, T, EC, NTU, DS, TS, NO3, Cl, PO4, As, Zn, Ca, Fe, K, Mg, Na, OG, E-Coli, C, Cd, Cr, Pb	Perak River Basin in Malaysia	International Journal of River Basin Management
Hameed et. al (Hameed et al., 2017)	2017	ANNs	Monthly	DO, BOD, COD, NH3-N, SS, pH	Langat River and Klang River in Peninsular Malaysia	Neural Computing and Applications
Pham et. al (Pham et al., 2017)	2017	HCA, CA, ANOVA	Seasonally	DO, BOD, COD, NH4, N, PO4, P, TSS, pH	The Upper Part of Dong Nai River Basin in Vietnam	Journal of Water Sustainability
Al-Musawi et. al (Al-Musawi et al., 2017)	2017	ANNs	Annually	pH, PO4, NO3, Mg, Ca, TH, Na, SO4, Cl, TDS, Alk, EC, Fe, NTU	Tigris River of Baghdad in Iraq	Applied Research Journal
Yaseen et. al (Yaseen et al., 2018)	2018	ANFIS	Monthly	DO, TS, NTU, Ca, BOD, COD, T, pH	Selangor River located in Malaysia	Water Resources Management
Wu et. al (Wu et al., 2018)	2018	SMLR	Seasonally	T, pH, DO, EC, NTU, N, P, NH4-N, NO3, NO3-N, Ca, Mg, Cl, SO4	In Lake Taihu Basin in China	Science of the Total Environment
Wang et. al (Wang, 2018)	2018	SVR, PSO-SVR	October 2016	COD, DO, pH, NTU, EC, TP, TN, NH4-N, NO2-N, NO3-N, Ca, Mg, Cl, SO4,T_{water}	Ebinur Lake in China	Nature, Scientific Reports
Tiwari et. al (Tiwari et al., 2018)	2018	ANFIS	Annually	DO, BOD, TDS, SS, NH3-N, N, TP, FC	River Satluj in India	Advances in Civil Engineering

Yilma et. al (Yilma et al., 2018)	2018	ANNs	Seasonally	TSS, N-NO3, N-NO2, TN, TA, TOC, COD, BOD, DO, T, EC, pH	Little Akaki River in Addis Ababa, Ethiopia	Modeling Earth Systems and Environment
Leong et. al (Leong et al., 2019)	2019	SVM, LS-SVM	Annually	DO, BOD, COD, SS, pH, AN	Perak State in Malaysia	International Journal of River Basin Management
Kumar et. al (Kumar et al., 2019)	2019	MSA	March 2012	pH, T, DO, BOD, COD, TN, NH4, TC, FC	Yamuna River in India	International Journal of River Basin Management
Kadam et. al (Kadam et al., 2019)	2019	ANNs, MLR	Pre and post-monsoon seasons	pH, EC, TDS, TH, Ca, Mg, Na, K, Cl, HCO3, SO4, NO3, PO4	Shivganga River Basin in India	Modeling Earth Systems and Environment
Kükrer et. al (Kükrer and Mutlu, 2019)	2019	MSA	Monthly	pH, T, EC, SS, BOD, TH, TA, Ca, N, NH3, Cu, DO	Saraydüzü Dam Lake in Turkey	Environmental Monitoring and Assessment
Ho et. al (Ho et al., 2019a)	2019	DT	Monthly	NH3-N, BOD, COD, DO, pH, SS	Klang River in Malaysia	Journal of Hydrology

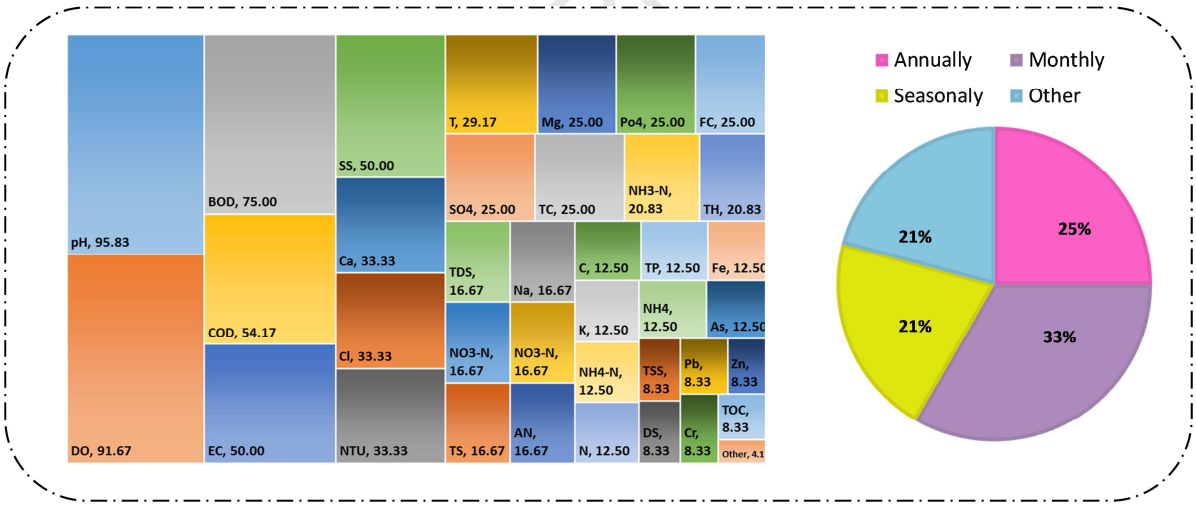


Figure 1. The participation of physicochemical parameters on the prediction of WQI extracted from the literature review between 2000 and 2019.

Chang et al. (Chang et al., 2001) considered three fuzzy synthetic evaluation approaches to model Taiwan river water quality at the Tseng-Wen river system. The results demonstrate that a fuzzy synthetic evaluation method could be useful for daily total maximum load prediction.

Khuan et al. (Khuan et al., 2002) predicted WQI for three years for rivers in Pahang and Selangor in Malaysia by using three algorithms of the ANN, including backpropagation, modular neural network, and radial basis function. Results showed that the RBFN algorithm had higher accuracy than the two other models. Khan et al. (Khan et al., 2003) assumed two different standard indexes namely British Columbia water quality index (BWQI) and Canadian water quality index (CWQI) to estimate WQI in specific watersheds of the region of Atlantic: the Point Wolfe River, the Mersey River, and the Dunk River of Canada. The results of this study asset each standard indexes.

Juahir et al. (Juahir et al., 2004) tested ANN, and multiple linear regression (MLR) approaches for modeling WQI in the site of the Langat River Basin, Malaysia. They showed that Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Ammoniacal-Nitrate (AN), Suspended Solids (SS), and pH were contributed to the estimation of WQI. The results indicate that with omitting two parameters, namely COD and pH as independent variables, the accuracy of the ANN model could be better. Ocampo-Duque et al. (Ocampo-Duque et al., 2006) stated fuzzy inference systems as a model for estimation WQI in Ebro River (Spain). The outcomes of this study have led to proper linking between fuzzy inference systems and parameter weighting approaches. Hore et al. (Hore et al., 2008) utilized the artificial neural network to estimate WQI by accessing waste and polluted water from industrial waste. Two algorithms, namely multilayer-perceptron (MLP) with a back-propagation, were used in this study. As a result, they found that ANN was a convincing method in WQI prediction. Lermontov et al. (Lermontov et al., 2009) used a fuzzy water quality index in order to predict WQI in Ribeira do Iguape River watershed in Brazil. They introduced a new Index as the

fuzzy water quality index (FWQI), and the results of this study showed a good correlation with the traditional calculated index.

Roveda et al. (Roveda et al., 2010) evaluated fuzzy logic in a case of Sorocaba River to model WQI, and they tried to compare the estimated WQI with CETESB WQI (Companhia de Tecnologia de Saneamento Ambiental, in Brazil). They found that it is better to use this estimated method instead of CETESB. Mahapatra et al. (Mahapatra et al., 2011) evaluated the Fuzzy Inference System for estimating WQI in India by utilizing two methods of Sugeno, Takagi, Mamdani, and Kang (TSK) models. The results of this study were compared with three international WQI criteria, and it was found that the cascaded fuzzy system has precious results. Gazzaz et al. (Gazzaz et al., 2012) presented ANN to model WQI for the Kinta River in Malaysia with three categorical variables, including watercolor, water level, and weather, and 32 parameters. The algorithm of ANN called quick propagation training algorithm was defined as the best algorithm to model WQI. Sinha & Saha (Sinha and (Saha), 2014) evaluated the reliability of artificial neural network and cluster analysis modeling in the case of the Hooghly River basin in India for WQI estimation. They tried to compare the results of these methods of DELPHI and CCME, and they found that the DELPHI method has the superior ability in WQI estimation rather than the CCME method.

Hameed et al. (Hameed et al., 2017) investigated artificial intelligence techniques with two different algorithms, namely BPNN and RBFNN, to model WQI in the tropical region in Malaysia. They have used six water quality parameters, including DO, NH3-N, COD, SS, BOD, and pH. Results demonstrated that the RBFNN algorithm performed better than BPNN, which has higher precision. Babbar & Babbar (Babbar and Babbar, 2017) evaluated water quality index applying techniques of data mining as flows: artificial neural networks, naive Bayes,

249 decision trees, k-nearest neighbors, and support vector machines. Parameters that are used to
250 their study were pH, chlorides, DO, BOD, total coliforms total dissolved solids (TDS), sulfates,
251 hardness, nitrates, and turbidity. They detected that the decision tree and support vector machine
252 classifiers are the best models among the other DDMs.

253 Kisi and Yaseen (2019) analyzed alternative hybrid models based on grid partition and
254 subtractive clustering models and adaptive neuro-fuzzy inference system (ANFIS) integrated
255 with fuzzy c-means data clustering, in order to model WQI in Selangor river basin in Selangor
256 by utilizing WQI parameters namely temperature, DO, BOD, turbidity (TU), total suspended
257 solids (TSS), calcium (Ca), COD, and pH. The results demonstrate that in the case of accuracy,
258 ANFIS-SC and ANFISFCM have better results in comparison to the ANFIS-GP model. Ho et al.
259 (2019a) employed decision tree machine learning techniques accompanied by different scenarios
260 (different inputs) for Klang river in Malaysia with six water quality parameters such as NH3-N,
261 DO, BOD, COD, SS, and pH in order to predict WQI. The results indicate that the number of
262 water quality parameters can be diminished as NH3-N, SS, and pH because of a less significant
263 outcome on WQI prediction in a monitoring process. Leong et al. (2019) outlined the use of a
264 support vector machine (SVM) and least-square SVM in WQI modeling. The DoE approach
265 (Malaysia formula to calculate WQI) was used and considered six variables, including DO, SS,
266 BOD, COD, AN, and pH value. As a result, it is found that the LS-SVM model performs better
267 than the SVM model.

268 By reviewing relevant literature, it is observed that most of them used ANN, ANFIS, and SVM
269 to predict WQI without considering uncertainty in models' parameters. That is modest changes
270 to these parameters can significantly alter the model output, making their uncertainty a serious
271 source for predict errors. Therefore, in this study, the feasibility of estimating parameters

simultaneously with the dynamical state is investigated using EnKF by means of state space augmentation. Monitoring and reducing the noise, non-stationary, non-linearity, and complexity of the time series data can be another gap that was not taken into account in previous studies. Prior to using input time series data in the model development process, the frequency components should be resolved to enhance the accuracy of the model. Hence, ITD is a decomposition tool available to address such issues to precisely reconstruct the original time series data and give an appropriate spectral separation of sub-series.

3. Case studies and available data

Adequate water resources are essential for overall economic prosperity in a developing country such as Malaysia (Najah Ahmed et al., 2019). However, some areas in Malaysia are currently experiencing water shortages, even though large amounts of water reserves are available (Naubi et al., 2016). This growing need for water is due to the growth in population, urbanization, industrialization, and irrigated agriculture have dramatically increased the demand for alternative water supplies (Ho et al., 2019b). During the monsoon season, most flood-prone areas experience flooding or flash floods that cause loss of lives, damage to property, and destruction of crops. According to (Ahmed et al., 2019), due to changing weather patterns, this situation will only get worse, and Malaysia has to improve its pre-disaster management systems in order to avoid further damage and other adverse effects caused by floods in the future.

Before focusing on the core of the study (developing water quality prediction model), it is necessary to provide an overview of the climate condition in the selected study area. This is due to the fact that the climate condition could be essential for the model generalization ability for future research. The climate condition for both river basins is the same as both are located in the tropical zone in Malaysia. In general, the Malaysian's climate is affected by several regional and

295 global phenomena such as El Nino, Indian Ocean Dipole (IOD), and monsoons (Suhaila et al.,
296 2010). These phenomena played a vital role in the hydrological formation of the whole country
297 and, more specifically, the extreme events of the rainfall along with the whole year. The annual
298 rainfall is almost 2000 mm, and the highest recorded rainfall was 330 mm that has been
299 experienced in November. On the other hand, the lowest record has occurred in June with almost
300 100 mm (Tangang et al., 2012).

301 Seasonally, two major monsoon regimes influenced the climate in Malaysia, namely; Northeast
302 (NE) and Southwest (SW) monsoon patterns. The SW monsoon season that is dominated by the
303 low-level south-westerly winds begins in May and lasts through August. On the other hand, the
304 NE monsoon season that is controlled by the northeast wind commences in November and ends
305 in February of the following year (Tan et al., 2019). In terms of the temperature pattern in both
306 study areas, the maximum temperature that has been recorded during the last 40 years ranged
307 between 32^{o}C and 35^{o}C, while the minimum temperature was ranged between 21^{o}C and 25^{o}C.
308 From these records, it could be noticed that the narrow changes in the range of temperature,
309 whether the maximum or the minimum ones, showed that the temperature might not play a
310 significant influence on the water quality pattern (Palizdan et al., 2015).

311 Recent research showed that there might not a significant change in the climate condition in
312 Malaysia in the short and medium terms. However, it is expected that there might be a gradual
313 change in the long-term trend changes in the rainfall and temperature patterns. In this context, in
314 the long-term, such significant changes in the seasonal or annual rainfall and the maximum and
315 minimum temperature could drive to changes in the water quality patterns. This is due to the fact
316 that such changes could influence on the flood and drought frequency and hence the availabilities

of the freshwater. Also, it has been observed that the monsoon phenomenon is the most powerful system on the climate condition in Malaysia (Soo et al., 2019).

In the present study, in order to examine the proposed models' performances, two different case studies were chosen, namely the Klang River and the Langat River. In the following subsections, details about the water resources and water quality for both rivers would be explained.

