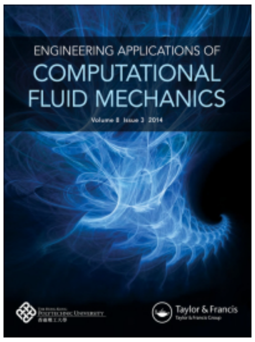


Engineering Applications of Computational Fluid Mechanics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tcfm20

Water level prediction using various machine learning algorithms: a case study of Durian Tunggal river, Malaysia

Ali Najah Ahmed, Ayman Yafouz, Ahmed H. Birima, Ozgur Kisi, Yuk Feng Huang, Mohsen Sherif, Ahmed Sefelnasr & Ahmed El-Shafie

To cite this article: Ali Najah Ahmed, Ayman Yafouz, Ahmed H. Birima, Ozgur Kisi, Yuk Feng Huang, Mohsen Sherif, Ahmed Sefelnasr & Ahmed El-Shafie (2022) Water level prediction using various machine learning algorithms: a case study of Durian Tunggal river, Malaysia, Engineering Applications of Computational Fluid Mechanics, 16:1, 422-440, DOI: 10.1080/19942060.2021.2019128

To link to this article: https://doi.org/10.1080/19942060.2021.2019128

Published online: 06 Feb 2022.

ENGINEERING APPLICATIONS OF COMPUTATIONAL FLUID MECHANICS
2022, VOL. 16, NO. 1, 422–440
https://doi.org/10.1080/19942060.2021.2019128

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Water level prediction using various machine learning algorithms: a case study of Durian Tunggal river, Malaysia

Ali Najah Ahmed [a], Ayman Yafouz [a], Ahmed H. Birima[b], Ozgur Kisi [c,d], Yuk Feng Huang [e], Mohsen Sherif [f,g], Ahmed Sefelnasr [g] and Ahmed El-Shafie [h]

[a]Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Selangor, Malaysia; [b]Department of Civil Engineering, College of Engineering, Qassim University, Unaizah, Saudi Arabia; [c]Department of Civil Engineering, Ilia State University, Tbilisi, Georgia; [d]Department of Civil Engineering, University of Applied Sciences, Lübeck, Germany; [e]Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang, Malaysia; [f]Civil and Environmental Engineering Department, College of Engineering, United Arab Emirates University, Al Ain, UAE; [g]National Water and Energy Center, United Arab Emirates University, Al Ain, UAE; [h]Department of Civil Engineering, Faculty of Engineering, University of Malaya (UM), Kuala, Lumpur, Malaysia

ABSTRACT

A reliable model to predict the changes in the water levels in a river is crucial for better planning to mitigate any risk associated with flooding. In this study, six different Machine Learning (ML) algorithms were developed to predict the river's water level, on a daily basis based on collected data from 1990 to 2019 which were used to train and test the proposed models. Different input combinations were explored to improve the accuracy of the model. Statistical indicators were calculated to examine the reliability of the proposed models with other models. The comparison of several data-driven regression methods indicate that the exponential Gaussian Process Regression (GPR) model offered better accuracy in predicting daily water levels with respect to different assessment criteria. The GPR model was then used to predict the water level after sorting the data based on 10 days maximum and minimum values of the water level, and the results proved the success of this model in catching the extremes of the water levels. In addition to that, based on two uncertainty indicators, it was concluded that the proposed model, the GPR, was capable of predicting the water level of the river with high precision and less uncertainty where the computed using the 95% prediction uncertainty (95PPU) and the d-factor were found to be equal to 98.276 and 0.000525, respectively. The findings of this study show the efficacy of the GPR model in capturing the changes in the water level in a river. Due to the importance of the water level of a river being an parameter for flood monitoring, this technique is likely beneficial to the design of the mitigation strategies for future flooding events.

ARTICLE HISTORY
Received 5 April 2021
Accepted 12 December 2021

KEYWORDS
Machine learning; Gaussian process regression; water level; Durian Tungal river; Malaysia

Introduction

A reliable model to predict the changes in a river water levels is crucial for the better planning fr mitigating risks associated with such changes, such as flooding (Mosavi et al., 2018; Rosli et al., 2015; Tikhamarine, Souag-Gamane, Najah Ahmed, et al., 2020; Olivia Muslim et al., 2018). Therefore, many studies have investigated various techniques to develop a robust model capable of capturing the water levels' fluctuations (Fotovatikhah et al., 2018; Yaseen et al., 2019). However, most traditional models such as the autoregressive model and the Box–Jenkins, have many limitations wherein many parameters are needed to be collected to build these models and many parameters require prior information and calibration (Ehteram et al., 2018; El-Shafie et al., 2014; Kumar et al., 2021). For example, one of the limitations associated with the physical-based models is the need for many assumptions, and there is also a need to have a sufficient and comprehensive understanding of the problem before making any assumptions (Carmo, 2020). For example, in Malaysia, a physical-based model was developed to assess one river's floodplain and water level (Mohamad et al., 2014). In order to build such a physical model, there was the need to collect massive data and information on top of the cost involved in building such a physical model. Over time, numerical models overcame the limitations of the physics-based models. For instance, a numerical model was developed by (Wu et al., 2014) to forecast water levels at the Yangtze River. However, a study conducted by (Guan et al., 2013) reported that many errors were encountered during the development of a numerical model. Therefore, despite the noticeable improvements in the numerical models, they still have limitations, such as the need to mimic some of

CONTACT Ali Najah Ahmed mahfoodh@uniten.edu.my

the physical phenomena to improve their accuracy and reliability.

Recently, data-driven techniques were shown to have overcome traditional models' drawbacks and proved to be more accurate in modeling complex engineering problems. For example, various machine learning (ML) algorithms were developed to predict different hydrological parameters. In (Allawi et al., 2019), two ML algorithms, namely the radial basis function(RBF) and the support vector model (SVM) were developed to predict reservoir evaporation over three different time horizons: daily, weekly and monthly. They found that the ML algorithms were more robust and capable of predicting the evaporation rates with a high accuracy level (daily). Similar findings were reported in (Kamel et al., 2021), where the ML algorithms were used to predict the sub-surface evaporation rates. In (Hipni et al., 2013), a support vector machine (SVM) was used to forecast the Klang gate reservoir's daily water level. They found that the SVM outperformed the ANFIS model. It could be concluded from the previous studies that the ML algorithms exhibit better accuracy compared with traditional methods; however, there are limitations with some of the conventional ML algorithms. For example, some of the ML algorithms' hyper-parameters such as weights, activation and transfer functions need to be optimized to achieve a high level of precision. Hyper-parameters are internal parameters inherent in the ML algorithms that can be tuned and optimized since they have a strongly impact on the ML models' performance (Ahmed et al., 2021). Therefore, many recent studies of hybridized ML algorithms with optimization techniques to overcome the conventional ML algorithms' limitations. For instance, in (Tikhamarine, Souag-Gamane, Ahmed, et al., 2020), ML algorithms were integrated with different meta-heuristics optimizers to predict the streamflow. They found that the hybridized models outperformed the standalone base model and exhibited better precision in capturing the rivers' streamflow changes. Similar findings have been reported and proved the superiority of the ML algorithms when they integrated with optimization algorithms (Attar et al., 2020; Banadkooki et al., 2020; Ehteram et al., 2020b; Osman et al., 2020; Pham et al., 2020; Rezaie-Balf et al., 2020). However, despite the achieved and perceived precision after introducing the hybridized models, there still are demerits in such hybridization. For example, some of the optimization techniques such as the particle swarm optimization and genetic algorithm have a low convergence rate, and the convergence tend to be not stable, while others might easily fall into the local optimum trap, and others can cause an increase in the computational time (Mohammadi et al., 2020). To deal with these shortcomings, new advanced ML algorithms have been developed. Tree regression models were proposed by many researchers as an alternative to the hybrid models to predict stochastic parameters. Tree family models proved to be less complex and accurate (Ghazvinian et al., 2019). In (Ridwan et al., 2020), three ML algorithms were developed, namely the boosted decision tree (BDT), the decision forest (DF), and the artificial neural network (ANN), to predict the daily precipitation. They found that the tree ML models outperformed the traditional model and could predict the precipitation with an acceptable accuracy level. In additional to that, in this study the tree regression models outperformed the standard ML models. For water level prediction, in (Sapitang et al., 2020), the boosted linear regression (BLR) and the boosted decision tree regression (BDTG) were proposed to predict the water level at a reservoir in Malaysia. They found that the tree-based ML algorithms were more reliable than the traditional and conventional ML algorithms such as the ANN and the RBF. Another good ML algorithm is the extreme gradient boosting algorithm (XGBoost) where this algorithm displays a high level of accuracy and speed in dealing with non-linear and complex parameters. For example, in (Ibrahem Ahmed Osman et al., 2021), the XGBoost was used to capture the sudden fluctuations in groundwater level in separate wells. They found that the XGBoost outperformed all other ML algorithms and captured the groundwater level fluctuations with high accuracy and in a short time. This ensemble machine learning technique is an effective and sophisticated enforcement of a gradient boosting framework (Melville, 2014). The XGBoost is a very operative and excessively utilized machine learning approach that analysts massively apply to obtain desirable results on various machine learning challenges. The algorithm progresses faster than the existing popular solutions on a single machine by ten times (Yafouz et al., 2021).

One of the recent supervised ML algorithms adopted by many researchers to deal with complex engineering problems is the Gaussian process regression model (GRP). The Ni-Based single-crystal superalloys' nonlinear deformation mechanism behavior was modeled precisely by the GRP (Zhang & Xu, 2021). The GRP showed superior performance in estimating wind speed after it was compared to a benchmarked model (Lio et al., 2021). In another study by (Wilkie & Galasso, 2021), they developed the GP to model the offshore wind turbine's fatigue damage. In their study, the GRP was developed and validated in three different wind farms. They found the GRP model to be more reliable and can reduce the computational time spent assessing fatigue damage using other conventional models.