3.1. The Klang River

The Klang River stretches approximately 120 km through the two most populated areas in Malaysia; the State of Selangor and the Wilayah Kuala Lumpur. The river flows from the Ulu Gombak Forest Reserve to Port Klang and on into the Straits of Malacca, one of the busiest shipping lanes in the world. The Klang River basin is the country's most inhabited region, with more than four million residents. This area contains several main cities of the Selangor State and Wilayah Kuala Lumpur, such as Klang, Shah Alam, Puchong, and Petaling Jaya. The biggest seaport in Malaysia, Port Klang, is also situated on the estuary of the Klang River (Juahir et al., 2004). The Klang River's watershed covers approximately 1,288 km^2 of the storage basin. This region has witnessed the country's strongest economic growth, and 35 % of the area has been built up for residential, commercial, industrial and institutional purposes. This region is also considered to be polluted as the extensive developments along the river basin due to the illegal discharge of unprocessed wastewater, as well as treatment plant and animal farming waste, which has deteriorated the water quality of the river. Figure 2 illustrates the location of the river in Malaysia and the location of the water quality monitoring stations.

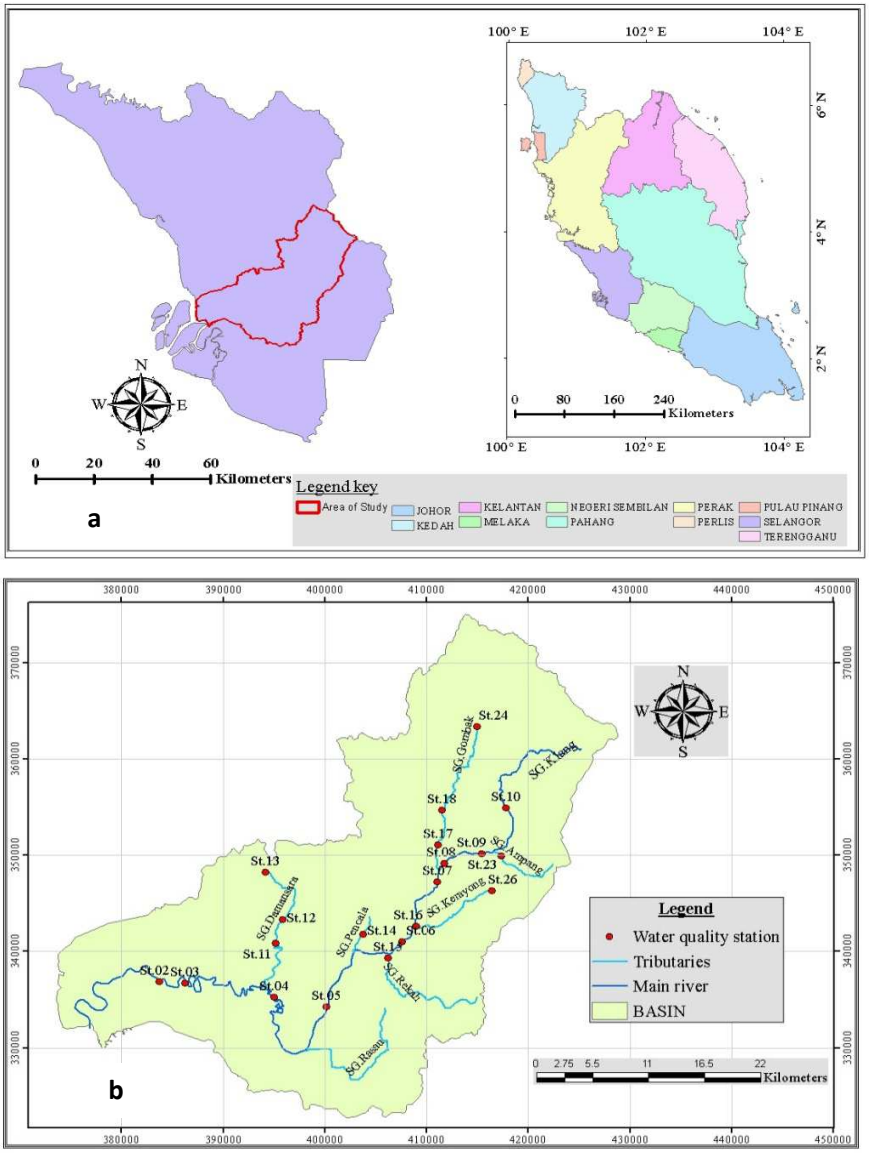


Figure 2. a) Location of Klang River Catchment and b) water quality monitoring stations

The Klang River serves as the primary water supply source for Selangor and Kuala Lumpur, providing nearly 1,128.4 million liters per day (DOE, 2007).

The data selected for the present study were monthly water quality parameter assessment data, summarized as WQI. The six physicochemical water quality parameters used to calculate the WQI was biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved

347 oxygen (DO), suspended solids (SS), pH and ammoniacal nitrogen (NH3-N). The data were
348 collected from monitoring stations situated on the Klang River (Yahya et al., 2019). A total of
349 305 data samples were accumulated for the duration between January 2005 and August 2016 for
350 this study (Palizdan et al., 2015).

351 **3.2. The Langat River**

352 One of Malaysia's most important rivers, the Langat River, is regarded as the primary source for
353 agriculture, consumption, farming, and fishing in the state of Selangor (see Figure 3). The Langat
354 River runs west across the Langat Basin to Kuala Langat from the highest point of the 1,493 m in
355 the Titiwangsa Range. It then flows into the Straits of Malacca. It is 78 km long discharges an
356 area of 2,350 km^2. The Langat River is mainly characterized by water bodies (e.g., natural lakes),
357 forests, agriculture, and urban residential and commercial areas. The types of forests within the
358 catchment area are mangrove, dipterocarp, and swamp. The dominant land-use within the
359 catchment area is for agricultural purposes.

360

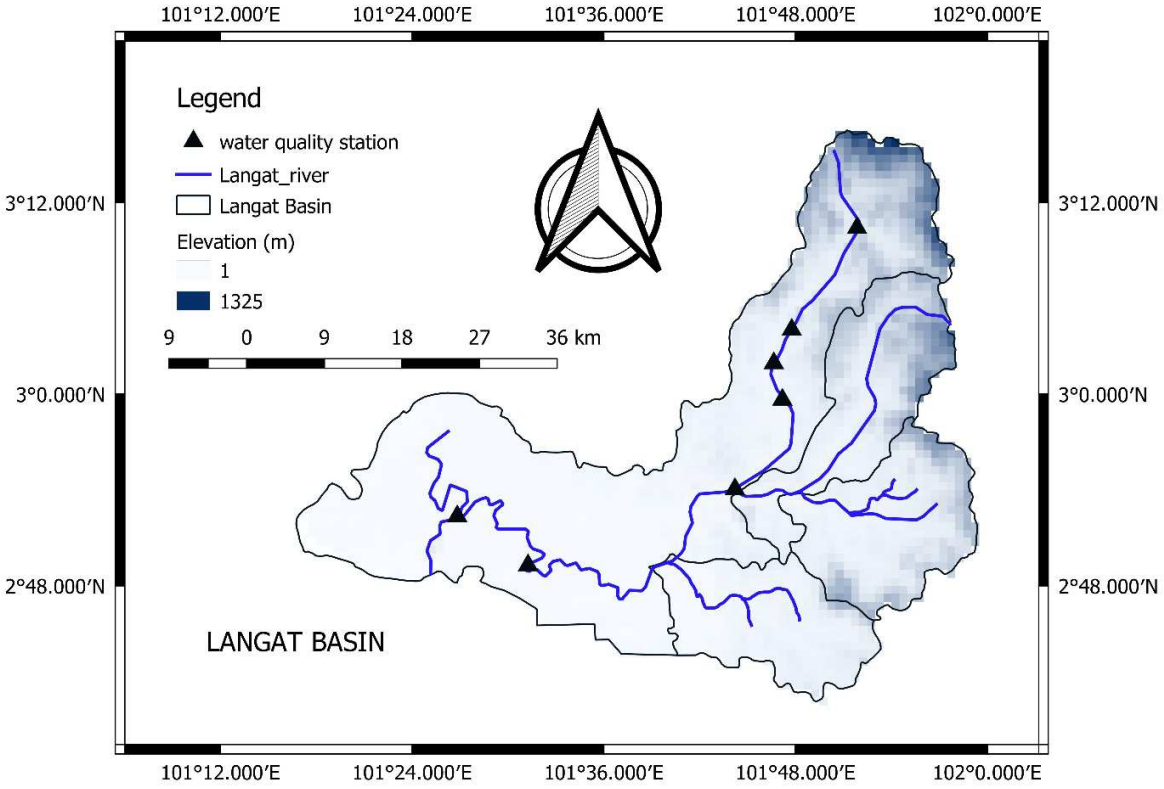


Figure 3. Location of the Langat River Catchment

Land-use practices along the banks of the river have led to the degradation of the water quality of the river (Najah et al., 2011). Research by Yahya et al. (Yahya et al., 2019) determined that the main factors contributing to differences in the quality of the Langat streamflow were the development of the wastewater treatment plants and the industrial waste (chemical effluents), as well as runoff from domestic and commercial areas.

3.3. Available data

In 1978 DoE established baselines to detect the water quality changes in river water quality and has since been extended to identifying pollution sources as well. Water samples are collected at regular intervals from Water samples that have been collected from designated stations for in--situ and laboratory analysis to determine physicochemical and biological characteristics. Water quality monitoring activities were privatized to ASMA (Alam Sekitar Malaysia Sdn Bhd) on 1st

January 1995, both manual & automatic monitoring. In 2005, 1064 manual stations in 146 river basins were monitored (Thorough review of river basins & monitoring stations in 2004). Parameter for in-situ measurement are DO (%), DO (mg/l), Turbidity (NTU), Conductivity (uS/cm), Salinity (ppt), pH, Temperature (T). While the parameter for lab analysis are: BOD, COD, SS, NH3-N, pH, DS, TS, NO3-N, Cl, PO4-P, O&G, MBAS, E.coli, Coliform, As, Hg, Cd, Cr, Pb, Zn, Ca, Fe, K, Mg, Na [24 chemical and biological parameters]. There are three types of monitoring stations that have been used by DoE to identify the water quality parameters. The first type is the baselines stations that allocated in the far upstream position of the river, which is only considered for reference and not for detecting the real water quality status of the river as the river did not affect by the water users. The second type is ambient stations used for monitoring the water quality, and their records are used to configure the change in the water quality parameters. These stations are located along with the whole river to detect the point and non-point sources of pollution. While the third type of the station is the impact station. This type is used for enforcement purposes and not for the calculation of the real water quality status of the river.

The data is available in, owned by DoE, and could be shared for research purposes. The data have been collected from DoE, who operates these monitoring stations for both Klang and Langat rivers, which is institute in charge to monitor the water quality for all rivers in Malaysia. In the present study, DoE (DOE, 2007), Malaysia provided the water quality records for the Langat River. The data of the water quality were recorded irregularly with respect to particular time intervals; therefore, quarterly data were instead used to expedite the study. Consequently, the present study utilized time-series water quality data (ranging from September 2002 to August 2016) at several monitoring stations for the required parameters. Because this research utilized

three-monthly record data, the data reflecting the first quarter were drawn from that quarter's last month, i.e., from March. Likewise, June data reflected the results for the second quarter.

On the other hand, for the last month of a specific quarter, if there were no data available, data were then taken from any of the other months within this quarter. For instance, the data representing Quarter 1 were obtained from either January or February. Similarly, data from April or May were used to reflect the data for Quarter 2. In order to ensure the development of a reliable model, it is required to utilize regular data monitoring. In this context, the proposed model in this study has been developed based on the steady acquired water quality data.

The main reason for the selection of these periods (2005 to 2016 for Klang River and 2012 to 2016 for Langat River) is that the monitoring program for the water quality parameters during these periods was more reliable. In fact, it is essential to develop the model relying on reliable data in order to achieve a successful model structure. In this context, it was decided to utilize the available reliable data during these periods to develop the proposed model.

4. Methods

This section is categorized into six parts including determination of WQI as a national index, an artificial neural network with its formulation, ensemble Kalman filter as a data assimilation technique, intrinsic time-scale decomposition method, description of ITD-based WQI prediction models and in the last part, the performance of the models was assessed.

4.1. Determination of WQI

As defined by the US Foundation of National Sanitation, WQI varies from 0 to 100, where high water quality results in the high value of WQI, and lower values of WQI represent the low quality of water (Said et al., 2004). In 1974, the DoE Malaysia endorsed an index to evaluate the surface water quality in Malaysia. Thoroughly, six parameters were chosen as chief water quality

variables to develop and calculate WQI, such as DO, COD, SS, NH3-N, BOD, and pH for surface water (Khan et al., 2003). These variables should be transformed into a non-dimensional parameter that the relationship for each parameter can be seen from (Gazzaz et al., 2012), which has the best-fit relations of parameters. WQI can be obtained considering the following equation (Khuan et al., 2002):

$$WQI = 0.22SI_{DO} + 0.19SI_{BOD} + 0.16SI_{COD} + 0.15SI_{\mathrm{NH3-N}} + 0.16SI_{SS} + 0.12SI_{pH} \quad (1)$$

4.2. Artificial Neural Networks

Artificial Neural Networks (ANNs) is one of the most fruitful and brain neurological based black-box methods in modeling environmental issues, specifically in water quality modeling (Gazzaz et al., 2015). The central aspect of ANNs is the estimation of nonlinear models with input data and resulted in the output data while using specific functions and algorithms by learning from an example (Maier et al., 2010). This model typically depends on the architecture of networks, the hidden layers, and nodes. ANN models can be categorized into varied categories according to which gradient descent model (GD) is one of the learning approaches. One of the most critical mathematical optimization methods in GD is backpropagation that is applied for learning the connection weights of algorithms in ANN models. The gradient descent model usually attempts to minimize the root-mean-square error (RMSE) by utilizing the backpropagation algorithm. As aforementioned, the ANN model influenced by the architecture of neural networks. Multi-layer perceptron is one of the most usual and popular feedforward types of ANNs architecture and functions for modeling in nonlinear occurrences (Rezaeian-Zadeh et al., 2012). Single neurons titled perceptron is the basis of this network. This architecture involves three layers, including input, hidden, and output layers. The hidden and

output layers contain a specific amount of neurons, but the input layer will vary by the data dimensions. The weights (w) connecting layers are defined by training algorithms that utilize the BP algorithm. Bearing above in mind, algorithms use the Levenberg-Marquardt algorithm for their function approximations. Further information about this typical neural network architecture could be found (Taud and Mas, 2018).

4.3. Ensemble Kalman filter

Data assimilation is the technique which is dealt with errors in model parameters, uncertainties in the models, boundary conditions errors and etc. One of the prominent data assimilation structures is the Kalman filter, which was proposed by (Kalman, 1960) for linear systems. Ensemble Kalman filter (EnKF) was first introduced by (Evensen, 1994) as an extended Kalman filter. The application of EnKF is based on an ensemble of simulations, which can represent the distribution of the system (Johns and Mandel, 2008). Unlike the Kalman filter, EnKF can suit for nonlinear models (e.g., hydrological models). The background of this model is based on the approach of Monte Carlo, in which probability density is the representation of the state (Clark et al., 2008). In the state of ensemble data assimilation generation, two errors should be worth considering, namely internal and external error. Consider the system of a stochastic, nonlinear, general model, M and the observations (Kalman, 1960),

$$x^f(t_k) = M[x^f(t_{k-1}), U(t_{k-1})] + \omega(t_{k-1}) \quad (2)$$

$$y^0(t_k) = H(t_k)x^f(t_k) + \vartheta(t_k) \quad (3)$$

where the forecast represented by $x^f(t_k) \in R^n$ of the system state at the time t_k, the system forcing $U(t_{k-1})$, and the one-time step of the model showed with M. It is compulsory to combine the measurement taken from the observation and modeled by Equation (2), to gain an optimum approximation. By using the information given by the system model Equation (3). The forecast state at the time t_k, denoted by $x^f(t_k)$, is the forecast from observation time t_{k-1} to observation time t_k by the following Equation,

$$x^f(t_k) = M[x^a(t_{k-1}), U(t_{k-1})] \quad (4)$$

where $x^a(t_{k-1})$ is the modeled system state. In t_k, an observation $y^0(t_k)$ is existed, and the investigation step restructures the model,

$$x^a(t_k) = x^f(t_k) + K[y^o(t_k) - H(t_k)x^f(t_k)] \quad (5)$$

where,

$$K(t_k) = P^f(t_k)H(t_k)^T[H(t_k)P^f(t_k)H(t_k)^T + R]^{-1} \quad (6)$$

are the minimum variance gain and the covariance matrix of modeled error represented by $P^f(t_k)$. The covariance in the EnKF algorithm can be estimated by randomly generated a finite number of system states. In order to estimate x_0 as an initial value, an ensemble $\varepsilon_i^f, i = 1, \dots, N$ of the uncertainty is stated for randomly generated states. Original model operator causes the ensemble members to are transmitted from a one-time step to another (Karunasingha and Liong, 2018),

$$\varepsilon_i^f(t_k) = M[\varepsilon_i^a(t_{k-1}), U(t_{k-1})] + \omega_i(t_{k-1}) \quad (7)$$

Where, $\omega_i(t_k)$ apprehensions of the noise process. This noise is a supplementary component for uncertain parts of the model to estimate the covariance between observation and modeled WQI

indices. More information and detailed mathematical background of the ensemble Kalman filter could be found at (Maxwell et al., 2018).

4.4. Intrinsic Time-scale Decomposition

Intrinsic Time-scale Decomposition (ITD)is a time-frequency representation presented by (Frei and Osorio, 2007) for complex and non-stationary time series hydrologic datasets. Proper Rotation Components (PRCs) are components in which the datasets are divided into them.

ITD process procedures could be divided into some steps. This method has an operator L, which extracts the baseline signal from the input signal $x(t)$ that resulted in an accurate rotation and lower frequency in residuals (Frei and Osorio, 2007). In which $Lx(t) = Lx(t)$ is the mean of the signal, written as $L(t)$. The proper rotation components (PRCs) are defined as $Hx(t) = (1 - L)\, x(t)$ which is written as $H(t)$. Then decomposed the input signal $x(t)$ as:

$$x(t) = Hx(t) + L(t) = (1 - L)\, x(t) \tag{8}$$

ITD algorithm follows four steps, including:

Step 1; Finding the corresponding occurrence time τ_k and the extreme points of input signal *x(t)*, where k = 0, 1, 2, ⋯. Considering τ0 = 0 as the first signal.

Step 2; Considering the input signal *x(t)* on the interval [0, τ_k + 2] and L*(t)* and *H(t)* as operators over the time interval [0, τ_k] in which the baseline-extracting operator L is defined as linear function on the interval [τ_k, τ_k + 1]. The baseline extraction operator is designed as:

$$Lx(t) = L(t) = L_k + \left(\frac{L_{k+1} - L_k}{x_{k+1} - x_k}\right)(x(t) - x_k), t \in (\tau_k, \tau_{k+1}), \tag{9}$$

and

$$L_{k+1} = \alpha\left[x_k + \frac{(\tau_{k+1} - \tau_k)}{\tau_{k+2} - \tau_k}(x_{k+1} - x_k)\right] + (1 - \alpha)x_{k+1} \tag{10}$$

Where $0 < \alpha < 1$ is a constant value and taken as fixed value of (α = ½). Linearly contraction of original signal built in order to make monotonic *x(t)* between the extrema points, which is necessary for PRCs.

Step3; The following operator, were defined for extracting PRCs:

$$H(t) = Hx(t) = x(t) - L(t) = x(t) - L(t) \quad (11)$$

The main purpose of ITD is to integrate higher signals into several PRCs. As shown in Equation 11, by subtracting the baseline from the input signal, PRCs can be attained. The advantages of ITD can be summarized in three concepts; low computational time, avoiding transient smoothing, solving the smearing in time-scale space, and constant sifting (this process is applied to data iteratively in order to generate optimum PRCs).

Step4; This process of equations 9 and 10 iteratively repeated until the baseline *L(t)* converts to a monotonic function in which the single signal can be divided into PRCs.

$$x(t) = \sum_{i=1}^{p} H^i(t) + L^p(t), \quad (12)$$

where *p* is the number of achieved PRCs.

4.5. Description of ITD-based WQI prediction models

The primary purpose of ITD-based DDMs is to predict the WQI using physicochemical parameters at two different rivers in Malaysia. The schematics of the ITD-EnKF-ANN, which is considered to predict WQI at two stations, is shown in Figure 4. Before starting three main steps of decomposition-based models, a total of physicochemical measurements and WQI over a monthly time-scale should be divided into two separate parts, calibration (a total of 75 % of data) and validation phases (the remaining 25 %), the ideal model is selected independent of the calibration stage.

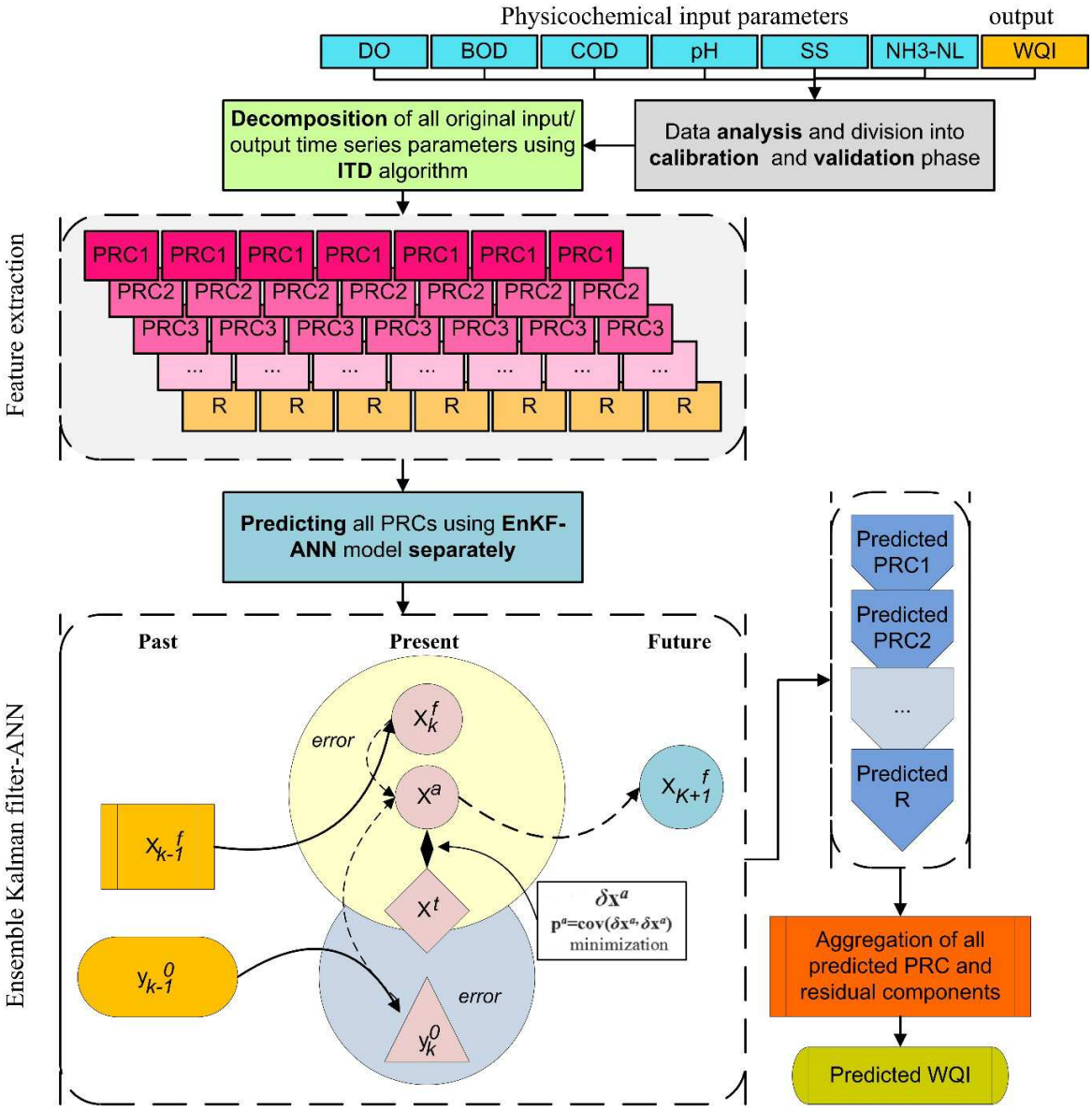


Figure 4. The schematic structure of the proposed hybrid (ITD-EnKF-ANN) model integrating intrinsic time-scale decomposition (ITD) pre-processing approach with an artificial neural network (ANN) model based on the ensemble Kalman filter (EnKF).

The number of the model parameters and the randomness of data is two factors which based on them, the number of data points could be calculated (Fijani et al., 2019). In this research, little

random variations of the data revealed that a reasonable model with a sufficient number of available observations could be estimated.

By considering (Rezaie-Balf et al., 2019a) study, three essential steps to enhance the performance of proposed models can be followed by:

Step 1: ITD procedure is used to break down the input and output datasets into several PRC and one residual components.

Step 2: The GP/EnKF-ANN models are proven as WQI estimation tools to calculate the decomposed PRC and to calculate each component using the same sub-series (PRC1) and the residual component of input variables respectively.

Step 3: The forecasted values of all extracted PRC and residue components using both GP/EnKF-ANN models are combined to generate the WQI.

To summarize, the ITD-based DDMs (i.e., ITD-GP-ANN and ITD-EnKF-ANN) emphasize the "*decomposition and ensemble*" idea. The decomposition is to facilitate the predicting procedure. Whereas, the ensemble is to formulate a consensus estimating on the original datasets. In this work, for verifying and making the pattern of the extracted PRCs and residual components to reflect the estimation technique and improve the prediction process, two rivers in Malaysia (e.g., Klang and Langat) are selected.

It should be mentioned that in step 2, how the ANN hybridized with a Kalman filter to predict each decomposed PRC and residuals components. In the combined approach, the state vectors are provided to the EnKF technique in which the output of the ANN will be considered as state vectors. The output of the ANN will correct by the EnKF to determine the best estimate of the analyzed state or system using the observation data. These states will have resumed all the inputs of the ANN model for the following time step. The inputs of this network have some differences

within them, which are related to feedback form loop or force. For more details, one of the hybridizations of EnKF and ANN the readers can be addressed to (Sharma and Lie, 2012).

4.6. Model's performance metrics

In this study, the newly implemented hybrid ITD-EnKF-ANN vs. ITD-GD-ANN, and standalone EnKF-ANN and GD-ANN models were evaluated by several standard statistical criteria during WQI prediction. Besides common criteria such as RMSE, Nash-Sutcliffe Efficiency (NSE), and Mean Absolute Error (MAE), to assess the fidelity of hybrid proposed models below indices were applied.

1. Kolmogorov-Smirnov distance (K-S distance): It measures the maximum distance D between two consecutive cumulative distribution functions (CDF) (Justel et al., 1997).

$$D_i = \max|F_{i-1}(x) - F_i(x)| \tag{13}$$

2. The ratio of RMSE to Standard Deviation (RSD): RSD metric, was first introduced by (Singh et al., 2005), which is a model evaluation metric to assess the variations between the predicted and observed WQI data. This metric is calculated based on two error metrics, namely, standard deviation (STDEV) and RMSE of the observed WQI data points. The lower value of RSD shows the higher performance of the model.

$$RSD = \frac{RMSE}{STDEV_{obs}} = \frac{\left[\sqrt{\sum_{i=1}^{N}(WQI_{obs} - WQI_{pre})^2}\right]}{\left[\sqrt{\sum_{i=1}^{N}(WQI_{obs} - \overline{WQI_{obs}})}\right]} \tag{14}$$

3. Uncertainty at 95 % (U95): U95 is considered as a 95 % uncertainty confidence of the model.

$$U_{95} = 1.96\sqrt{(STDEV^2 + RMSE^2)} \tag{15}$$

581 4. Reliability of model (%): This statistical metric indicates the satisfactory state of the model's
582 prediction rate by the probability.

$$Reliability = \frac{\sum_{i=1}^{N} K_i}{N} \times 100\ \% \tag{16}$$

$$K_i = \begin{cases} 1, & if (RAE_i \leq \delta) \\ 0, & else \end{cases} \tag{17}$$

$$RAE_i = \frac{\left|WQI_{pre}(i) - WQI_{obs}(i)\right|}{WQI_{obs}(i)} \times 100\ \%, \quad RAE_i \geq 0 \tag{18}$$

583 5. The resilience of model (%): This indicator defines how rapidly the model forecast is likely to
584 recover once an unqualified prediction has followed (Zhou et al., 2017).

$$Resilience = \begin{cases} 100\ \%, & if\ (Reliability = 100\ \% \\ \frac{\sum_{i=1}^{N-1} R_i}{N - \sum_{i=1}^{N} K_i} \times 100\ \%, & else \end{cases} \tag{19}$$

$$R_i = \begin{cases} 1, & if\ (RAE_i > \delta\ and\ RAE_{i+1} \leq \delta \\ 0, & else \end{cases} \tag{20}$$

585 Where $F_j(x)$ and $F_{j-1}(x)$ are the CDF of *i* interval and the previous interval (*i-1*). WQI_{obs} and
586 WQI_{pre} denote the observed and predicted values, respectively; $\overline{WQI_{pre}}$ is an average of
587 observed values, and *N* is the number of the dataset. RAE_i is the ith value for the data, K_i is the
588 number of periods that the threshold value (δ) of the qualified forecast is greater than or equal to
589 RAE value. According to the Chinese standard, the δ is set to 20% (GB/T22482, 2008). R_i is the
590 number of periods in which model prediction is likely to transfer from unqualified into qualified
591 prediction in the ith data (Rezaie-Balf et al., 2019a).