Therefore, in this study, six different ML algorithms were developed to predict the river's water level at the

Durian Tungal river in the state of Malacca, chosen as the study area due to its importance. It has been reported that the Durian Tungal river's water levels rose many times during the last two years and overflowed the river banks (Arbain & Wibowo, 2012). Different input combinations were explored to improve the model's accuracy and gauge the sensitivity of the model to these inputs. In addition, statistical indicators were calculated to examine the reliability of the proposed model with other models. Finally, an uncertainty analysis was carried out to examine the consistency of the proposed model. The best-proposed model would next be used to predict the extreme maximum and minimum changes in water level over a 10-days lead. The performance of the proposed model was then compared with other models developed for the same purpose, found in the literature. One of the contributions in this study is the development of the six widely used ML algorithms to capture the changes in the water levels for the chosen study area. And the introduction of two ML algorithms, namely the XGBoost and the GPR models, has never been used before to predict the level of water in rivers. In addition, this study carried out a comprehensive analysis to prove the robustness of the Gaussian Process Regression (GPR) model, which can be used as a reliable tool for predicting the river water level. However, one of the limitations in this study is the availability, completeness and consistency of data. Most of the data from monitored stations do come with a problem, such as many missing data. Unfortunately, this predicament seems to appear too often at many areas all over the world, worth investigating. But nevertheless, however daunting, there should be a start with the best possible data scenario, and hopefully in future, further and more advance studies can be continued at site, when the data situation improves. Therefore, only data from a single station was used at our study area, after ensuring that the data set is complete and consistent. The authors would like to acknowledge the access of data from the Department of Irrigation and Drainage Malaysia (JPS). The data of this study can be requested directly from JPS. The code of this study is available on request from the corresponding author.

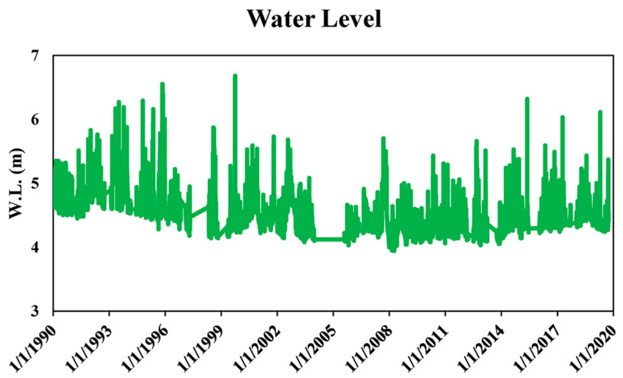
Methodology

Regarding the study area and data acquisition, the Durian Tungal river investigated in this study, located in Malacca state in Malaysia, is as shown in Figure 1. Malacca is one of the Malaysia states which has been experiencing annual flash flood as other states (Dashti Latif et al., 2020). Monitoring stations network to monitor the status of the river in terms of the river's water level and the rainfall data have been measured and collected by the department of drainage and irrigation in Malaysia. In this study, the data used for training and testing are collected on a daily basis from 1990 to 2019. As mentioned previously, data was collected from only one monitoring station installed and monitored by the Department of Irrigation and Drainage of Malaysia, and the station ID is 2322415 and is located at Latitude equal to 2.3236 deg and Longitude equal to 102.2931 deg. Table 1 and

Figure 1. Location of the study area.

Table 1. Data description for the input and output parameters.

Input parameter	Water Level (W.L.) (m) Above mean sea level	Rainfall (R.F.) (mm)
Count	8453	8453
Mean	4.487770	2.913817
Std	0.303889	4.896668
Min	3.940000	0.000000
25%	4.270000	0.062500
50%	4.410000	0.941667
75%	4.620000	3.687500
Max	6.680000	98.637500



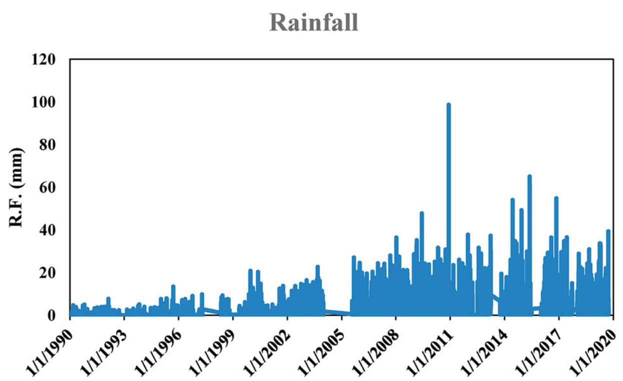


Figure 2. Water levels and rainfall series at the station ID 2322415.

Figure 2 depict the primary statistical analysis of the two parameters, water level and rainfall.

In order to optimize the accuracy of the proposed model, choosing the right input combinations is vital, (Lai et al., 2019), (Ali Najah et al., 2011). Therefore, in this study, the partial autocorrelation function was used to choose the best lag-time to be considered for both the rainfall and water level, as shown in Figure 3.

Based on the figure above, there is a decrease in the partial autocorrelation function value when the lag-time increases and it dropped drastically after the lag exceeds 5 days for the rainfall dataset and 7 days for the water level dataset. Therefore, to measure the sensitivity of each model, in this study, four different input combinations were investigated based on different lag-time combinations, as shown in Table 2.

Due to the wide variation in the precipitation over the chosen period (1990–2019), rainfall was included as input to the model as representative to the effect of the historical climate change at that study area.

It is true that the influences of the rainfall of the previous days have been reflected in the current water level; however, this is only the direct surface runoff effects, which is considered as the direct effect of the rainfall on the water level. In fact, there is another component that the rainfall affects the water level indirectly, and which is the sub-surface drainage. The sub-surface drainage is considered as another component of the rainfall-runoff spectacle and phenomenon that might take longer time to reflect on the water level. Since, the sub-surface drainage amount is embedded and dependent to the rainfall intensity which will definitely contribute indirectly to the water level, it is better to include the rainfall of the previous days to indirectly include the effect of the sub-surface drainage on the water level in the model.

The second step after selecting the four different scenarios is the date splitting. In literature, different splitting ratios were proposed (Abdullah et al., 2019; Ehteram et al., 2020a; Khan et al., 2020). However, the most used ratio is by splitting the data into two subsets where four-fifths of the data were used to train the model, while the second subset, which consists of the remaining samples from the entire dataset, was used to test the model in predicting unseen data (A. A. Najah et al., 2011; Najah Ahmed et al., 2012; Napi et al., 2020, 2021). In this study, 80% of the data was used to train the models (1990–2015), while 20% (2016–2019) was used to test the proposed models' reliability in predicting the changes in the level of river's water. Six different ML models and techniques were investigated in this study. The first is the linear regression model, which was developed with four different techniques.

Linear regression models have linear/non-linear parameters and make prediction uncomplicated because it does not rely on the degree of the model sophistication. Due to that, many researchers begin exploring this model before moving on to a more complex model (El-Shafie et al., 2011). However, such a model exhibits a low level of accuracy in solving the non-linearity in engineering parameters (Abdullah et al., 2019). In this study, four different linear regression models (linear, interaction, robust and stepwise) were developed during the study's first phase. A robust model is usually used to reduce the model's sensitivity if there are outliers in the data using a robust objective function. Using linear regression models proves that the fluctuations in the level of water in the river are difficult to capture by such models (Štefelová et al., 2021).

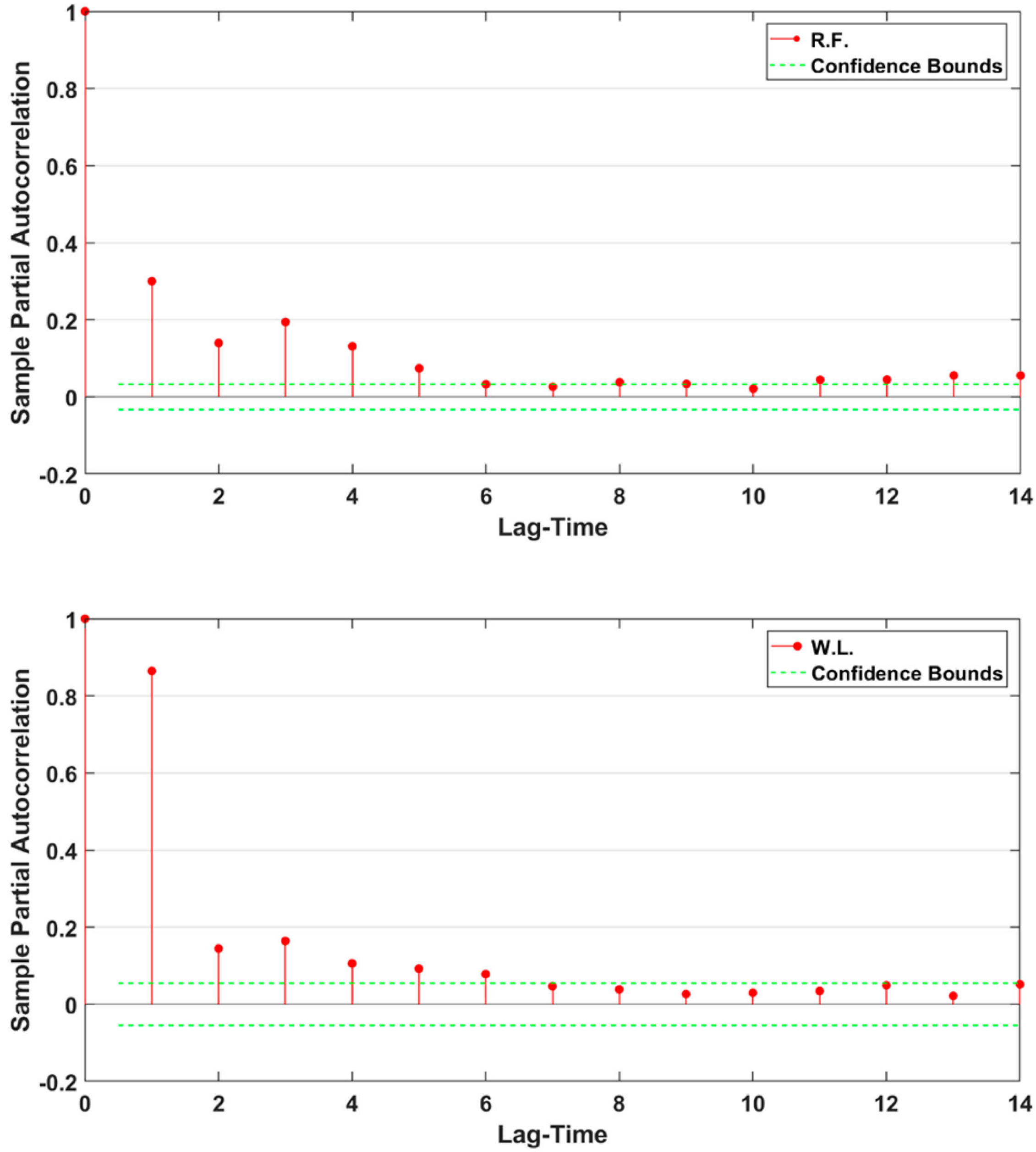


Figure 3. Partial autocorrelation function with different lag-time (day).