5. Results and discussion

The result section starts with the input data screening, which is finding the correlation between them, along with determining the trend following variance estimation. Then this section continuous with results of stand-alone and hybrid models for both Klang and Langat stations. The discussion part gives information about the comparison between proposed models. This section ends with the current study limitation and suggestion for future works.

5.1. Physiochemical–covariate correlation of source data

The monthly WQI co-variability with the river physiochemical parameters BOD, DO, SS, COD, NH3-NL, and pH are evaluated using the Pearson coefficient, which is known as parametric correlation analysis and primary check, in order to investigate the dependence between multiple variables at the same time. In order to assess the data relationships, the correlation factor was used in which it varies from -1 to +1, where -1 showed a negative correlation, and +1 defines positive correlation. In this study, a graphical correlation matrix is plotted to show a linear dependence between two variables for both Klang and Langat stations (Figure 5). According to this matrix, the monthly WQI has a positive, statistically significant correlation with monthly DO (0.74) and pH (0.28) for Klang and DO (0.82) for the Langat River. Also, a strong negative relationship is attained from the matrix between WQI as the corresponding target and BOD (-0.68) and COD (-0.62) for Klang, and all independent variables except DO for Langat river.

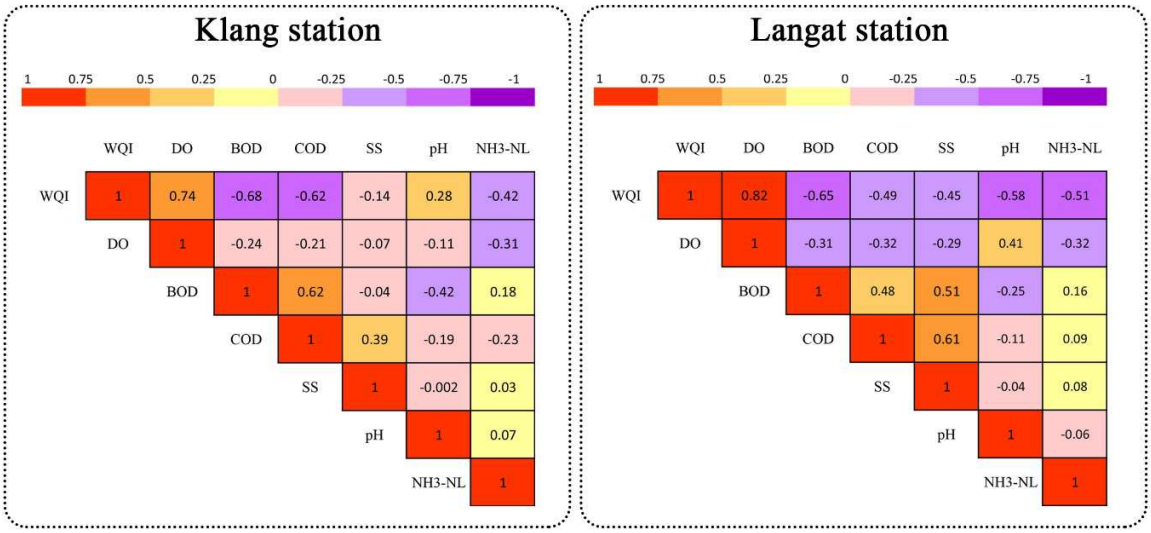


Figure 5. Pearson correlation matrix measures the linear relationship between each physicochemical parameters and corresponding WQI. The color scale indicates the direction of the correlation which means that purple color represents negatively correlated statistics, and orange color positively correlated statistics

5.2. Monotonic trends detection of source data

The monotonic trends in monthly WQI are examined by a standard nonparametric method called Spearman's rank-order correlation coefficient, denoted by ρ, to assess the fact that how two data sets are linked to each other. In other words, the absence of trends is verified by this method in both nonlinear and linear trends (Rezaie-Balf et al., 2019c). The null hypothesis of this analyze the identical distribution and the independence of two variables, and the alternative hypothesis is the existence of decreasing or increasing trends. Considering the position order for identical values, the rank order is assigned. For instance, ρ could take -1 to +1.

As proven in Table 2, correlations between modeled WQI and input (physicochemical parameters of the river) variables are estimated by the Spearman's rank correlation test for two stations. When the P values <0.05, the confidence level for the correlation test is selected. Spearman's rank correlation coefficients for the input variables at two stations of Klang and Langat are more than 0.5, and the confidence level (P-values) is less than 0.05. So, the null

hypothesis in which the two populations are independent is rejected at a level of 5 % of significance, and the modeled WQI is judged to be dependent on input variables.

Table 2. Values of the correlation coefficient between the WQI and the physicochemical input variables

Station	Parameter value	Input variable					
		DO	BOD	COD	SS	pH	NH3-NL
Klang	R	0.73**	-0.59**	-0.78**	-0.64**	0.76**	-0.49**
	Sig	0.00	0.002	0.00	0.001	0.00	0.005
Langat	R	0.87**	-0.79**	-0.86**	-0.73**	0.84**	-0.69**
	Sig	0.00	0.00	0.00	0.00	0.00	0.00

** Marked correlations are significant at P > 0.05 level.

5.3. Statistical analysis of variance of source data

The assessment of the effects of the variables (dependent and independent) is a significant issue in testing the data. Analysis of variance (ANOVA) is one of these tests and allows the modeler to indicate whether independent variables have an influence or the effect of the interaction between these variables on the dependent variable (Lam et al., 2016). The GLM-ANOVA which stands for general linear model analysis of variance is one of the diagnostic tools and critical statistical analysis, which reduce the error variance. In this research, the significance level of 0.05 was utilized in order to recognize the statistical significance of physicochemical variables. These variables, including DO, BOD, COD, SS, pH, and NH3-NL, were selected as the independent variables in this analysis. The GLM-ANOVA was implemented on data for each variable; the results are showed their degree of freedom, the sequential sum of squares, and their contribution percentage of physicochemical properties at each station. The effect of the individual independent variable on WQI (dependent variable) was valued by defining the null hypothesis of equality of independent variances or significance test at probability level (p-values). As shown in

649 Table 3, the significance of physicochemical variables was obtained from the comparison of p-
650 values with the significance level factor (0.05). All of the physicochemical variables were
651 defined as a significant variable because of their p-value <= 0.05. The contribution of each
652 physiochemical data was also shown in these two selected stations. For Klang station, NH3-NL
653 (90.22%) and pH (64.82%), respectively, were defined as the highest and the lowest contributors.
654 However, in Langat station, SS with 95.21 % and NH3-NL with 83.52 % has the highest and
655 lowest contribution, respectively.

656 **Table 3.** Analysis of variance (ANOVA) results for physicochemical variables of river

Station	**Statistical parameters**						
	Source of Variation	DF	Seq. SS	Computed F	P-value	Significance	Co. (%)
Klang	DO	227	67671.22	2.83	0.00	Yes	89.46
	BOD	33	6210.99	13.58	0.00	Yes	82.41
	COD	82	1425.42	6.6	0.00	Yes	70.99
	SS	176	17.86	1.54	0.005	Yes	88.04
	pH	140	2.57	1.41	0.017	Yes	64.82
	NH3-NL	238	93.84	2.52	0.00	Yes	90.22
	Error	13	258.02	-	-	-	-
Langat	DO	104	40865.38	1.4	0.02	Yes	87.95
	BOD	24	3141.46	14.9	0.00	Yes	79.54
	COD	50	153.43	11.26	0.00	Yes	89.5
	SS	103	33.28	2.51	0.03	Yes	95.21
	pH	93	3.672	4.01	0.00	Yes	94.19
	NH3-NL	66	29.72	3.84	0.00	Yes	83.52
	Error	116	44223	-	-	-	-

657 DF: the degree of freedom; Seq. SS: Sequential sum of squares; Co.: Contribution.

658 **5.4. Results for standalone and hybrid models**

659 This section has highlighted the results for standalone and hybrid models in two subsections for
660 Klang and Langat stations in both calibration and validation stages. The initial attempt is to
661 investigate which training algorithm in ANN will be suitable for the given task. Besides gradient
662 descent that is one of the typical training algorithms, the result of another implementation with
663 EnKF assimilation is obtained in order to consider the effect of assimilation for predicting WQI.
664 Afterward, the pre-processing technique, ITD, integrating with the models mentioned above, is

665 applied in order to improve models' accuracy. Diagnostic evaluation of the statistical error
666 metrics, including NSE, RMSE, MAE, RSD, U95, reliability, resiliency, non-parametric
667 Kolmogorov- Smirnov (K-S) distance statistic, and visual plots such as scatter plot, time-series
668 plot, Taylor diagram, and error bar for predicted and measured WQI are employed to assess
669 models' performance.
670

671 **5.4.1. Klang Station**

672 According to performance measures, the prediction of well-designed hybrid model ITD-EnKF-
673 ANN *vs.* ITD-GD-ANN, EnKF-ANN, and GD-ANN models is numerically evaluated in this
674 sub-section. As shown in Table 4, at the calibration stage, in terms of NSE, the accuracy of both
675 GD-ANN and EnKF-ANN integrating the ITD approach is increased from 0.74 to 0.79 and 0.92
676 to 0.935, respectively. By considering RMSE, the combined model errors were decreased by 45
677 % for the ITD-GD-ANN model and 44 % for the ITD-EnKF-ANN model. In the case of MAE,
678 ITD-EnKF-ANN (2.92) performed better than ITD-GD-ANN (3.42), and RSD explains the
679 decrease by 0.238 and 0.206 for hybrid GD-ANN and EnKF-ANN models, respectively.
680 Therefore, it shows the satisfactory results that hybrid models outperformed the standalone
681 models. U95 shows that 95 % of the uncertainty confidence of the models; the results confirmed
682 the decreasing trend in both hybrid models (ITD-GD-ANN=31.906 and ITD-EnKF-
683 ANN=31.73). Further comparison of these models by reliability and resilience percentages
684 showed a notable increase for both models. Focusing on K-S distance between observed and
685 modeled WQI data, the ITD-EnKF-ANN model has the lowest amount of distance (0.068), among
686 other models for the Klang river. This shows the preference of this model in which the observed

687 and modeled values are closer. All the above statistics show the quicker and satisfactory WQI
688 forecast by using ITD-GD-ANN and ITD-EnKF-ANN.

689 **Table 4.** Evaluation metrics of the proposed models in the calibration and validation stages at Klang
690 station

Models	Statistical error indices			
	GD-ANN	EnKF-ANN	ITD-GD-ANN	ITD-EnKF-ANN
Total available data in the calibration stage				
NSE	0.74	0.788	0.92	0.935
RMSE	7.97	7.21	4.31	3.97
MAE	6.23	5.704	3.42	2.92
RSD	0.508	0.459	0.27	0.253
U95	34.51	33.856	31.906	31.73
Reliability (%)	78.29	82.55	95.74	97.87
Resilience (%)	56.66	66.09	90.17	93.42
K-S distance	0.115	0.085	0.072	0.068
Total available data in the validation stage				
NSE	0.51	0.71	0.682	0.81
RMSE	10.05	7.66	8.055	6.17
MAE	8.16	6.51	5.97	5.069
RSD	0.69	0.532	0.559	0.42
U95	34.403	31.95	32.33	30.69
Reliability (%)	73.91	84.05	89.85	94.30
Resilience (%)	77.75	90.91	85.71	91.97
K-S distance	0.319	0.256	0.289	0.217

691
692
693 Regarding results presenting in the validation stage, all the statistical indices marked that ITD
694 based ANN model with the help of EnKF assimilation performed better than sole models. For
695 instance, the NSE increased by 0.172 and 0.1, reliability 15.94 and 10.25 and resiliency 7.96 and
696 1.06 for ITD-GD-ANN and ITD-EnKF-ANN respectively. The decrease in other error indices
697 such as RMSE, MAE also reveals that the coupled models have better results in the prediction of
698 WQI. In accordance with this, MAE has the highest decrease in values for both ITD-GD-ANN
699 (26.83 %) and ITD-EnKF-ANN (22.13 %) models, however, for RMSE values, 19.8 %, and
700 19.45 % deduction were noticed for coupled GD-ANN and EnKF-ANN models respectively. By
701 considering RSD as a mathematical index, by combining the ITD algorithm with GD-ANN, this

index was decreased from 0.69 to 0.559, and by combining the ITD algorithm with EnKF-ANN models, this index decreased from 0.532 to 0.42. This shows that the error diminished, and the prediction could be more accurate. K-S statistic in the validation stage also depicts the lowest distance (0.217) for the ITD-EnKF-ANN method, which is the result of concordance between input and output data of WQI.

The scatter plots between observed and the predicted WQI (Figure 6) reveal that at the calibration stage, the ITD-GD-ANN model (R^2=0.92) relatively superior to standalone GD-ANN models (R^2=0.74). The same results for EnKF-ANN (R^2=0.81) as standalone and ITD-EnKF-ANN models (R^2=0.94) showed that combined models performed better than standalone models. Comparison of model accuracies in the validation stage, also the priority of coupled models (ITD-GD-ANN and ITD-EnKF-ANN), was seen. In another view of scatter plots, the regression equation, which is based on modeled and observed values of the WQI index, is found from $y(W_m)=aW_o+b$.

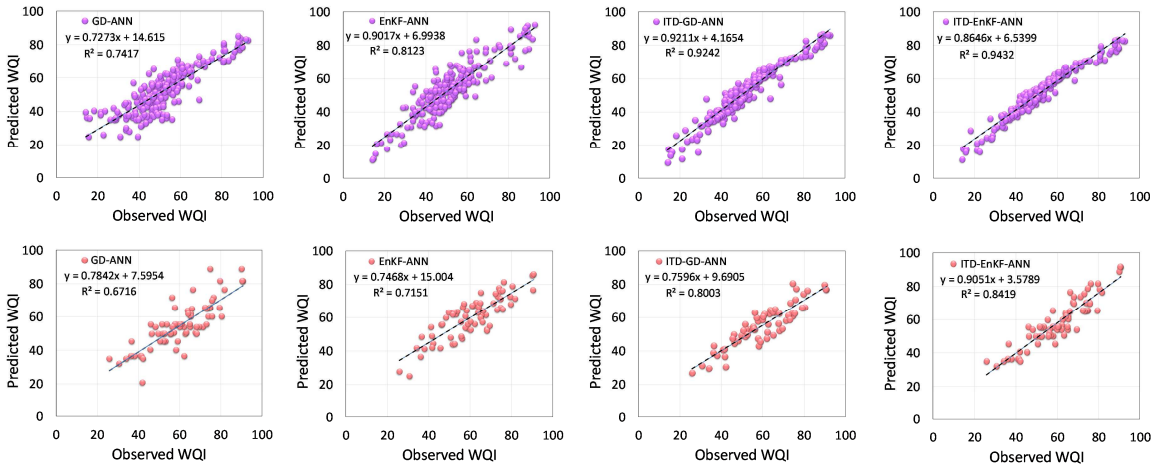
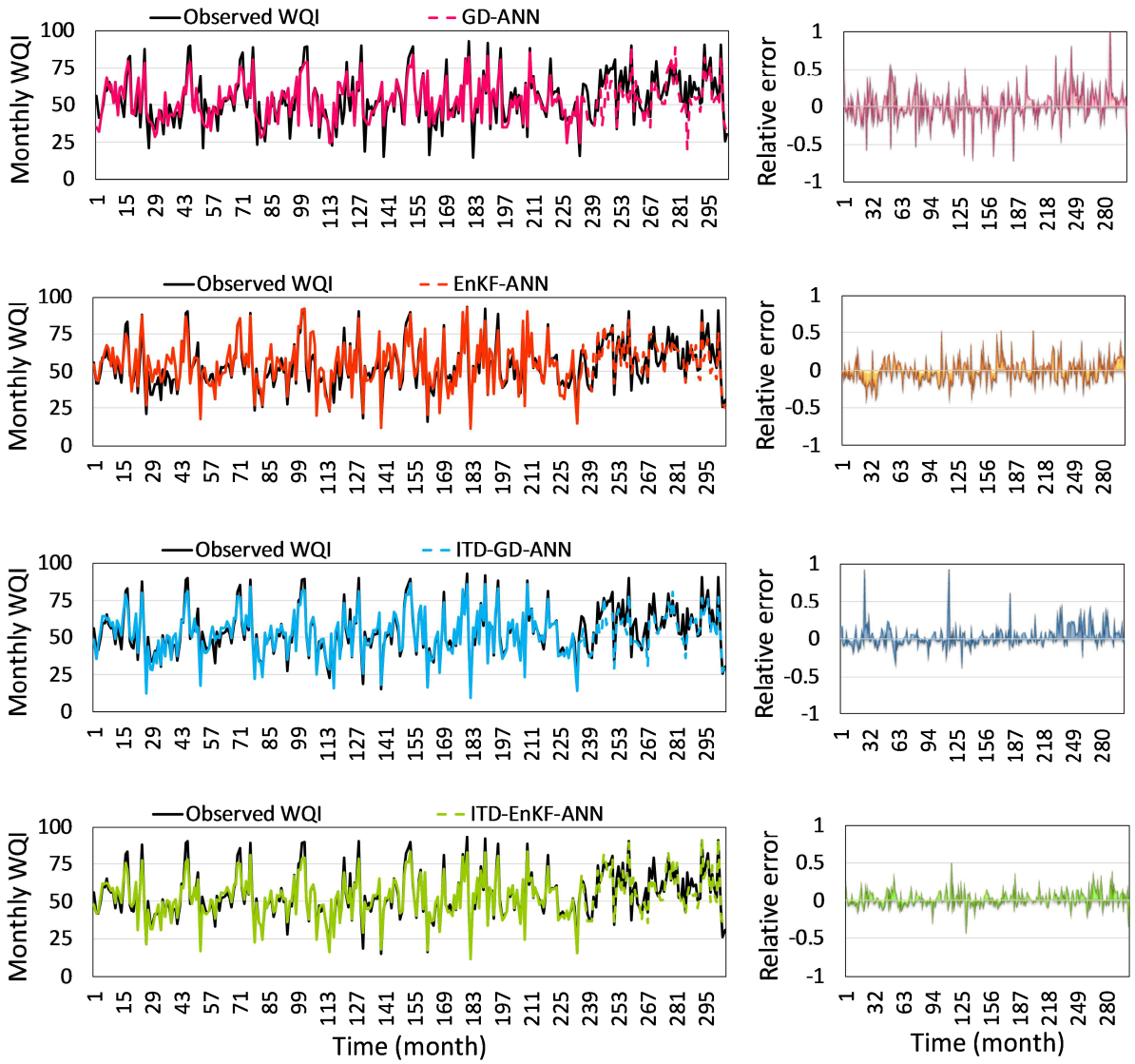


Figure 6. Scatter plots between the observed and the predicted value of WQI for standalone and hybrid models at Klang station in calibration (up) and validation (down) stages for all proposed models.

720

721 In this equation, W_m shows the modeled WQI, and W_o illustrates the observed WQI. The
722 accurate model based on the values of a, b, and R^2 could be selected. In this research at the
723 calibration stage, the ITD-EnKF-ANN with a=0.864, b=6.5399, and R^2=0.9432 was selected as
724 the best model. The same analysis for the validation stage reveals that with a=0.9051, b=3.5789,
725 and R^2=0.8419, the ITD-EnKF-ANN model was selected as the accurate model for predicting
726 WQI in Klang station.
727 Although predicting all quantities of WQI is helpful for various practices, such as drinking,
728 agriculture, and industry, the low and high values of this index are more crucial because of its
729 direct impact on public health and the environment. In this regard, by considering time-series
730 plots along with relative error plots which their x-axis showed their cumulative time (month) and
731 y-axis for predicted monthly WQI high and low values of WQI, which is predicted by selected
732 hybrid ITD-EnKF-ANN along with other models are analyzed (Figure 7).
733 Error criteria (relative error) and graphical analyses were used for evaluating the proposed
734 methods of WQI prediction for standalone (GD-ANN and EnKF-ANN) and integrated (ITD-GD-
735 ANN and ITD-EnKF-ANN) models. In the plots that relative error is calculated, the accuracy of
736 models was analyzed, and it is shown that in the GD-ANN model, the peak values reach to one
737 value, while for ITD-GD-ANN, most of the values are closer to zero. Besides, with EnKF-ANN,
738 the error values fluctuate between 0.5 and -0.5, also in ITD-EnKF-ANN. The values of relative
739 error were close to zero, and the fluctuations are extremely low. Considering time series plots
740 and relative error plots, ITD-EnKF-ANN was found to be more suitable for WQI prediction.
741 However, ANN, with the help of the GD algorithm, had poor accuracy and was not a reliable
742 model.

743



744

745 **Figure 7.** The hydrographs of observed *vs.* predicted monthly WQI using standalone and hybrid
746 models for calibration (solid lime) and validation (dash line) stages and relative error plot for
747 Klang station.

748

749 **5.4.2. Langat Station**

750 Similar to the previous section, Table 5 exhibits the mathematical indexes in the calibration stage
751 and validation stages for proposed models. As shown in the calibration stage, NSE value was

increased from 0.82 to 0.89 in ITD-GD-ANN and 0.87 to 0.92in ITD-EnKF-ANN in comparison with their sole-models. The error values of RMSE, MAE, and RSD were decreased by 22 %, 25 %, 24 % when the GD-ANN model was combined with ITD algorithms. By comparing the uncertainty of models, ITD-EnKF-ANN (U95=41.75) performed better than ITD-GD-ANN (U95=42.38) at 95 % of confidence. In the validation stage, the same indices provide adequate proof that combines ITD-EnKF-ANN outperformed the ITD-GD-ANN models. For example, considering NSE, it is increased by 1.328 for the GD-ANN model also 0.14 for the EnKF-ANN model by a combination of ITD algorithms. By considering the error indices, RMSE, MAE, RSD were decreased by their values in coupled ITD models. In the other aspect, the lowest difference between two consecutive cumulative distribution functions (CDF) of input and output WQI data in the calibration stage for Langat station is 0.106, which is belong for ITD-EnKF-ANN model. This shows that the hybrid ITD-EnKF-ANN model performed better than the other models. The same outcome of the preference of the ITD-EnKF-ANN model with a distance of 0.217 was calculated for the validation stage. These results reveal that the ITD algorithm as a pre-processing algorithm performed better while combining to DDMs.

Table 5. Evaluation metrics of the proposed models in the calibration and validation stages at Langat station

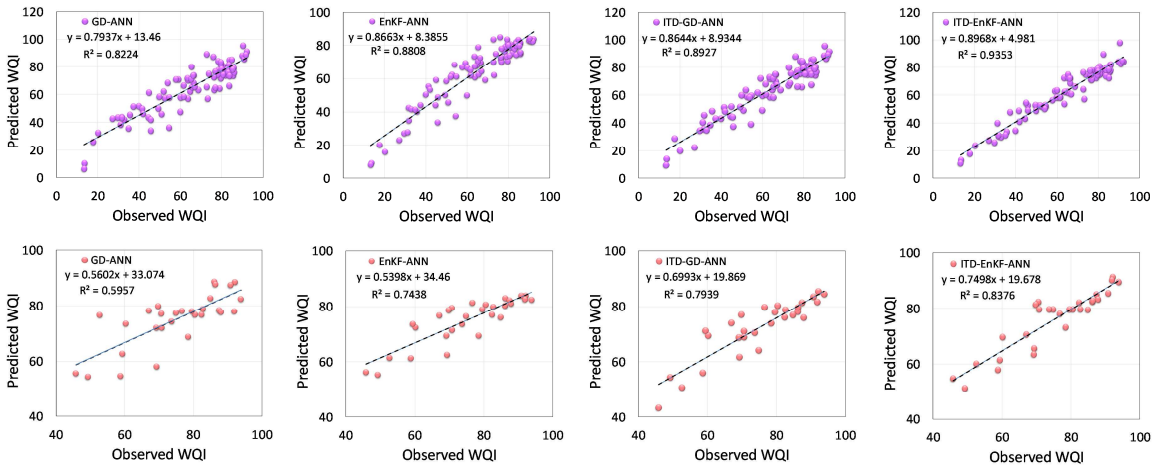
Models	Statistical error indices			
	GD-ANN	EnKF-ANN	ITD-GD-ANN	ITD-EnKF-ANN
Total available data in the calibration stage				
NSE	0.82	0.87	0.89	0.92
RMSE	8.66	7.09	6.74	5.61
MAE	7.21	5.91	5.42	4.59
RSD	0.42	0.345	0.32	0.27
U95	43.71	42.607	42.38	41.75
Reliability (%)	72.73	86.36	89.77	95.45
Resilience (%)	66.49	83.33	77.78	89.56
K-S distance	0.159	0.148	0.125	0.106
Total available data in the validation stage				
NSE	0.598	0.69	0.73	0.83
RMSE	8.28	7.27	6.69	5.403

MAE	6.504	6.16	5.72	4.38
RSD	0.622	0.54	0.503	0.406
U95	30.72	29.72	29.19	28.15
Reliability (%)	89.65	89.65	94.55	97.68
Resilience (%)	67.58	67.58	91.48	94.15
K-S distance	0.310	0.276	0.241	0.217

769

770

771 Figure 8 showed the scatter plot of the proposed models in order to assess the best accuracy for
772 WQI prediction. Considering Figure 8 in detail, at the calibration stage, the correlation
773 coefficient for ITD-GD-ANN was increased by 0.07 in comparison with stand-alone GD-ANN.
774 Also, the coefficient of determination for ITD-EnKF-ANN increased by 0.09 compared with
775 stand-alone EnKF-ANN. In the validation stage, the R^2 for ITD-GD-ANN and ITD-EnKF-ANN
776 were increased by 25 % and 11 % in comparison with their sole models. The outcomes from the
777 stand-alone and combined models reveal that in both calibration and validation stages, the ITD-
778 EnKF-ANN was confirmed the best model in WQI prediction.



779

780 **Figure 8.** Scatter plots between the observed and the predicted value of WQI for standalone and
781 hybrid models at Langat station in calibration (upper row) and validation (lower row) stages for
782 all proposed models.