Table 2. The proposed four scenarios are based on different input combinations to gauge the models' sensitivity.

	Input combination	Output
Scenario-1	R.F.(t-1), W.L.(t-1) and R.F.(t)	W.L.(t)
Scenario-2	R.F.(t-1), W.L.(t-1), R.F.(t-2), W.L.(t-2), R.F.(t-3), W.L.(t-3) and R.F.(t)	W.L.(t)
Scenario-3	R.F.(t-1), W.L.(t-1), R.F.(t-2), W.L.(t-2), R.F.(t-3), W.L.(t-3), R.F.(t-4), W.L.(t-4), R.F.(t-5), W.L.(t-5) and R.F.(t)	W.L.(t)
Scenario-4	R.F.(t-1), W.L.(t-1), R.F.(t-2), W.L.(t-2), R.F.(t-3), W.L.(t-3), R.F.(t-4), W.L.(t-4), R.F.(t-5), W.L.(t-5), R.F.(t-6), W.L.(t-6), R.F.(t-7), W.L.(t-7) and R.F.(t)	W.L.(t)

The Support vector regression model was employed by many researchers to account for the non-linearity in the data and delivers a more accurate prediction (Banadkooki et al., 2019; Ghazvinian et al., 2019; A. Najah et al., 2011). In this study, six different models of this technique were developed: namely linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian models. Three different families from tree regression models were introduced in this study, such as both the standard and the ensemble and the extreme gradient boosting models.

An extreme gradient boosting model is proposed for improving comparison, and increasing searches for sequences of decision nodes to reduce the tree prior values. Such an advantage makes this algorithm superior and more accurate, and faster to learn compared to the other tree regression models (Aljanabi et al., 2020). On

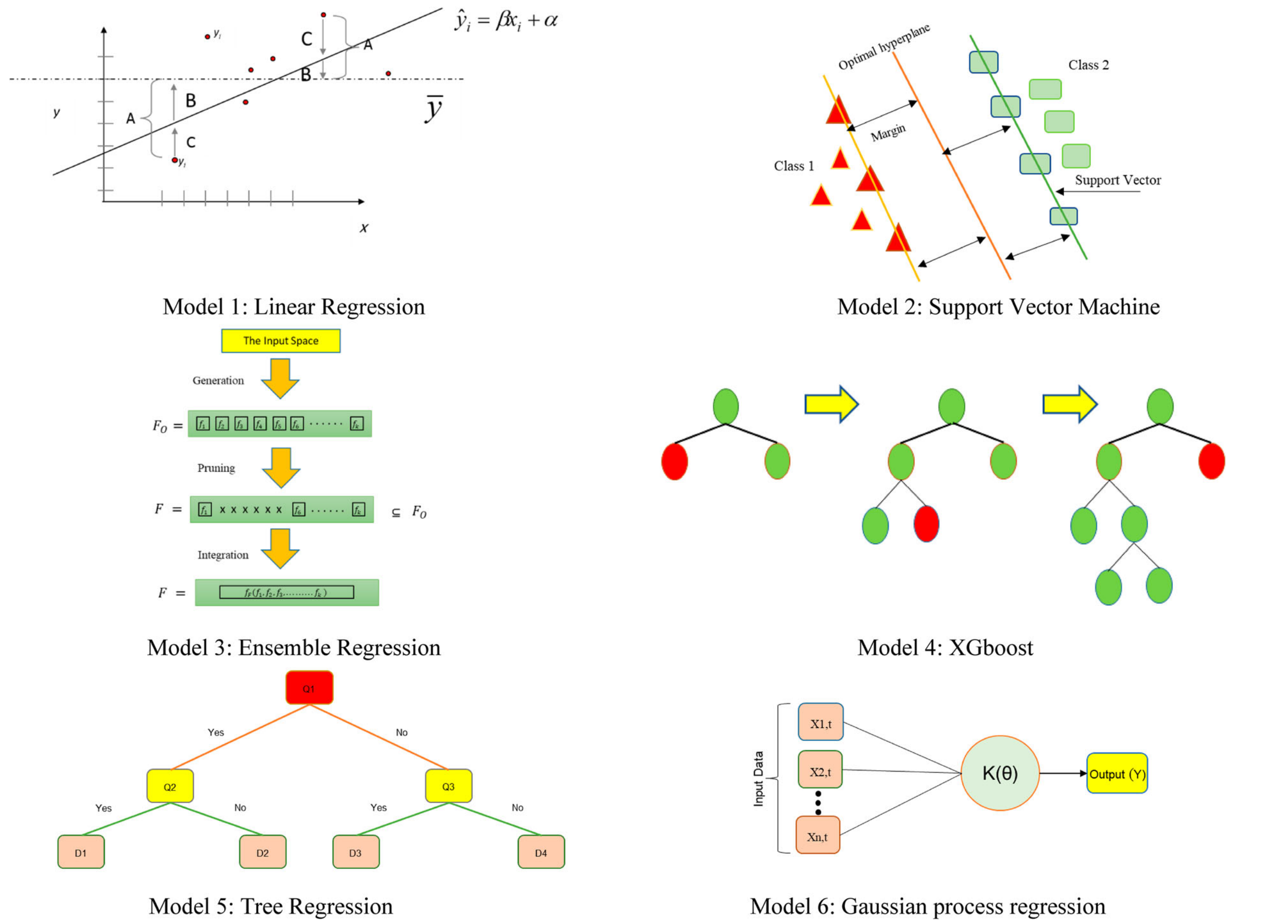


Figure 4. Architectures of the proposed ML models in this study.

top of that, this algorithm proves to be reliable in dealing with data with many missing values and can overcome the over-fitting problems (Ibrahem Ahmed Osman et al., 2021). Therefore, this algorithm was utilized to predict the river's water level, and a comprehensive comparison between all the algorithms would be carried out in this study.

The Gaussian process regression algorithm is one of the non-parametric kernel-based probabilistic techniques. More comprehensive details about this algorithm can be found in (Rasmussen, 2004). One of this algorithm's advantages is the use of the covariance function for regulation purposes, making this algorithm more reliable and robust than other ML algorithms (Sun et al., 2014). However, the one limitation in this algorithm is the loss of its efficiency if there are many numbers of features (Band et al., 2021). Such a limitation is negligible since the high dimensionality is not present in the data used in this study. Figure 4 demonstrates an example from the architecture of the proposed models.

During the training phase, the loss was used as a reliable stopping criteria technique to avoid overfitting (Najah Ahmed et al., 2019). In addition to that, a cross-validation technique with different folds (3, 5, 7 and 9) would be carried out. Regarding the hyper-parameters tuning, during the training phase, all of the six models would be optimized. For example, for the Gaussian Process Regression Model, four different kernel's functions were explored to optimize the accuracy of the model. The explored kernel functions are the squared exponential, Matern 5/2, rational quadratic, and exponential function. Uncertainty analysis was carried out using two techniques, namely the 95PPU and the d-factor. Finally, to validate the reliability of each model, in this study, four different statistical indicators were used to validate and compare the accuracy of each model. Training speeds of the models were also compared and are presented in the discussion section.

Coefficient of determination (R^2)

The R^2 is the ratio of the disparity in the dependent variable which is predicted with the aid of the independent variable. It is an indicator generated from regression analysis. In linear regression, the R^2 is also equal to the square of the correlation (R) between predicted *P*-value and actual O values; owing to this, it is a value from 0 to 1 (Nur Adli Zakaria et al., 2021). In the no correlation case, $R^2 = 0$, which means the independent variable is incapable of forecasting the dependent variable. In the case of no correlation error, $R^2 = 1$, which means the independent variable is capable of forecasting the dependent variable with no error. If R^2 is between 0 and 1, it refers to what extent the ability to predict the dependent variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O_m)^2} \tag{1}$$

where O_i is the true value at time i; O_m is the mean number of the recorded result, P_i is the anticipated result at time i and n is the total samples number.

Mean absolute error

The Mean Absolute Error (MAE) is a statistical tool to calculate the errors between the data which is representing a specific event (Kamel et al., 2021). It is the comparisons of O_i observed versus P_i predicted, such as subsequent time versus initial time. Also, it can be used for one measurement technique versus another measurement technique. The MAE equation is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \tag{2}$$

Mean square error

The Mean Square Error (MSE) or mean squared deviation (MSD) is an indicator for the predicting data by measuring the average of the squares of the errors, which is obtained from the average squared difference among the predicted data and the observed data (Ali Najah et al., 2021). Indeed, it describes how the set of a point near the regression line is (Sami et al., 2021). As shown in Equation (3), it does this by taking the distances from the points to the regression line (these distances are the 'errors') and squaring them. The squaring is necessary to remove any negative signs.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2 \tag{3}$$

Root mean square error

The Root Mean Square Error (RMSE) is known as the residuals' standard deviation residuals (forecast errors). Residuals are the distance of data points from the regression line. The RMSE is a statistical tool to identify how these residuals are spread out. Indeed, it demonstrates how the data is allocated surround the best fit line (Chong et al., 2021; Teo et al., 2021). RMSE is usually utilized in the weather field and analysis of regression cases to validate the results of an experiment. The equation is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{4}$$

Figure 5, which shows the flow-chart of the proposed method in this study, demonstrates all the proposed methodology phases to attain all the objectives of the current study.

Results and discussion

The viability of the different data-driven methods; the linear regression (LR), interaction regression (IR), robust regression (RR), stepwise regression (SR), support vector regression (SVR), boosted trees ensemble regression (BOOSTER), bagged trees ensemble regression (BAGER), XGBoost, tree regression (TR) and Gaussian process regression (GPR), was investigated for daily water level prediction. Table 3 sums up the test performances of the LR, IR, RR and SR in predicting water level. Training speeds of the models are also compared in the last columns of the table. As is clear from Table 3, the IR performs superior to the others in all input scenarios while the RR provides the worst performance. Among the IR models, scenario 4 has the lowest RMSE (0.13334 m) and the highest R^2 (0.83). As expected, the LR generally has the lowest training time while the SR may need long duration for training (e.g. 323 s for the scenario 4). It can be said that the increasing number of lagged inputs improves the models' accuracy and scenario 4 generally produces the best accuracy; decrease in RMSE of IR is from 0.14455 m (scenario 1) to 0.13334 m (scenario 4) and increase in R^2 from (0.80 to 0.83) for the corresponding scenarios.