783

784 Figure 9 depicts the time series of predicted vs. observed values of WQI calibration and
785 validation stages. In standalone GD-ANN and EnKF-ANN models, the maximum value of
786 relative error belongs to 61st month with RE=1 and RE=0.7, respectively. By comparing the two
787 combined models, ITD-GD-ANN and ITD-EnKF-ANN, it is shown that the maximum value for
788 relative error was between 0.5 and -0.5, and the error values are close to zero in ITD-EnKF-ANN
789 model. Thus, this also reaffirms that the ITD-EnKF-ANN hybrid model has better predictive
790 skill than the other combined and standalone models considered in this research.
791 Furthermore, the utilization of such a modeling procedure does not only predict water quality
792 index accurately but also can improve the water quality monitoring programs by reducing the
793 costly experimental testing and time-consuming issues.
794

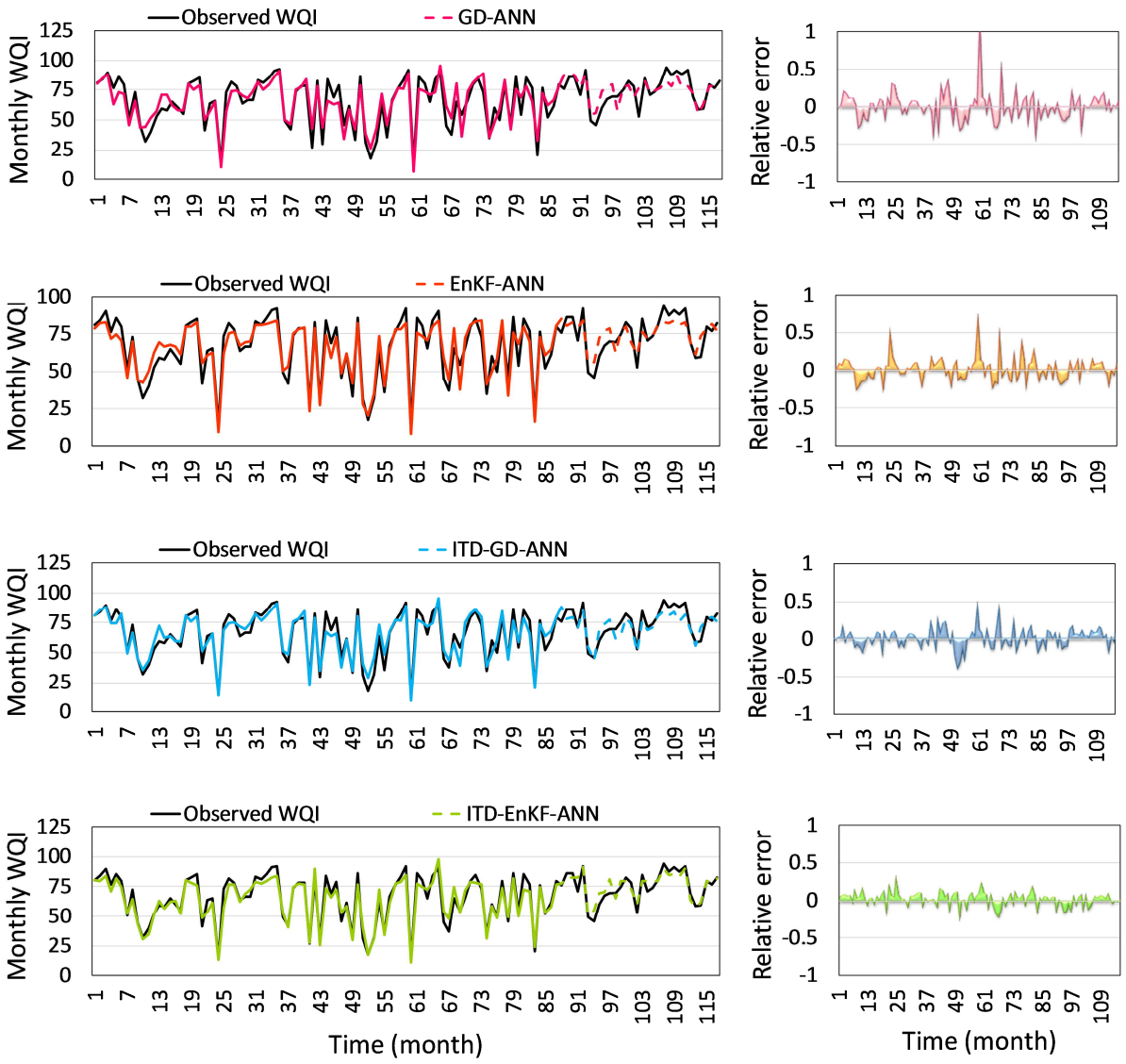


Figure 9. The hydrographs of observed *vs.* predicted monthly WQI using standalone and hybrid models for calibration (solid lime) and validation (dash line) stages and relative error plot for Langat station.

5.5. Further comparison among proposed models

Based on peak values of predicted monthly WQI with the observed extreme values of each station, the best models can be identified. For this aim, Table 6 demonstrates the ten highest extreme values of predicted WQI for two stations considering the GD-ANN, EnKF-ANN, ITD-GD-ANN, and ITD-EnKF-ANN models. As shown in Table 6, the maximum difference between the extreme value belongs to the EnKF-ANN model while the minimum difference goes to ITD-

EnKF-ANN for Klang station. Again for Langat station, the highest value for WQI observation was 93.84, while the peak values for the models were 82.506, 82.436, 84.484, and 89.538 for GD-ANN, EnKF-ANN, ITD-GD-ANN, and ITD-EnKF-ANN models, respectively. This resulted that the ITD-EnKF-ANN model outperformed other models in the view of extreme values.

Table 6: Accuracy evaluation of different models for predicting extreme WQI values (Klang and Langat stations)

Observed value	GD-ANN	EnKF-ANN	ITD-GD-ANN	ITD-EnKF-ANN
		Klang station		
92.86	88.252	82.528	85.916	90.471
91.77	83.628	83.134	86.181	92.699
90.78	86.110	81.664	76.881	91.664
90.15	76.608	81.664	77.14	91.664
90.13	85.3826	88.766	81.499	88.766
90.12	85.139	78.468	81.283	88.431
89.87	77.866	78.892	81.306	78.006
89.42	91.660	79.195	82.079	89.152
89.30	87.959	83.68	86.561	83.098
89.00	87.207	80.529	84.054	86.649
		Langat station		
93.84	82.506	82.436	84.484	89.538
92.2	91.099	83.838	91.389	86.101
92.12	88.549	83.980	85.437	91.274
91.85	77.995	82.551	81.692	90.269
91.71	86.74	82.248	89.072	86.618
90.91	84.721	82.452	86.299	85.442
90.83	84.5	83.960	83.497	86.411
90.37	95.478	84.063	95.396	92.93
89.93	87.006	82.896	88.47	86.262
87.90	77.701	81.121	79.935	83.305

Figure 10 demonstrates the Taylor diagram, which is used to quantify the degree of correspondence between modeled and observed WQI in the tested data in terms of three primary statistics on a single diagram. It shows the RMSE, the correlation coefficient, and the standard deviation for GD-ANN, EnKF-ANN, ITD-GD-ANN, and ITD-EnKF-ANN models for both Klang and Langat stations.

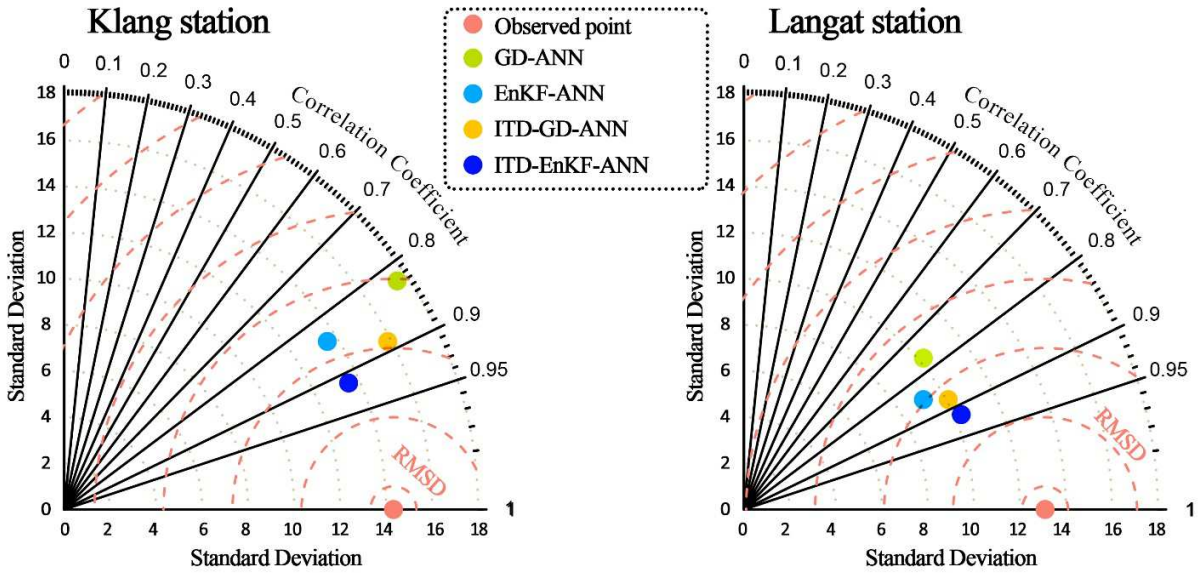


Figure 10. Taylor plots indicating the correlation coefficient and standard deviation in the validation stage based on the standalone models vs. the hybrid-assimilated models for predicting monthly water quality index at two candidate study stations.

Concurring with earlier results, it was evident that the ITD-EnKF-ANN model in both stations is closer to the optimum reference point when a combined visual valuation of the statistics is made. As evident from this diagram, the coupled ITD-EnKF-ANN model has a higher correlation and inversely a lower standard deviation for both stations in the prediction of WQI. However, the GD-ANN model lies much farther to the line representing the centered root-mean-square difference, while the standard deviation of the GD-ANN model remains modestly farther than other models to reference.

The empirical cumulative distribution function (ECDF) was plotted at different predicting abilities (Figure 11), which predicted error of monthly WQI in the x-axis and the percentage of the distribution function in the y-axis for each model. According to the plot, it is evident that the

834 ITD-EnKF-ANN hybrid model was gently better than ITD-GD-ANN for WQI predicting at both
835 stations, and both decomposed-based models were superior to the original models.
836 Based on the percentage of errors in the minimum error bracket (i.e., from 0 to 5) for the Langat
837 station clearly confirms that the ITD-EnKF-ANN was the most responsive model in predicting
838 water quality index (50 %) compared to 44 % for the ITD-EnKF-ANN, 36 % for the EnKF-
839 ANN, and 29 % for the standalone GD-ANN model. Inferior performances were demonstrated
840 when the non-ITD/DA mechanisms were utilized. Therefore, the highest performance with the
841 lowest predicted error resulted from the GD-ANN model. The results of these ECDF plots are
842 consistent with the subject that WQI prediction has a better result when using an ensemble
843 Kalman filter ANN model, which is combined with ITD pre-processing techniques.
844
845
846
847
848

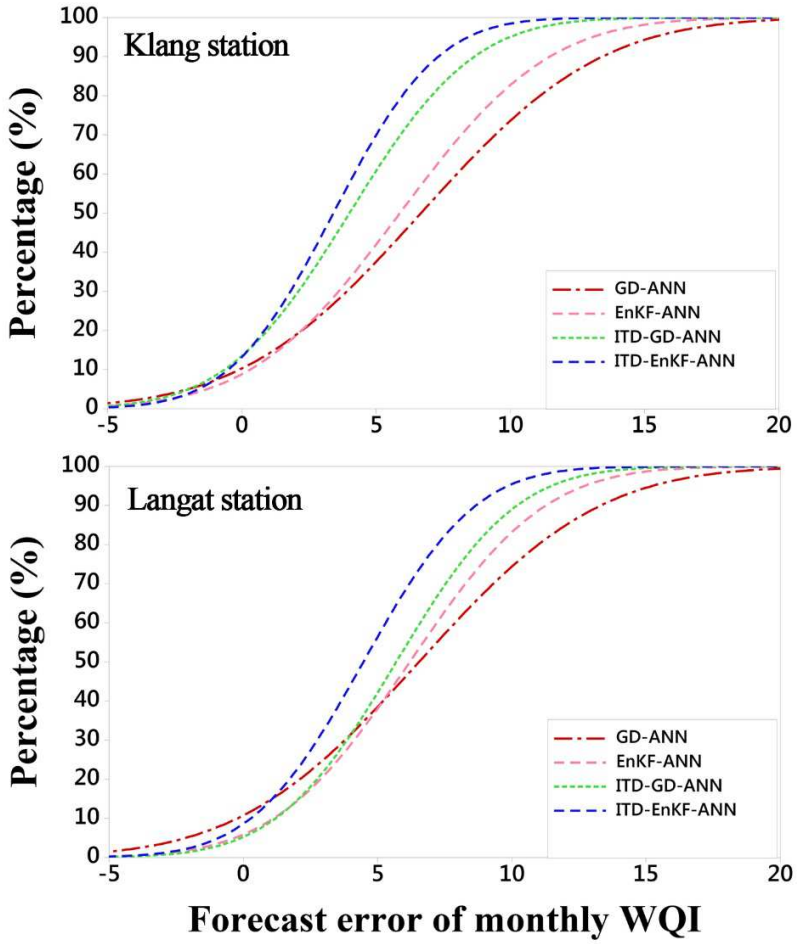


Figure 11. Empirical cumulative distribution (ECDF) of the absolute forecasted error |FE| for the ITD-EnKF-ANN model compared to the other models at Klang and Langat stations in the validation phase.

5.6. Current study limitations

This study consistent with the concept of water quality index modeling by using neural networks modeling, which is used the ITD pre-processing data algorithm for the first time. Despite using a standard training algorithm, namely gradient descent in this research, ensemble Kalman filter algorithm as an assimilation algorithm used in this research in order to eliminate GD algorithm drawbacks for improving the model's accuracy in terms of prediction. For more satisfaction on

the result of WQI prediction, data decomposition technique, ITD, proposed to extract input/output variables into different sub-signals in order to overcome the non-stationarity features in the time series real data.

The present study has shortcomings that create an opportunity for follow-up research in the field of hydrology. Implementation of the ITD pre-processing technique integrated with EnKF-ANN is time-consuming because it produces a large number of PRCs. Follow-up studies can consider another pre-processing method to reduce computational cost or to implement ITD-EnKF-ANN all together in one main source code, at least to reduce the time of development. Another limitation through the study was the lack of meteorological data in some months for both stations, and this drawback may provide uncertainty on the prediction of water quality index. In this regard, it is suggested that future studies might use the satellite-based dataset in order to analyze the data for WQI prediction. In addition, as mentioned above, source data was limited in terms of predicting WQI and was the three-month timescale. Hence, a follow-up study could investigate the model's skill for better temporal resolution (e.g., hourly, daily, weekly, and monthly) with satellite-based prediction.

6. Conclusions

This paper underlined the importance of water quality modeling for human health. In this study, as the first step, a comprehensive literature review was carried out on the current state of river WQI modeling. It was found that pH and DO as the physicochemical parameters with the 95.83 and 91.67, respectively, were the most influential parameters researchers considered for the studies.

Besides the GD algorithm that was initially used for finding the minimum of a function in ANN, the Ensemble Kalan Filter (EnKF) assimilation approach that is one of the best solutions to nonlinear problems, is used to merge ANN model prediction with assimilating production data at two famous polluted rivers in Malaysia, namely Klang and Langat. Considering evaluation metrics, using EnKF to predict WQI could improve the accuracy of the standalone ANN model by 39 % and 17 %, respectively, for Klang and Langat stations in terms of NSE compared with GD training algorithms. In addition, predicting error was reduced to 7.66 and 6.51 in terms of RMSE and MAE, respectively, by augmenting the state space with model parameters (using DA technique) compared to no assimilation at Klang station.

As a further attempt, the performance of a newly constructed ensemble hybrid decomposition model embedded with the Intrinsic Time-scale Decomposition (ITD) as a pre-processing technique integrated with the ANN model was adopted. That is, the physicochemical time series and the corresponding target using the ITD algorithm were extracted (decomposed), resulting in improved performance of the standalone models. In this respect, the RSD and U95 values of the ITD-EnKF-ANN model for WQI estimation were reduced to 25.3 % and 5.2 %, respectively, compared with the EnKF-ANN model at Langat station. Considering the plotted empirical cumulative distribution function (ECDF) at different predicting abilities in both stages of calibration and validation along with non-parametric statistics, namely Kolmogorov-Smirnov (K-S) Distance method in Klang and Langat rivers, the hybrid assimilated ITD-EnKF-ANN performed better than the other models.

Overall, the achieved results indicated that the hybrid assimilated ITD-EnKF-ANN model would be a robust approach to predict WQI on the monthly timescale since the results were favorable for both Malaysian stations. It is also can be proposed as a possible solution in order to reduce

905 the noise in highly nonlinear hydrological phenomena such as the prediction of streamflow, solar
906 radiation, etc.

907 In order to widen the scope of the study, the ITD-EnKF-ANN model could be improved with
908 ensemble-based uncertainty testing via a bootstrapping and the Bayesian model averaging
909 techniques, although the proposed model had a precise prediction. One possibility for future
910 study is to consider other DDMs such as gene expression programming, extreme learning
911 machine, etc. for integrating with ensemble Kalman filter to perform an accurate model in the
912 prediction of the hydrological processes (i.e., streamflow, rainfall, water stage, groundwater,
913 etc.).

914 With the aim of the accuracy of WQI modeling, it is better to consider more data samples and
915 various input variables such as heavy metals, pollutants, and radioactive samples from different
916 rivers in Malaysia. The water quality can also be affected by their background basin, so this can
917 affect the concentration of each quality parameter. Thus, for future works, the authors suggested
918 assessing basin effects too. Finally, it can be suggested as a potential alternative to enhance the
919 forecasting accuracy using other pre-processing approaches, complete ensemble local mean
920 decomposition with adaptive noise, variational mode decomposition, complete ensemble
921 empirical mode decomposition(CEEMD), improved CEEMD, local mean decomposition
922 (LMD), and ensemble LMD.

Acknowledgments

The authors appreciate so much the facilities support by the Civil Engineering Department, Faculty of Engineering, University of Malaya, Malaysia, and the financial support received from research grant coded GPF082A-2018 funded by the University of Malaya and 2020106TELCO grant by the Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional (UNITEN) and Telkom University, Indonesia. Besides, the authors would like to thank the Department of Environment (DoE) for providing data and technical support.

References

Abbaszadeh, P., Moradkhani, H., Yan, H., 2017. Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov Chain Monte Carlo. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2017.11.011

Ahmad, Z., Rahim, N.A., Bahadori, A., Zhang, J., 2017. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int. J. River Basin Manag. 15, 79–87. https://doi.org/10.1080/15715124.2016.1256297

Ahmed, A.N., Hayder, G., Rahman, R.A.B.A., Borhana, A.A., 2019. Rainfall-runoff forecasting utilizing genetic programming technique. Int. J. Civ. Eng. Technol. 10, 1523–1534.

Al-Musawi, N., Al-Rubaie, F.M., Al-Musawi, N.O., 2017. prediction and assessment of water quality index using neural network model and gis case study: tigris river in baghdad city chlorine decay view project prediction and assessment of water quality index using neural network model and gis case study AND GIS. Appl. Res. J. 3, 343–353.

Amornsamankul, S., Road, S.A., Road, S.A., Road, S.A., n.d. Modified WQI Model using Fourier series and Genetic Algorithm Technique. Recent Res. Autom. Control Electron. Modif. 73–76.

Attar, N.F., Khalili, K., Behmanesh, J., Khanmohammadi, N., 2018. On the reliability of soft computing methods in the estimation of dew point temperature: The case of arid regions of Iran. Comput. Electron. Agric. 153, 334–346. https://doi.org/10.1016/j.compag.2018.08.029

Attar, N.F., Pham, Q.B., Nowbandegani, S.F., Rezaie-Balf, M., Fai, C.M., Ahmed, A.N., Pipelzadeh, S., Dung, T.D., Nhi, P.T.T., Khoi, D.N., El-Shafie, A., 2020. Enhancing the Prediction Accuracy of Data-Driven Models for Monthly Streamflow in Urmia Lake Basin Based upon the Autoregressive Conditionally Heteroskedastic Time-Series Model. Appl. Sci. 10, 571. https://doi.org/10.3390/app10020571

Babaei, S., F, Hassani, A H, Torabian, A, Karbassi, A R, Hosseinzadeh, L., F, 2011. Water quality index development using fuzzy logic: A case study of the Karoon River of Iran. African J. Biotechnol. 10, 10125–10133. https://doi.org/10.5897/AJB11.1608

Babbar, R., Babbar, S., 2017. Predicting river water quality index using data mining techniques. Environ. Earth Sci. 76. https://doi.org/10.1007/s12665-017-6845-9

960 Bagheri, M., Mirbagheri, S.A., Bagheri, Z., Kamarkhani, A.M., 2015. Modeling and
961 optimization of activated sludge bulking for a real wastewater treatment plant using hybrid
962 artificial neural networks-genetic algorithm approach. Process Saf. Environ. Prot. 95, 12–
963 25. https://doi.org/10.1016/J.PSEP.2015.02.008
964 Chang, N. Bin, Chen, H.W., Ning, S.K., 2001. Identification of river water quality using the
965 fuzzy synthetic evaluation approach. J. Environ. Manage. 63, 293–305.
966 https://doi.org/10.1006/jema.2001.0483
967 Chau, K., 2006. A review on integration of artificial intelligence into water quality modelling.
968 Mar. Pollut. Bull. 52, 726–733. https://doi.org/10.1016/J.MARPOLBUL.2006.04.003
969 Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J.,
970 Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use
971 of streamflow observations to update states in a distributed hydrological model. Adv. Water
972 Resour. 31, 1309–1324. https://doi.org/10.1016/J.ADVWATRES.2008.06.005
973 DOE, 2007. Malaysia Environmental Quality Report. Malaysia Env. Rep.1–86.
974 Dong, X., Lian, J., Wang, H., 2019. Vibration source identification of offshore wind turbine
975 structure based on optimized spectral kurtosis and ensemble empirical mode decomposition.
976 Ocean Eng. 172, 199–212. https://doi.org/10.1016/j.oceaneng.2018.11.030
977 El-Kowrany, S.I., El- Zamarany, E.A., El-Nouby, K.A., El-Mehy, D.A., Abo Ali, E.A., Othman,
978 A.A., Salah, W., El-Ebiary, A.A., 2016. Water pollution in the Middle Nile Delta, Egypt:
979 An environmental study. J. Adv. Res. 7, 781–794.
980 https://doi.org/10.1016/J.JARE.2015.11.005
981 Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using
982 Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99.
983 Fijani, E., Barzegar, R., Deo, R., Tziritis, E., Skordas, K., 2019. Design and implementation of a
984 hybrid model based on two-layer decomposition method coupled with extreme learning
985 machines to support real-time environmental monitoring of water quality parameters. Sci.
986 Total Environ. 648, 839–853. https://doi.org/10.1016/J.SCITOTENV.2018.08.221
987 Frei, M.G., Osorio, I., 2007. Intrinsic time-scale decomposition: time–frequency–energy analysis
988 and real-time filtering of non-stationary signals. Proc. R. Soc. A Math. Phys. Eng. Sci. 463,
989 321–342. https://doi.org/10.1098/rspa.2006.1761
990 Gaafar, M., Mahmoud, S.H., Gan, T.Y., Davies, E.G.R., 2019. A practical GIS-based hazard
991 assessment framework for water quality in stormwater systems. J. Clean. Prod.
992 https://doi.org/10.1016/j.jclepro.2019.118855
993 Gazzaz, N.M., Yusoff, M.K., Aris, A.Z., Juahir, H., Ramli, M.F., 2012. Artificial neural network
994 modeling of the water quality index for Kinta River (Malaysia) using water quality
995 variables as predictors. Mar. Pollut. Bull. 64, 2409–2420.
996 https://doi.org/10.1016/j.marpolbul.2012.08.005
997 Gazzaz, N.M., Yusoff, M.K., Ramli, M.F., Juahir, H., Aris, A.Z., 2015. Artificial Neural
998 Network Modeling of the Water Quality Index Using Land Use Areas as Predictors. Water
999 Environ. Res. 87, 99–112. https://doi.org/10.2175/106143014X14062131179276
1000 GB/T22482-2008, n.d. GB/T 22482-2008 - Standard for hydrological information and
1001 hydrological forecasting (TEXT OF DOCUMENT IS IN CHINESE) [WWW Document].
1002 URL https://webstore.ansi.org/standards/spc/gb224822008 (accessed 3.23.20).
1003 Gurjar, S.K., Tare, V., 2019. Spatial-temporal assessment of water quality and assimilative
1004 capacity of river Ramganga, a tributary of Ganga using multivariate analysis and QUEL2K.
1005 J. Clean. Prod. 222, 550–564. https://doi.org/10.1016/j.jclepro.2019.03.064

1006 Hameed, M., Sharqi, S.S., Yaseen, Z.M., Afan, H.A., Hussain, A., Elshafie, A., 2017.
1007 Application of artificial intelligence (AI) techniques in water quality index prediction: a
1008 case study in tropical region, Malaysia. Neural Comput. Appl. 28, 893–905.
1009 https://doi.org/10.1007/s00521-016-2404-7
1010 Ho, J.Y., Afan, H.A., El-Shafie, A.H., Koting, S.B., Mohd, N.S., Jaafar, W.Z.B., Lai Sai, H.,
1011 Malek, M.A., Ahmed, A.N., Mohtar, W.H.M.W., Elshorbagy, A., El-Shafie, A., 2019a.
1012 Towards a time and cost effective approach to water quality index class prediction. J.
1013 Hydrol. 575, 148–165. https://doi.org/10.1016/j.jhydrol.2019.05.016
1014 Ho, J.Y., Afan, H.A., El-Shafie, A.H., Koting, S.B., Mohd, N.S., Jaafar, W.Z.B., Lai Sai, H.,
1015 Malek, M.A., Ahmed, A.N., Mohtar, W.H.M.W., Elshorbagy, A., El-Shafie, A., 2019b.
1016 Towards a time and cost effective approach to water quality index class prediction. J.
1017 Hydrol. 575, 148–165. https://doi.org/10.1016/j.jhydrol.2019.05.016
1018 Hore, A., Dutta, S., Datta, S., Bhattacharjee, C., 2008. Application of an artificial neural network
1019 in wastewater quality monitoring: prediction of water quality index. Int. J. Nucl. Desalin. 3,
1020 160. https://doi.org/10.1504/IJND.2008.020223
1021 Ishikawa, Y., Murata, M., Kawaguchi, T., 2019. Globally applicable water quality simulation
1022 model for river basin chemical risk assessment. J. Clean. Prod. 239.
1023 https://doi.org/10.1016/j.jclepro.2019.118027
1024 Johns, C.J., Mandel, J., 2008. A two-stage ensemble Kalman filter for smooth data assimilation.
1025 Environ. Ecol. Stat. 15, 101–110. https://doi.org/10.1007/s10651-007-0033-0
1026 Juahir, H., Zain, S.M., Toriman, M.E., Mokhtar, M., Man, H.C., 2004. Application of Artificial
1027 Neural Network Models for Prediction Water Quality Index. Quality 16, 42–55.
1028 Justel, A., Peña, D., Zamar, R., 1997. A multivariate Kolmogorov-Smirnov test of goodness of
1029 fit. Stat. Probab. Lett. 35, 251–259. https://doi.org/10.1016/s0167-7152(97)00020-5
1030 Kadam, A.K., Wagh, V.M., Muley, A.A., Umrikar, B.N., Sankhua, R.N., 2019. Prediction of
1031 water quality index using artificial neural network and multiple linear regression modelling
1032 approach in Shivganga River basin, India. Model. Earth Syst. Environ. 1–12.
1033 https://doi.org/10.1007/s40808-019-00581-3
1034 Kalman, R.E., 1960. A New Approach to Linear Filtering and Prediction Problems. J. Basic Eng.
1035 82, 35. https://doi.org/10.1115/1.3662552
1036 Karunasingha, D.S.K., Liong, S.-Y., 2018. Enhancement of chaotic hydrological time series
1037 prediction with real-time noise reduction using Extended Kalman Filter. J. Hydrol. 565,
1038 737–746. https://doi.org/10.1016/J.JHYDROL.2018.08.044
1039 Kashif Gill, M., Kemblowski, M.W., McKee, M., 2007. Soil Moisture Data Assimilation Using
1040 Support Vector Machines and Ensemble Kalman Filter. J. Am. Water Resour. Assoc. 43,
1041 1004–1015. https://doi.org/10.1111/j.1752-1688.2007.00082.x
1042 Khalid, Samina, Murtaza, B., Shaheen, I., Ahmad, I., Ullah, M.I., Abbas, T., Rehman, F., Ashraf,
1043 M.R., Khalid, Sana, Abbas, S., Imran, M., 2018. Assessment and public perception of
1044 drinking water quality and safety in district Vehari, Punjab, Pakistan. J. Clean. Prod. 181,
1045 224–234. https://doi.org/10.1016/j.jclepro.2018.01.178
1046 Khan, F., Husain, T., Lumb, A., 2003. Water Quality Evaluation and Trend Analysis in Selected
1047 Watersheds of the Atlantic Region of Canada. Environ. Monit. Assess. 88, 221–248.
1048 https://doi.org/10.1023/A:1025573108513
1049 Khuan, L.Y., Hamzah, N., Jailani, R., 2002. Prediction of water quality index (WQI) based on
1050 artificial neural network (ANN). 2002 Student Conf. Res. Dev. Glob. Res. Dev. Electr.
1051 Electron. Eng. SCOReD 2002 - Proc. 157–161.