Test statistics of the support vector regression (SVR) models are compared in Table 4 for the daily water level prediction. As seen from the table, the cubic based SVR outperforms the other alternatives with the lowest RMSE (0.1239 m) and MAE (0.062764 m) and the highest R^2 (0.86) while the linear based SVR has the worst accuracy (RMSE: 0.14136 m, MAE: 0.073663 m and R^2: 0.81 for the best model) as expected. However, the cubic based SVR has the highest training duration (ranges from 313.29 to 548.84 s) among the different SVR models. Here, also the best accuracies come from scenario 4 in general and increasing lagged inputs of R.F. and W.L. has positive effect on models' performances; decrease in RMSE of cubic SVR is from 0.1478 m (scenario 1) to 0.1239 m (scenario 4) and increase in R^2 from 0.79 to 0.86.

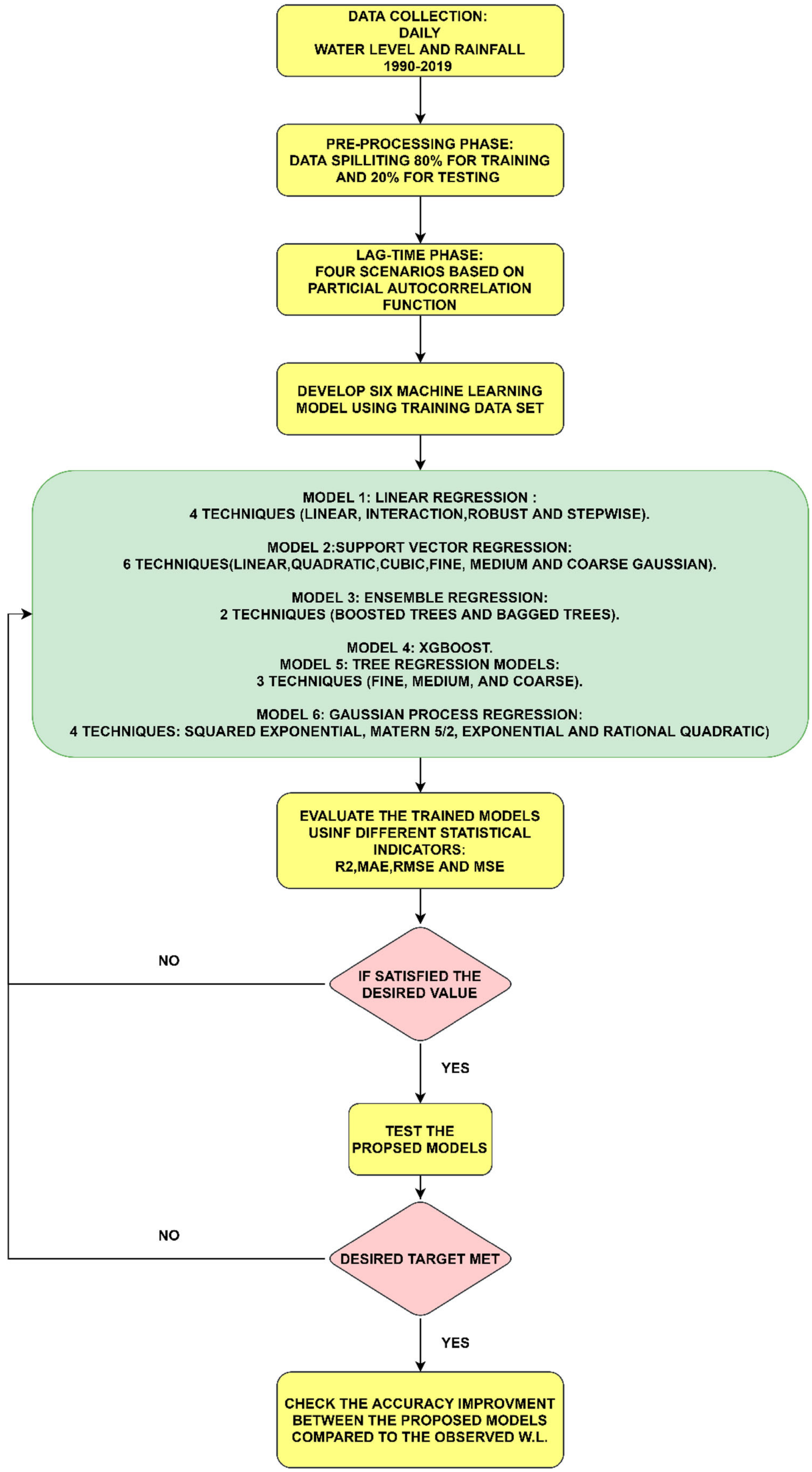


Figure 5. Flow-chart of the proposed method in this study.

Table 3. Performances of the linear regression models.

MODEL 1	Performance during testing				
Linear Regression Models	RMSE	R-Squared	MAE	Prediction speed (obs/sec)	Training time (sec)
Linear					
Scenario-1	0.14527	0.8	0.079943	150000	6.6466
Scenario-2	0.14134	0.81	0.078519	79000	11.424
Scenario-3	0.13992	0.82	0.077749	150000	4.4912
Scenario-4	0.13911	0.82	0.077376	90000	26.936
Interaction					
Scenario-1	0.14455	0.8	0.078839	150000	5.9487
Scenario-2	0.13754	0.82	0.07621	70000	10.323
Scenario-3	0.13519	0.83	0.075405	120000	4.1478
Scenario-4	0.13334	0.83	0.074818	170000	2.5597
Robust					
Scenario-1	0.14875	0.79	0.075521	120000	5.5355
Scenario-2	0.1487	0.79	0.075068	68000	9.7939
Scenario-3	0.14732	0.8	0.074371	140000	3.9217
Scenario-4	0.14639	0.8	0.073897	69000	37.932
Stepwise					
Scenario-1	0.14461	0.8	0.078885	220000	13.821
Scenario-2	0.13771	0.82	0.076283	65000	23.209
Scenario-3	0.13567	0.83	0.075493	110000	47.555
Scenario-4	0.13452	0.83	0.075413	180000	323.75

Through Table 5, the statistics of ensemble regression models are reported and compared in predicting water level for the test stage. The superiority of the BAGER over the BOOSTER is apparent for all scenarios. Among the BAGER models, scenario 3 has the lowest RMSE (0.1055 m) and MAE (0.05605 m) and the highest R^2 (0.90). There is a marginal difference between the training durations of the BAGER and BOOSTER methods. The accuracy difference between the scenario 3 and scenario 1 shows the positive influence of the lagged inputs in prediction daily water level; improvement in the RMSE of the BAGER is from 0.12502 m (scenario 1) to 0.1055 m (scenario 3) and in R^2 from (0.85 to 0.90).

Table 6 shows the test performance of the XGBoost models at water level prediction. The model having scenario 2 inputs, R.F.(t-1), W.L.(t-1), R.F.(t-2), W.L.(t-2), R.F.(t-3), W.L.(t-3) and R.F.(t), produces the best testing accuracy (RMSE: 0.13927 m, MAE: 0.07840 m and R^2: 0.7907); whilst the worst performance belongs to scenario 1 (RMSE: 0.14105 m, MAE: 0.07840 m and R^2: 0.7853) as is also found for the other methods. The main advantage of this method is however, its low training duration. In case of a fast decision, this method can be considered as an alternative method in the prediction of daily water levels.

Test performances of the tree regression (TR) models are illustrated in Table 7 in predicting daily water levels. Fine TR models act better than the medium and coarse TR models, with respect to the RMSE, MAE and R^2 criteria. The fine TR model with scenario 4 inputs, R.F.(t-1), W.L.(t-1), R.F.(t-2), W.L.(t-2), R.F.(t-3), W.L.(t-3), R.F.(t-4), W.L.(t-4), R.F.(t-5), W.L.(t-5), R.F.(t-6), W.L.(t-6), R.F.(t-7), W.L.(t-7) and R.F.(t), offers the best accuracy (RMSE: 0.09119 m, MAE: 0.04783 m and R^2: 0.92) while the worst accuracy is produced by the coarse TR with scenario 4 (RMSE: 0.13436 m, MAE: 0.074887 m and R^2: 0.83). The contribution of the including lagged inputs on models' accuracy is clearly seen when compared to the considered scenarios; a decrease in RMSE of fine TR is

Table 4. Performances of the support vector regression models.

MODEL 2	Performance during testing				
Support vector Regression Models	RMSE	R-Squared	MAE	Prediction speed (obs/sec)	Training time (sec)
Linear					
Scenario-1	0.14695	0.8	0.075973	7700	27.368
Scenario-2	0.1437	0.81	0.074845	37000	31.573
Scenario-3	0.14219	0.81	0.074046	38000	13.598
Scenario-4	0.14136	0.81	0.073663	40000	40.655
Quadratic					
Scenario-1	0.14321	0.81	0.07339	52000	105.24
Scenario-2	0.13761	0.82	0.070409	22000	125.4
Scenario-3	0.13649	0.82	0.069437	33000	63.884
Scenario-4	0.13501	0.83	0.068938	3700	139.42
Cubic					
Scenario-1	0.1478	0.79	0.075201	59000	356.71
Scenario-2	0.13729	0.82	0.068943	27000	548.84
Scenario-3	0.13084	0.84	0.065703	44000	313.29
Scenario-4	0.1239	0.86	0.062764	3400	270.75
Fine Gaussian					
Scenario-1	0.13839	0.82	0.067967	2200	19.438
Scenario-2	0.129	0.84	0.055497	7400	33.518
Scenario-3	0.13188	0.84	0.054506	8600	20.423
Scenario-4	0.13478	0.83	0.05508	7500	48.892
Medium Gaussian					
Scenario-1	0.14056	0.81	0.072062	20000	21.824
Scenario-2	0.13255	0.83	0.06659	10000	36.361
Scenario-3	0.12956	0.84	0.06466	14000	25.323
Scenario-4	0.12953	0.84	0.064027	1300	63.786
Coarse Gaussian					
Scenario-1	0.1441	0.8	0.073761	24000	31.934
Scenario-2	0.13982	0.82	0.072316	9300	37.765
Scenario-3	0.13894	0.82	0.071965	14000	28.271
Scenario-4	0.1387	0.82	0.071656	15000	70.67

Table 5. Performances of the ensemble regression models.

MODEL 3	Performance during testing				
Ensemble Regression Models	RMSE	R-Squared	MAE	Prediction speed (obs/sec)	Training time (sec)
Boosted Trees					
Scenario-1	0.2359	0.48	0.19848	210000	30.971
Scenario-2	0.23272	0.49	0.1956	110000	40.348
Scenario-3	0.23184	0.49	0.19521	150000	30.828
Scenario-4	0.23161	0.5	0.19499	240000	77.215
Bagged Trees					
Scenario-1	0.12502	0.85	0.068283	69000	34.639
Scenario-2	0.11099	0.88	0.059877	36000	41.7
Scenario-3	0.1055	0.9	0.05605	43000	6.0972
Scenario-4	0.10625	0.89	0.056419	67000	79.123

Table 6. Performances of the XGBoost models.

MODEL 4	Performance during testing			
XGBoost	RMSE	R-Squared	MAE	Training time (sec)
Scenario-1	0.14105	0.7853	0.07743	3.917
Scenario-2	0.13927	0.7907	0.0784	4.115
Scenario-3	0.13956	0.7898	0.07919	4.195
Scenario-4	0.13926	0.7907	0.07899	4.431

Table 7. performances of the tree regression models.

MODEL 5	Performance during testing				
Tree Regression Models	RMSE	R-Squared	MAE	Prediction speed (obs/sec)	Training time (sec)
Fine					
Scenario-1	0.11117	0.88	0.06365	360000	13.539
Scenario-2	0.097853	0.91	0.052972	260000	18.198
Scenario-3	0.093665	0.92	0.049495	400000	7.1821
Scenario-4	0.09119	0.92	0.04783	750000	1.733
Medium					
Scenario-1	0.13017	0.84	0.074142	420000	13.179
Scenario-2	0.12161	0.86	0.068699	210000	17.11
Scenario-3	0.11916	0.87	0.06699	440000	6.8155
Scenario-4	0.11889	0.87	0.066371	310000	33.961
Coarse					
Scenario-1	0.13988	0.82	0.077408	340000	12.924
Scenario-2	0.13515	0.83	0.074955	250000	16.671
Scenario-3	0.13516	0.83	0.075421	530000	6.5431
Scenario-4	0.13436	0.83	0.074887	260000	33.5

from 0.11117 m (scenario 1) to 0.09119 m (scenario 4) increase in R^2 from 0.88 to 0.92. The differences among the training duration of the fine, medium and coarse TR models are slight.

Table 8 compares the Gaussian process regression (GPR) models for the prediction of daily water levels. Among the four different GPR methods, the exponential GPR outperforms the other methods (e.g. squared exponential GPR, matern 5/2 GPR and rational quadratic GPR). The exponential GPR with the scenario 3 inputs has the lowest RMSE (0.0251116 m) and MAE (0.014062 m) and the highest R^2 (0.99), while the squared exponential GPR offered the worst accuracy (RMSE: 0.12611 m, MAE: 0.069673 m and R^2: 0.85). Involving lagged R.F. and W.L. data as inputs improved the accuracy as was also found for the other methods; decrease in the RMSE of fine exponential GPR is from 0.11907m (scenario 1) to 0.025116 m (scenario 3) and an increase in R^2 from 0.87 to 0.99. The exponential GPR with scenario 4 inputs can also be considered as another good alternative because it has much lower training duration (42.3 s).

Comparison of different methods including the LR, IR, RR, SR, SVR, ensemble regression, XGBoost, TR and the GPR models clearly indicates that the exponential GPR model acts superior to the other models. It considerably improves the accuracy in the prediction of daily

Table 8. performances of the Gaussian process regression models.

MODEL 6	Performance during testing				
Gaussian Process Regression Models	RMSE	R-Squared	MAE	Prediction speed (obs/sec)	Training time (sec)
Squared exponential					
Scenario-1	0.13444	0.83	0.072932	5000	104.78
Scenario-2	0.13108	0.84	0.071924	3100	109.48
Scenario-3	0.12611	0.85	0.069673	4800	77.144
Scenario-4	0.13133	0.84	0.072834	2600	198.47
Matern 5/2					
Scenario-1	0.14012	0.82	0.075046	4000	127.88
Scenario-2	0.1782	0.85	0.070332	1800	120.48
Scenario-3	0.12022	0.86	0.066661	2200	109.89
Scenario-4	0.12793	0.85	0.071165	1000	244.03
Exponential					
Scenario-1	0.11907	0.87	0.06611	5700	171.78
Scenario-2	0.080128	0.94	0.045133	3100	179.86
Scenario-3	**0.025116**	**0.99**	**0.014062**	**4700**	**127.14**
Scenario-4	0.054888	0.97	0.031103	4300	42.301
Rational quadratic					
Scenario-1	0.13972	0.82	0.074956	3500	197.74
Scenario-2	0.13055	0.84	0.071716	2400	217.15
Scenario-3	0.09	0.92	0.049998	2400	172.2
Scenario-4	0.12558	0.85	0.069925	3500	375.58

water level; improvement in the RMSE accuracy is by 454%, 431%, 483%, 436%, 463%, 438%, 393%, 414%, 416%, 452%, 822%, 320%, 455%, 263%, 373%, 435%, 402%, 379%, 258% compared to the best LR (scenario 4), IR (scenario 4), RR (scenario 4), SR (scenario 4), linear based SVR (scenario 4), quadratic based SVR (scenario 4), cubic based SVR (scenario 4), fine Gaussian based SVR (scenario 2), medium Gaussian based SVR (scenario 4), coarse Gaussian based SVR (scenario 4), BOOSTER (scenario 4), BAGER (scenario 3), XGBoost (scenario 2), fine TR (scenario 4), medium TR (scenario 4), coarse TR (scenario 4), squared exponential GPR (scenario 3), matern 5/2 GPR (scenario 3) and the rational quadratic GPR (scenario 3) models, respectively. As is also clear from the best input scenarios, involving longer lag data as input is necessary for the better accuracy in the prediction of daily water levels.

Figure 6 illustrates the errors produced by the best models from each group (e.g. model 1: IR with scenario 4, model 2: cubic based SVR with scenario 4, model 3: BAGER with scenario 3, model 4: XGBoost with scenario 2, model 5: fine TR with scenario 4 and model 6: exponential GPR with scenario 3). The x-axis of the figure represents the daily water level. The formula for calculating the relative error percentage is shown in Equation (5).

$$E(\%) = \frac{(\text{Predicted}_{\text{Wl}} - \text{Actual}_{\text{Wl}})}{\text{Actual}_{\text{Wl}}} \times 100\% \quad (5)$$

As is clearly observed from the graphs, the exponential GPR produces the least errors for all ranges of the time series, and this shows the consistency of this method in

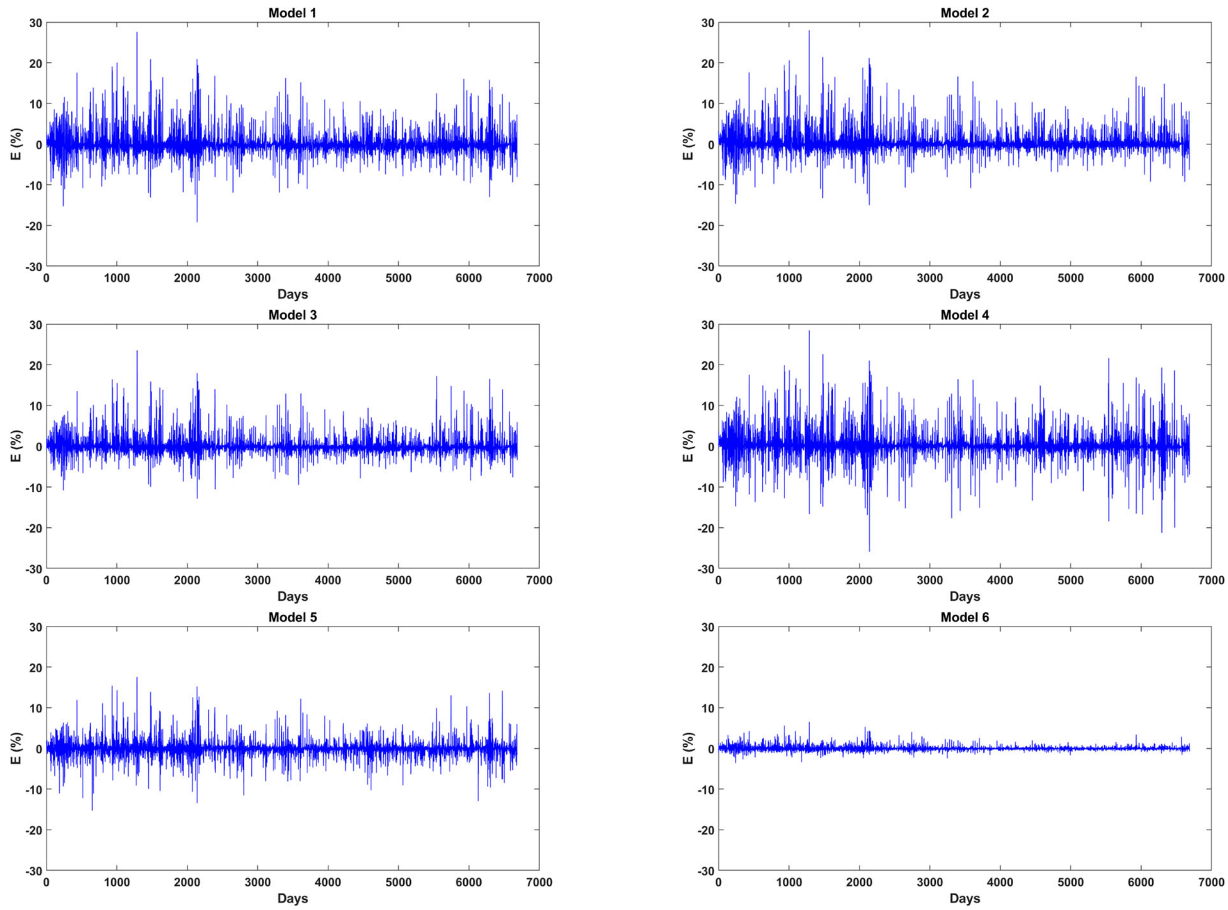


Figure 6. Errors produced by the best models from each group.

predicting daily water levels. In addition to that, Figure 7 compares the relative error, defined by Equation 6, for the best (Model-6) and worst (Model-1) models.

$$E(\%) = \frac{(\text{Predicted WL} - \text{Actual WL})}{(\text{average water depth})} \times 100\% \quad (6)$$

It could be observed from Figure 7 that the performance of the Model-1 could provide prediction value for water level with accuracy ranging between −12% and +40%. On the other hand, while examining the performance of Model-6, it could be depicted that the prediction accuracy has been significantly enhanced through the reduced range of the error to be between −4% and +10%. In addition, in general, it could be noticed that both models Model-1 and Model-6 could provide better prediction accuracy for the recent data (as appearing at the tail-end of the figure), while the accuracy were relatively low for the old data (as appearing at the beginning of the figure). This is due to the fact that the water level records at the beginning were highly fluctuating, and hence, complex and difficult for the model to predict with high accuracy. At the other end, when the water level became more stable in recent years, the prediction accuracy has been improved and the errors significantly reduced. Such observations showed that, generally, both models able to detect the complexity of the water level patterns under different fluctuation conditions but within the given range of accuracy, which is relatively acceptable for river management.

Figure 8 compares the predictions of the six best models on the scatter diagrams. It is apparent that the exponential GPR provides the least scattered predictions with higher variance (99%) explained by the model. With regards to the sensitivity analysis, the exponential GPR model exhibits a high level of accuracy when input combinations of scenario 3 used. It can be concluded that, among the four investigated kernel functions, the exponential kernel exhibits the highest level of precision, a similar finding of the superiority of the exponential kernel has been reported recently by (Alghamdi et al., 2020). Since kernel choice is problem dependent, the exponential kernel function should be prioritized in dealing with the fluctuations in water level or a river. In addition, there is a need to examine that all the proposed modeling approaches can be adequately used for modeling the water level. Therefore, two statistical indices

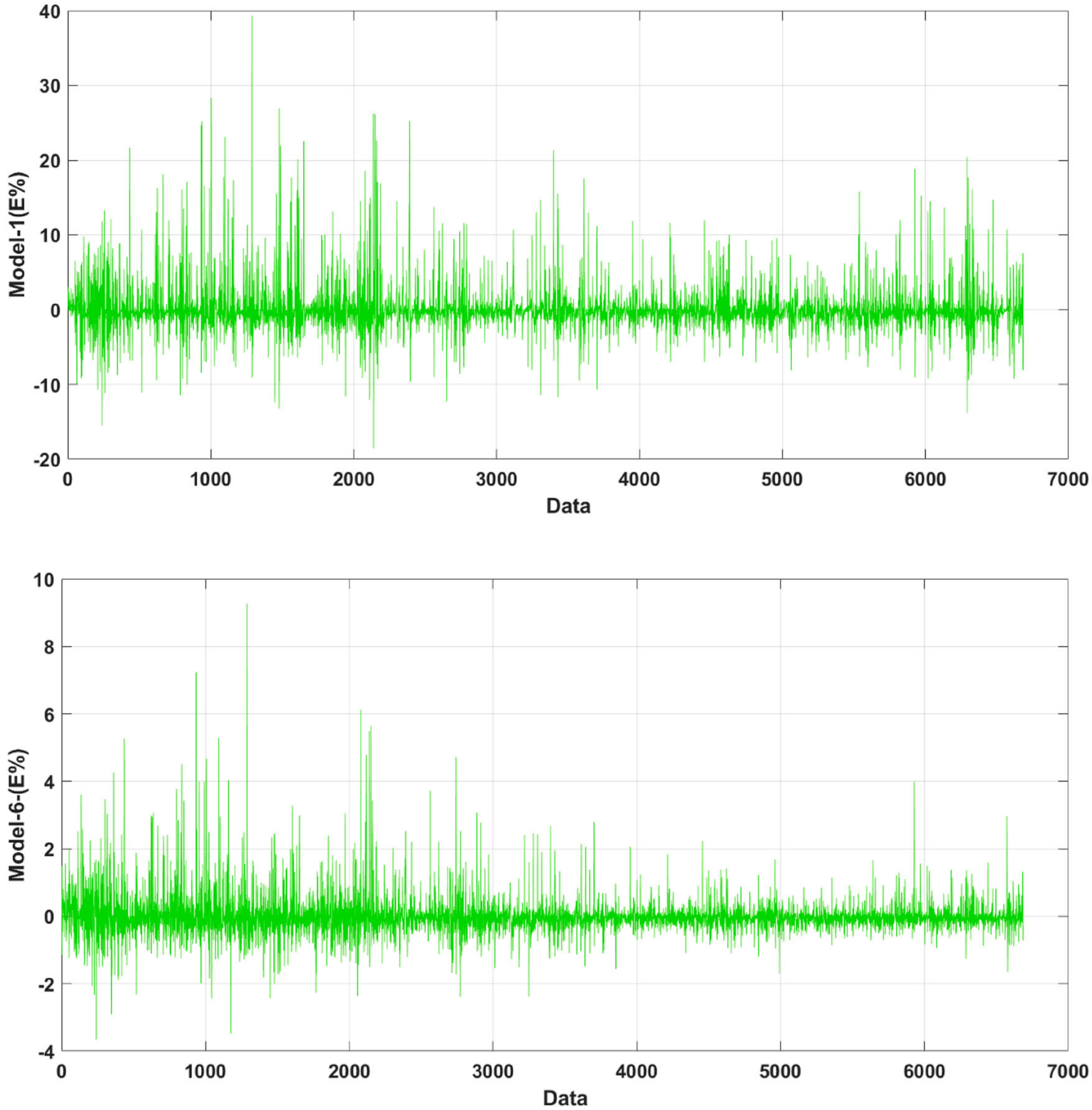


Figure 7. Relative error percentage comparison between the best model (Model-6) and the lowest accuracy model (Model-1).

namely, the mean and the standard deviation were used. It should be noticed here that if the resulting mean and the standard deviation of actual data and the model output are closed, it indirectly showed that the model mathematical procedure could successfully capture the feature of the original pattern of the data. However, it should be noted here that these two statistical indices are not sufficient to judge the superiority of a certain model over the others. Table 9 demonstrates the observed values versus the performances of each model with respect to the mean and standard deviation. As could be observed from Table 9, the mean and standard deviation values of the actual data when compared to the achieved outputs from all the proposed modeling approaches, are almost similar. Thus this reflects the success of all the models.

Table 9. the mean and SD for the actual versus predicted.

	Mean	SD
Actual	4.48777	0.30389
Model-1	4.49807	0.29754
Model-2	4.48132	0.29734
Model-3	4.49821	0.29434
Model-4	4.48463	0.31304
Model-5	4.48754	0.30868
Model-6	4.48776	0.30388

Figure 9 shows the overall performance of the exponential GPR in capturing the changes in the river's water level. It can be concluded that the model exhibits a high precision.

It could be noted from Figure 9 that the proposed Model-6 could detect the fluctuations pattern of the water level experienced along the whole historical records. Although the model input-output was developed based on the previous records of the water level, the model

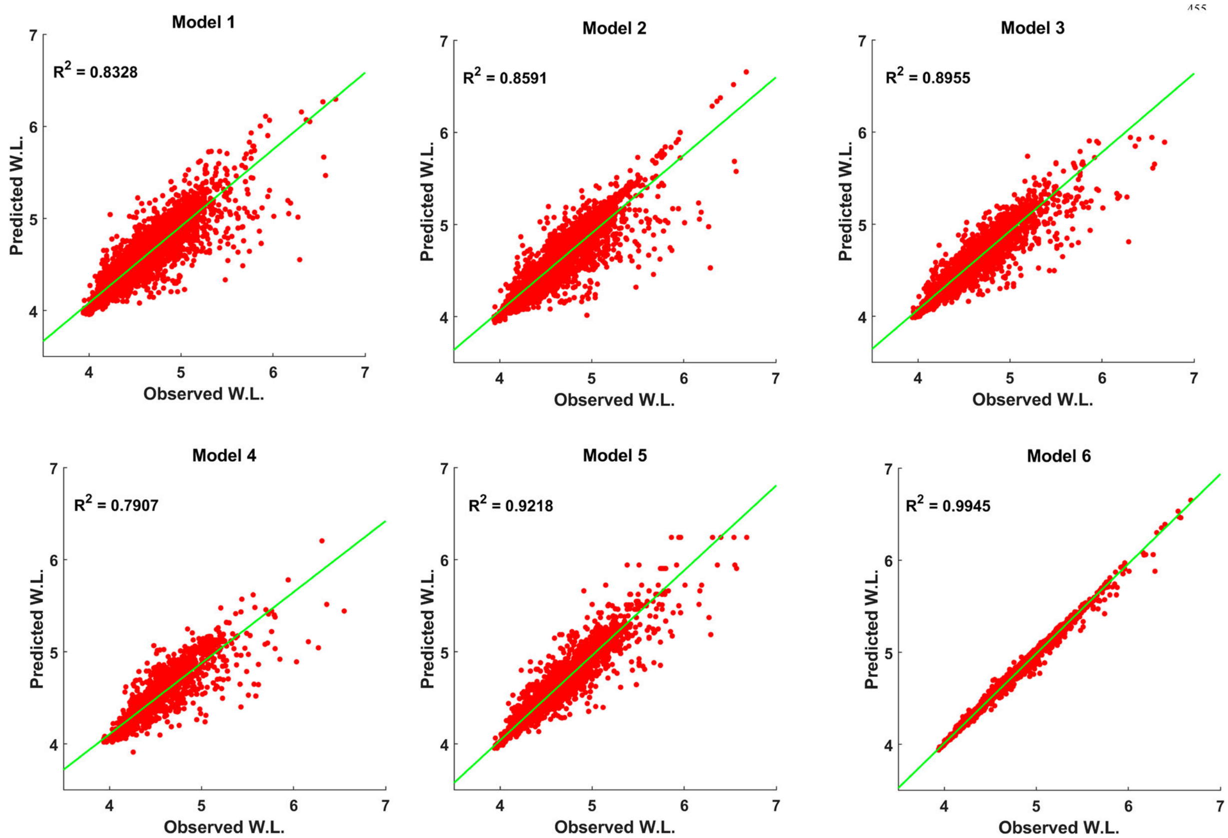


Figure 8. Scatterplots for the 6 proposed models.

could mimic the water level pattern even with absence of other vital variables that affect the water level value. Such observations shows that proposed model could be able to predict the water level fluctuation without explicit characterization and quantification of the physical properties and conditions of such phenomena. In fact, it is one of the major advantages of the data-driven modeling approach that it does not require accurate representation of governing physical laws. For further details, zoom-in Figure 9 for certain period of testing data that experienced most of possible ranges of water level, the zoomed-in sub-figure would show that the model could smoothly follow the original actual water level changes.

One of the reliable measures to prevent overfitting is the cross-validation method (Nur Adli Zakaria et al., 2021). Therefore, in this study, different folds (3,5,7 and 9) were used, as shown in Table 10. In additional, one of the advantages of GPR model is its capability in avoiding the over-fitting problem (Asante-Okyere et al., 2018). Moreover, it has been reported that, the GPR with exponential is better than other kernel functions in dealing with the over-fitting problem (Mohammed & Cawley, 2017). It has to be noted that, the problem of over-fitting usually occurs when a small dataset is used to develop a model, however in the current study the length of data used would be considered acceptable in avoiding such a problem (daily basis data from 1990 to 2019).

The next step is to use the exponential GPR model in capturing the extreme events that occurred in 10 days periods. To achieve that, the water level data were sorted based on 10 days maximum and minimum, and the results were illustrated in Figure 10. This graph also proves the success of the exponential GPR model in capturing the extremes of the water levels.

Finally, an uncertainty analysis was carried out, which is vital to measure how reliable is the developed model (the exponential GPR model) is with its prediction. Different techniques have been used to assess the uncertainty of ML algorithms (Ehteram et al., 2019; Ho et al., 2019). In this study, the 95PPU technique was computed from the predicted dataset's cumulative distribution at two levels: 2.5% and 97.5%. The D-factor is then computed by measuring the average distance between the two levels of the 95PPU. The best result is when 100% of the predicted dataset are bracketed between the two levels of 95PPU, and the d-factor is equal to zero (Jumin et al., 2020). Based on these two indicators, it can be observed that the developed model (the exponential GPR model) is capable of predicting the water level of the

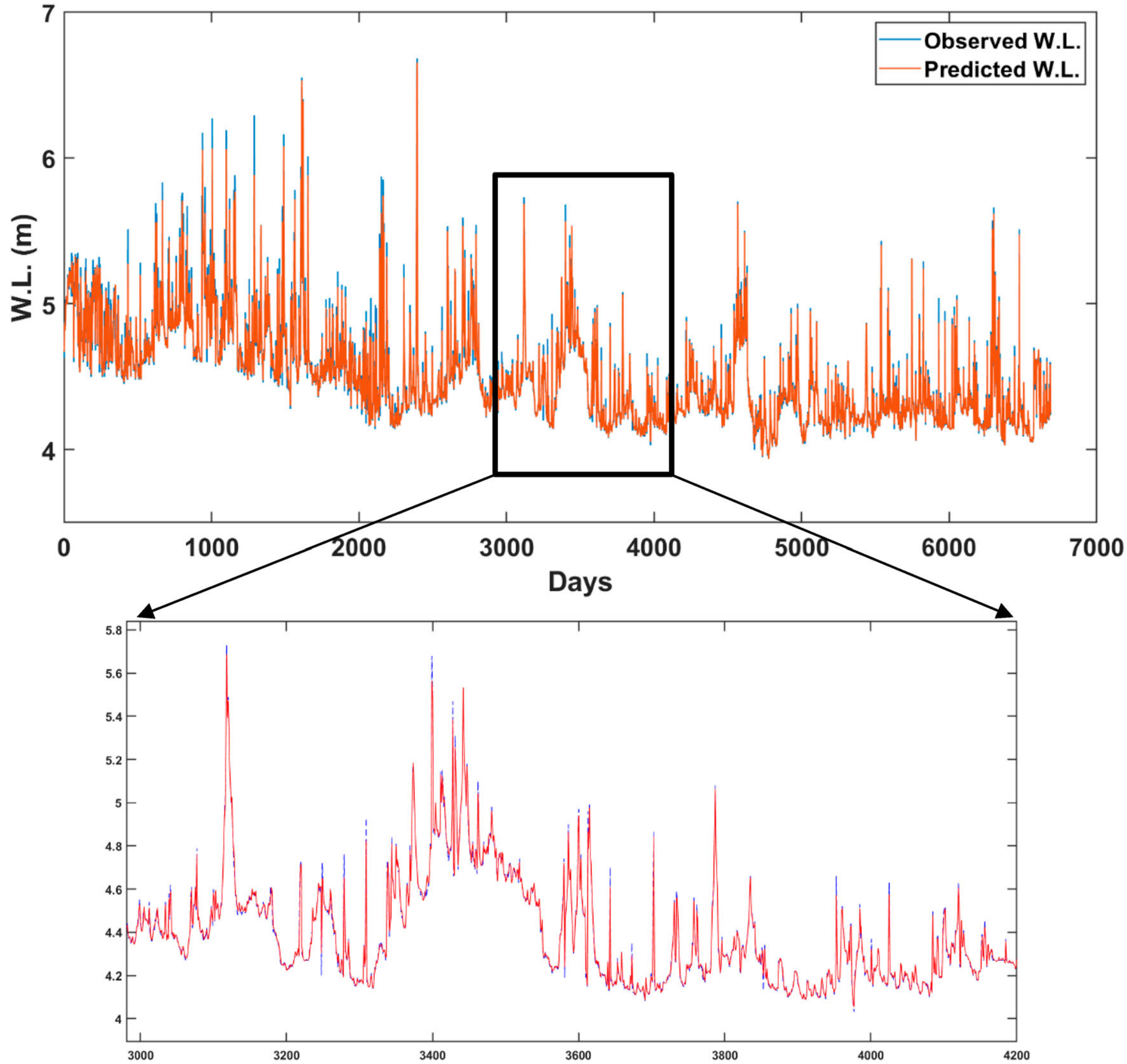


Figure 9. Actual versus output the exponential GPR model.

river with high precision and less uncertainty; where the computed 95PPU and d-factor were found to be equal to 98.27689362 and 0.000525, respectively. The performance of the developed model (the exponential GPR model) was compared with other developed models in other recent studies found from the literature source. A combined ML model with ARIMA was developed by (Phan & Nguyen, 2020) to predict hourly water level of Red River. The highest accuracy was achieved when the Random Forest was combined with ARIMA ($R = 0.98$). The developed model in our study exhibits slightly better performance ($R = 0.997$) and can be considered less complex since there is no need to combine two techniques to achieve such a level of precision. Another study by (Garcia et al., 2017) predicts the water level at the Cagayan River, in the Philippines; using the Random Forest model. The average performance with respect to r-squared was 0.95. Our study achieved better performance with the developed exponential GPR model where r-squared was equal to 0.995. It can be concluded that the developed model could be considered a reliable tool for predicting the river water level. Albeit, the developed model approach and structure had proved suitability to predict the river water level, however, it is still pertinent and necessary to examine the application of the model in different case studies with different climate conditions to assure its suitability worldwide including different climatic zones.

Table 10. Time series cross-validation.

k-fold cross-validation	R-Squared
3	0.98
5	0.96
7	0.99
9	0.98

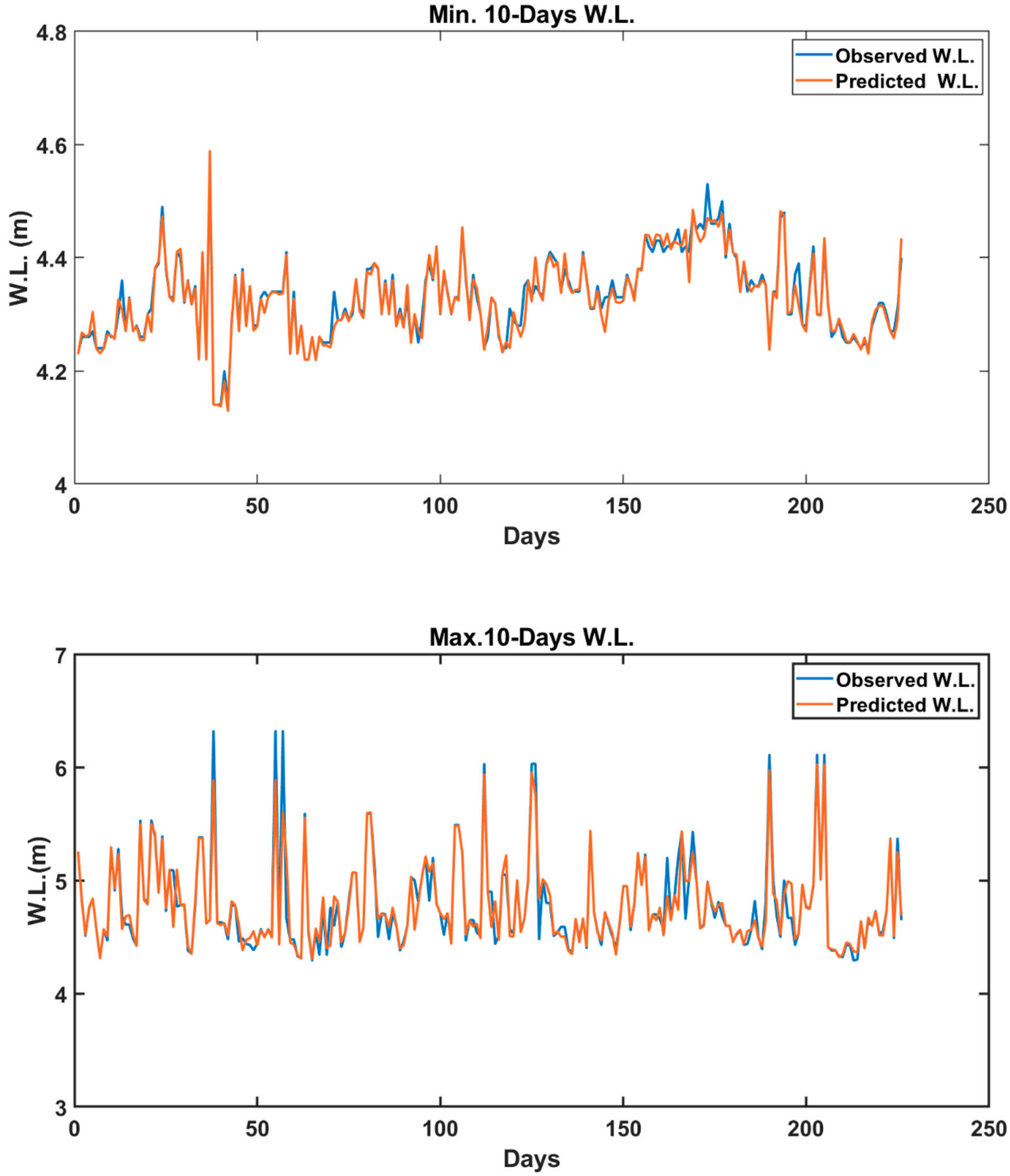


Figure 10. Model performances in predicting extreme events.

Conclusions

The capability of different machine learning models such as the linear regression (LR), interaction regression (IR), robust regression (RR), stepwise regression (SR), support vector regression (SVR), boosted trees ensemble regression (BOOSTER), bagged trees ensemble regression (BAGER), XGBoost, tree regression (TR) and Gaussian process regression (GPR), was investigated for the prediction of water level at a river in Malaysia. Four different input scenarios were investigated, considering correlation analysis. Generally, the comparison of several data-driven regression methods indicated that the exponential GPR model offered better accuracy in predicting daily water levels with respect to different assessment criteria. The findings of this study show the success of the GPR model in capturing the changes in the water level of a river; thus paving the way for which the model can be used in works to mitigate potential risk that may occur in the future from natural events. However, unfortunately as in many instances, one of the limitations in this study is data availability; albeit a year of complete data set was the only possible usage for our study. This should not be a deterrent to initiate research work. Nevertheless, future work could and should be continued by testing the Exponential Gaussian Regression model (the exponential GPR model) using data from different gauge stations within Malaysia in another region with a similar or different climate, when opportunity arises. Moreover, future research can be carried out to hybridize the GPR algorithm with optimization techniques (that are currently gathering at a fast pace) that could be investigated to improve the accuracy. For example, the augmentation of ML algorithms with meta-heuristic optimization techniques, is worthy of a detailed look-into. In addition to that, deploying the developed model using data from rivers in different countries with a similar climate like Malaysia and with different climate conditions could be explored for generalization purposes. Moreover, future work could be carried out by integrating the developed model with high quality sensors to forecast water level in the river in real-time.

Acknowledgements

This research was supported by Ministry of Education (MOE) through Fundamental Research Grant Scheme (FRGS, Grant No.: FRGS/1/2019/TK01/UNITEN/02/3). We also want to thank to the Department of Irrigation and Drainage Malaysia (JPS) for providing this study with the data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Ali Najah Ahmed http://orcid.org/0000-0002-5618-6663
Ayman Yafouz http://orcid.org/0000-0002-0932-1295
Ozgur Kisi http://orcid.org/0000-0001-7847-5872
Yuk Feng Huang http://orcid.org/0000-0001-6565-947X
Mohsen Sherif http://orcid.org/0000-0002-6368-8143
Ahmed Sefelnasr http://orcid.org/0000-0002-0281-6455
Ahmed El-Shafie http://orcid.org/0000-0001-5018-8505

References

Abdullah, S., Nasir, N. H. A., Ismail, M., Ahmed, A. N., & Jarkoni, M. N. K. (2019). Development of Ozone prediction model in urban area. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 2263–2267. https://doi.org/10.35940/ijitee.J1127.0881019

Ahmed, A. N., Van Lam, T., Hung, N. D., Van Thieu, N., Kisi, O., & El-Shafie, A. (2021). A comprehensive comparison of recent developed meta-heuristic algorithms for streamflow time series forecasting problem. *Applied Soft Computing*, *105*, 107282. https://doi.org/10.1016/j.asoc.2021.107282

Alghamdi, A. S., Polat, K., Alghoson, A., Alshdadi, A. A., & Abd El-Latif, A. A. (2020). Gaussian process regression (GPR) based non-invasive continuous blood pressure prediction method from cuff oscillometric signals. *Applied Acoustics*, *164*, 107256. https://doi.org/10.1016/j.apacoust.2020.107256

Aljanabi, M., Shkoukani, M., & Hijjawi, M. (2020). Ground-level Ozone prediction using machine learning techniques: A case study in Amman, Jordan. *International Journal of Automation and Computing*, *17*(5), 667–677. https://doi.org/10.1007/s11633-020-1233-4

Allawi, M. F., Othman, F. B., Afan, H. A., Ahmed, A. N., Hossain, M. S., Fai, C. M., & El-Shafie, A. (2019). Reservoir evaporation prediction modeling based on artificial intelligence methods. *Water (Switzerland)*, *11*, 6. https://doi.org/10.3390/w11061226

Arbain, S., & Wibowo, A. (2012). Time series methods for water level forecasting of Dungun river in Terengganu Malaysia. *International Journal of Engineering Science and Technology*, *4*(4), 1803–1811.

Asante-Okyere, S., Shen, C., Ziggah, Y. Y., Rulegeya, M. M., & Zhu, X. (2018). Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability. *Energies*, *11*(12), 3261. https://doi.org/10.3390/EN11123261

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., 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. *Applied Sciences*, *10*(2), 571. https://doi.org/10.3390/app10020571

Banadkooki, F. B., Ehteram, M., Ahmed, A. N., Fai, C. M., Afan, H. A., Ridwam, W. M., Sefelnasr, A., & El-Shafie, A. (2019). Precipitation forecasting using multilayer neuralNetwork and support vector machine optimization based on flow regime algorithm taking intoAccount uncertainties of soft computing models. *Sustainability (Switzerland)*, *11*(23), 6681. https://doi.org/10.3390/su11236681

Banadkooki, F. B., Ehteram, M., Ahmed, A. N., Teo, F. Y., Ebrahimi, M., Fai, C. M., Huang, Y. F., & El-Shafie, A. (2020). Suspended sediment load prediction using artificial neural network and ant lion optimization algorithm. *Environmental Science and Pollution Research*, *27*(30), https://doi.org/10.1007/s11356-020-09876-w

Band, S. S., Heggy, E., Bateni, S. M., Karami, H., Rabiee, M., Samadianfard, S., Chau, K.-W., & Mosavi, A. (2021). Groundwater level prediction in arid areas using wavelet analysis and Gaussian process regression. *Engineering Applications of Computational Fluid Mechanics*, *15*(1), 1147–1158. https://doi.org/10.1080/19942060.2021.1944913

Carmo, J. S. A. d. (2020). *Physical modelling vs. numerical modelling: Complementarity and learning*. https://doi.org/10.20944/PREPRINTS202007.0753.V1

Chong, K. L., Lai, S. H., Ahmed, A. N., Zaafar, W. Z. W., Rao, R. V., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2021). Review on dam and reservoir optimal operation for irrigation and hydropower energy generation utilizing meta-heuristic algorithms. *IEEE Access*, *9*, 19488–19505. https://doi.org/10.1109/access.2021.3054424

Dashti Latif, S., Najah Ahmed, A., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2020). Reservoir water balance simulation model utilizing machine learning algorithm. *Alexandria Engineering Journal*, https://doi.org/10.1016/j.aej.2020.10.057

Ehteram, M., Ahmed, A. N., Ling, L., Fai, C. M., Latif, S. D., Afan, H. A., Banadkooki, F. B., & El-Shafie, A. (2020a). Pipeline scour rates prediction-based model Utilizing a multilayer perceptron-colliding body algorithm. *Water*, *12*(3), 902. https://doi.org/10.3390/w12030902

Ehteram, M., Ghotbi, S., Kisi, O., Ahmed, N., Hayder, G., Fai, M., Krishnan, M., Afan, A., & EL-Shafie, A. (2019). Investigation on the potential to integrate different artificial intelligence models with metaheuristic algorithms for improving river suspended sediment predictions. *Applied Sciences*, *9*(19), 4149. https://doi.org/10.3390/app9194149

Ehteram, M., Othman, F. B., Yaseen, Z. M., Afan, H. A., Allawi, M. F., Malek, M. B. A., Ahmed, A. N., Shahid, S., Singh, V. P., & El-Shafie, A. (2018). Improving the Muskingum flood routing method using a hybrid of particle swarm optimization and bat algorithm. *Water (Switzerland)*, *10*(6), 1–21. https://doi.org/10.3390/w10060807

Ehteram, M., Yenn Teo, F., Najah Ahmed, A., Dashti Latif, S., Feng Huang, Y., Abozweita, O., Al-Ansari, N., & El-Shafie, A. (2020b). Performance improvement for infiltration rate prediction using hybridized adaptive neuro-fuzzy inferences system (ANFIS) with optimization algorithms. *Ain Shams Engineering Journal*, https://doi.org/10.1016/j.asej.2020.08.019

El-Shafie, A., Mukhlisin, M., Najah, A. A. A., & Taha, M. R. R. (2011). Performance of artificial neural network and regression techniques for rainfall-runoff prediction. *International Journal of Physical Sciences*, *6*(8), 1997–2003. https://doi.org/10.5897/IJPS11.314

El-Shafie, A., Najah, A., Alsulami, H. M., & Jahanbani, H. (2014). Optimized neural network prediction model for potential evapotranspiration utilizing ensemble procedure. *Water Resources Management*, *28*(4), 947–967. https://doi.org/10.1007/s11269-014-0526-1

Fotovatikhah, F., Herrera, M., Shamshirband, S., Chau, K., Faizollahzadeh Ardabili, S., & Piran, M. J. (2018). Survey of computational intelligence as basis to big flood management: Challenges, research directions and future work. *Engineering Applications of Computational Fluid Mechanics*, *12*(1), 411–437. https://doi.org/10.1080/19942060.2018.1448896

Garcia, F. C. C., Retamar, A. E., & Javier, J. C. (2017). Development of a predictive model for on-demand remote river level nowcasting: Case study in Cagayan River Basin, Philippines. In *IEEE region 10 Annual International Conference, Proceedings/TENCON* (pp. 3275–3279). https://doi.org/10.1109/TENCON.2016.7848657

Ghazvinian, H., Mousavi, S.-F., Karami, H., Farzin, S., Ehteram, M., Hossain, M. S., Fai, C. M., Hashim, H. B., Singh, V. P., Ros, F. C., Ahmed, A. N., Afan, H. A., Lai, S. H., & El-Shafie, A. (2019). Integrated support vector regression and an improved particle swarm optimization-based model for solar radiation prediction. *PLoS ONE*, *14*(5), e0217634. https://doi.org/10.1371/journal.pone.0217634

Guan, M., Wright, N. G., & Sleigh, P. A. (2013). A robust 2D shallow water model for solving flow over complex topography using homogenous flux method. *International Journal for Numerical Methods in Fluids*, *73*(3), 225–249. https://doi.org/10.1002/fld.3795

Hipni, A., El-shafie, A., Najah, A., Karim, O. A., Hussain, A., & Mukhlisin, M. (2013). Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). *Water Resources Management*, *27*(10), 3803–3823. https://doi.org/10.1007/s11269-013-0382-4

Ho, J. Y., Afan, H. A., El-Shafie, A. A. H., Koting, S. B., Mohd, N. S., Jaafar, W. Z. B., Lai Sai, H., Malek, M. A., Ahmed, A. N., Mohtar, W. H. M. W., Elshorbagy, A., & El-Shafie, A. A. H. (2019). Towards a time