https://doi.org/10.1109/SCORED.2002.1033081
Kisi, O., Yaseen, Z.M., 2019. The potential of hybrid evolutionary fuzzy intelligence model for suspended sediment concentration prediction. Catena 174, 11–23. https://doi.org/10.1016/j.catena.2018.10.047
Kükrer, S., Mutlu, E., 2019. Assessment of surface water quality using water quality index and multivariate statistical analyses in Saraydüzü Dam Lake, Turkey. Environ. Monit. Assess. 191. https://doi.org/10.1007/s10661-019-7197-6
Kumar, B., Singh, U.K., Ojha, S.N., 2019. Evaluation of geochemical data of Yamuna River using WQI and multivariate statistical analyses: a case study. Int. J. River Basin Manag. 17, 143–155. https://doi.org/10.1080/15715124.2018.1437743
Lam, T.C., Ge, H., Fazio, P., 2016. Energy positive curtain wall configurations for a cold climate using the Analysis of Variance (ANOVA) approach. Build. Simul. 9, 297–310. https://doi.org/10.1007/s12273-016-0275-6
Leong, W.C., Bahadori, A., Zhang, J., Ahmad, Z., 2019. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). Int. J. River Basin Manag. 0, 1–8. https://doi.org/10.1080/15715124.2019.1628030
Lermontov, A., Yokoyama, L., Lermontov, M., Machado, M.A.S., 2009. River quality analysis using fuzzy water quality index: Ribeira do Iguape river watershed, Brazil. Ecol. Indic. 9, 1188–1197. https://doi.org/10.1016/j.ecolind.2009.02.006
Li, S., Gu, S., Tan, X., Zhang, Q., 2009. Water quality in the upper Han River basin, China: The impacts of land use/land cover in riparian buffer zone. J. Hazard. Mater. 165, 317–324. https://doi.org/10.1016/j.jhazmat.2008.09.123
Liou, S.-M., Lo, S.-L., Wang, S.-H., 2004. A Generalized Water Quality Index for Taiwan. Environ. Monit. Assess. 96, 35–52. https://doi.org/10.1023/B:EMAS.0000031715.83752.a1
Mahapatra, S.S., Nanda, S.K., Panigrahy, B.K., 2011. A Cascaded Fuzzy Inference System for Indian river water quality prediction. Adv. Eng. Softw. 42, 787–796. https://doi.org/10.1016/j.advengsoft.2011.05.018
Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ. Model. Softw. 25, 891–909. https://doi.org/10.1016/j.envsoft.2010.02.003
Martis, R.J., Acharya, U.R., Tan, J.H., Petznick, A., Tong, L., Chua, C.K., Kwee Ng, E.Y., 2013. Application of intrinsic Time-scale decomposition (ITD) to EEG signals for automated seizure prediction. Int. J. Neural Syst. 23, 1350023. https://doi.org/10.1142/S0129065713500238
Maxwell, D.H., Jackson, B.M., McGregor, J., 2018. Constraining the ensemble Kalman filter for improved streamflow forecasting. J. Hydrol. 560, 127–140. https://doi.org/10.1016/J.JHYDROL.2018.03.015
Mijares, V., Gitau, M., Johnson, D.R., 2019. A Method for Assessing and Predicting Water Quality Status for Improved Decision-Making and Management 509–522.
Moeeni, H., Bonakdari, H., 2017. Impact of Normalization and Input on ARMAX-ANN Model Performance in Suspended Sediment Load Prediction. Water Resour. Manag. 1–19. https://doi.org/10.1007/s11269-017-1842-z
Mohammadpour, R., Shaharuddin, S., Zakaria, N.A., Ghani, A.A., Vakili, M., Chan, N.W., 2016. Prediction of water quality index in free surface constructed wetlands. Environ. Earth Sci. 75, 139. https://doi.org/10.1007/s12665-015-4905-6

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28, 135–147. https://doi.org/10.1016/J.ADVWATRES.2004.09.002

Motagh, M., Shamshiri, R., Haghshenas Haghighi, M., Wetzel, H.-U., Akbari, B., Nahavandchi, H., Roessner, S., Arabi, S., 2017. Quantifying groundwater exploitation induced subsidence in the Rafsanjan plain, southeastern Iran, using InSAR time-series and in situ measurements. Eng. Geol. 218, 134–151. https://doi.org/10.1016/J.ENGGEO.2017.01.011

Najah, A., El-Shafie, A., Karim, O.A., Jaafar, O., 2011. Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations. Hydrol. Earth Syst. Sci. 15, 2693–2708. https://doi.org/10.5194/hess-15-2693-2011

Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., Elshafie, A., 2019. Machine learning methods for better water quality prediction. J. Hydrol. 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Nath, B.K., Chaliha, C., Bhuyan, B., Kalita, E., Baruah, D.C., Bhagabati, A.K., 2018. GIS mapping-based impact assessment of groundwater contamination by arsenic and other heavy metal contaminants in the Brahmaputra River valley: A water quality assessment study. J. Clean. Prod. 201, 1001–1011. https://doi.org/10.1016/j.jclepro.2018.08.084

Naubi, I., Zardari, N.H., Shirazi, S.M., Ibrahim, N.F.B., Baloo, L., 2016. Effectiveness of water quality index for monitoring Malaysian river water quality. Polish J. Environ. Stud. 25, 231–239. https://doi.org/10.15244/pjoes/60109

Nayak, P.C., Rao, Y.R.S., Sudheer, K.P., 2006. Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. Water Resour. Manag. 20, 77–90. https://doi.org/10.1007/s11269-006-4007-z

Nguyen Hien Than, Che Dinh Ly, Pham Van Tat, Nguyen Ngoc Thanh, 2016. Application of a Neural Network Technique for Prediction of the Water Quality Index in the Dong Nai River, Vietnam. J. Environ. Sci. Eng. B 5, 363–370. https://doi.org/10.17265/2162-5263/2016.07.007

Ocampo-Duque, W., Ferré-Huguet, N., Domingo, J.L., Schuhmacher, M., 2006. Assessing water quality in rivers with fuzzy inference systems: A case study. Environ. Int. 32, 733–742. https://doi.org/10.1016/j.envint.2006.03.009

Palizdan, N., Falamarzi, Y., Huang, Y.F., Lee, T.S., Ghazali, A.H., 2015. Temporal precipitation trend analysis at the langat river Basin, Selangor, Malaysia. J. Earth Syst. Sci. 124, 1623–1638. https://doi.org/10.1007/s12040-015-0636-z

Pham, H., Rahman, M.M., Nguyen, N.C., Vo, P. Le, Van, T. Le, Ngo, H., 2017. Assessment of Surface Water Quality Using the Water Quality Index and Multivariate Statistical Techniques-A Case Study: The Upper Part of Dong Nai River Basin, Vietnam. J. Water Sustain. 7, 225–245. https://doi.org/10.11912/jws.2017.7.4.

Rezaeian-Zadeh, M., Zand-Parsa, S., Abghari, H., Zolghadr, M., Singh, V.P., 2012. Hourly air temperature driven using multi-layer perceptron and radial basis function networks in arid and semi-arid regions. Theor. Appl. Climatol. https://doi.org/10.1007/s00704-012-0595-0

Rezaie-Balf, M., Fani Nowbandegani, S., Samadi, S.Z., Fallah, H., Alaghmand, S., 2019a. An Ensemble Decomposition-Based Artificial Intelligence Approach for Daily Streamflow Prediction. Water 11, 709. https://doi.org/10.3390/w11040709

Rezaie-Balf, M., Kim, S., Fallah, H., Alaghmand, S., 2019b. Daily river flow forecasting using ensemble empirical mode decomposition based heuristic regression models: Application on

the perennial rivers in Iran and South Korea. J. Hydrol. 572, 470–485. https://doi.org/10.1016/j.jhydrol.2019.03.046

Rezaie-Balf, M., Naganna, S.R., Kisi, O., El-Shafie, A., 2019c. Enhancing streamflow forecasting using the augmenting ensemble procedure coupled machine learning models: case study of Aswan High Dam. Hydrol. Sci. J. 1–18. https://doi.org/10.1080/02626667.2019.1661417

Robert K, H., 1965. An index number system for rating water quality. ournal Water Pollut. Control Fed. 3, 300–306.

Roveda, S.R.M.M., Bondança, A.P.M., Silva, J.G.S., Roveda, J.A.F., Rosa, A.H., 2010. Development of a water quality index using a fuzzy logic: A case study for the sorocaba river. 2010 IEEE World Congr. Comput. Intell. WCCI 2010. https://doi.org/10.1109/FUZZY.2010.5584172

Sahoo, M.M., Patra, K.C., Khatua, K.K., 2015. Inference of Water Quality Index Using ANFIA and PCA. Aquat. Procedia 4, 1099–1106. https://doi.org/10.1016/j.aqpro.2015.02.139

Said, A., Stevens, D.K., Sehlke, G., 2004. An Innovative Index for Evaluating Water Quality in Streams. Environ. Manage. 34, 406–414. https://doi.org/10.1007/s00267-004-0210-y

Sarkar, S.K., Saha, M., Takada, H., Bhattacharya, A., Mishra, P., Bhattacharya, B., 2007. Water quality management in the lower stretch of the river Ganges, east coast of India: an approach through environmental education. J. Clean. Prod. 15, 1559–1567. https://doi.org/10.1016/j.jclepro.2006.07.030

Sharma, D., Lie, T.T., 2012. Wind speed forecasting using hybrid ANN-Kalman Filter techniques, in: 2012 10th International Power & Energy Conference (IPEC). IEEE, pp. 644–648. https://doi.org/10.1109/ASSCC.2012.6523344

Singh, J., Knapp, H.V., Arnold, J.G., Demissie, M., 2005. Hydrological modeling of the Iroquois River watershed using HSPF and SWAT. J. Am. Water Resour. Assoc. 41, 343–360. https://doi.org/10.1111/j.1752-1688.2005.tb03740.x

Sinha, K., (Saha), P. Das, 2014. Assessment of water quality index using cluster analysis and artificial neural network modeling: a case study of the Hooghly River basin, West Bengal, India. Desalin. Water Treat. 54, 28–36. https://doi.org/10.1080/19443994.2014.880379

Soo, E.Z.X., Jaafar, W.Z.W., Lai, S.H., Othman, F., Elshafie, A., Islam, T., Srivastava, P., Hadi, H.S.O., 2019. Evaluation of bias-adjusted satellite precipitation estimations for extreme flood events in Langat river basin, Malaysia. Hydrol. Res. 51, 105–126. https://doi.org/10.2166/nh.2019.071

Suhaila, J., Deni, S.M., Zawiah Zin, W.A.N., Jemain, A.A., 2010. Trends in Peninsular Malaysia rainfall data during the southwest monsoon and northeast monsoon seasons: 1975-2004. Sains Malaysiana 39, 533–542.

Tan, M.L., Samat, N., Chan, N.W., Lee, A.J., Li, C., 2019. Analysis of precipitation and temperature extremes over the Muda River Basin, Malaysia. Water (Switzerland) 11, 1–17. https://doi.org/10.3390/w11020283

Tangang, F.T., Juneng, L., Salimun, E., Sei, K.M., Le, L.J., Muhamad, H., 2012. Climate change and variability over Malaysia: Gaps in science and research information. Sains Malaysiana 41, 1355–1366.

Tasneem Abbasi, Shahid A., A., 2012. Water quality indices.

Taud, H., Mas, J.F., 2018. Multilayer Perceptron (MLP). Springer, Cham, pp. 451–455. https://doi.org/10.1007/978-3-319-60801-3_27

Tiwari, S., Babbar, R., Kaur, G., 2018. Performance Evaluation of Two ANFIS Models for

Predicting Water Quality Index of River Satluj (India). Adv. Civ. Eng. 2018, 1–10. https://doi.org/10.1155/2018/8971079

Wang, Z., 2018. Hourly Solar Radiation Forecasting Using a Volterra-Least Squares Support Vector Machine Model Combined with Signal Decomposition. Energies 11, 68. https://doi.org/10.3390/en11010068

Wu, Z., Wang, X., Chen, Y., Cai, Y., Deng, J., 2018. Assessing river water quality using water quality index in Lake Taihu Basin, China. Sci. Total Environ. 612, 914–922. https://doi.org/10.1016/j.scitotenv.2017.08.293

Yahya, A.S.A., Ahmed, A.N., Othman, F.B., Ibrahim, R.K., Afan, H.A., El-Shafie, A., Fai, C.M., Hossain, M.S., Ehteram, M., Elshafie, A., 2019. Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. Water (Switzerland) 11. https://doi.org/10.3390/w11061231

Yaseen, Z.M., Ramal, M.M., Diop, L., Jaafar, O., Demir, V., Kisi, O., 2018. Hybrid Adaptive Neuro-Fuzzy Models for Water Quality Index Estimation. Water Resour. Manag. 32, 2227–2245. https://doi.org/10.1007/s11269-018-1915-7

Yilma, M., Kiflie, Z., Windsperger, A., Gessese, N., 2018. Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia. Model. Earth Syst. Environ. 0, 0. https://doi.org/10.1007/s40808-018-0437-x

Yu, C., Li, Y., Zhang, M., 2017. Comparative study on three new hybrid models using Elman Neural Network and Empirical Mode Decomposition based technologies improved by Singular Spectrum Analysis for hour-ahead wind speed forecasting. Energy Convers. Manag. 147, 75–85. https://doi.org/10.1016/j.enconman.2017.05.008

Zhang, Q., Li, Z., 2019. Development of an interval quadratic programming water quality management model and its solution algorithms. J. Clean. Prod. 119319. https://doi.org/10.1016/j.jclepro.2019.119319

Zhang, X., Zhang, Q., Zhang, G., Nie, Z., Gui, Z., Que, H., Zhang, X., Zhang, Q., Zhang, G., Nie, Z., Gui, Z., Que, H., 2018. A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition. Int. J. Environ. Res. Public Health 15, 1032. https://doi.org/10.3390/ijerph15051032

Zhou, Y., Chang, F.-J., Guo, S., Ba, H., He, S., 2017. A robust recurrent ANFIS for modeling multi-step-ahead flood forecast of Three Gorges Reservoir in the Yangtze River. Hydrol. Earth Syst. Sci. Discuss. 1–29. https://doi.org/10.5194/hess-2017-457

Highlight

As a comprehensive review, pH and DO were the most influential parameters for WQI prediction.

Ensemble Kalman Filter as the DA technique is applied to generate an accurate state estimation.

For improving the physicochemical data to noise ratio, ITD approach hybridized with EnKF-ANN.

The new ITD-EnKF-ANN generally outperformed other standalone and hybrid DDMs for the prediction of WQI.

Credit Author Statement

Mohammad Rezaie-Balf: Conceptualization, Software, Supervision.

Nasrin Fathollahzadeh Attar: Methodology, Writing- Original draft preparation.

Ardashir Mohammadzadeh: Software, Methodology.

Muhammad Ary Murti: Resources, Writing- Original draft preparation.

Ali Najah Ahmed: Writing- Reviewing and Editing.

Chow Ming Fai: Resources, Writing- Original draft preparation.

Narjes Nabipour: Writing and Reviewing, Formal analysis.

Sina Alaghmand: Writing- Reviewing and Editing.

Ahmed El-Shafie: Visualization, Supervision.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: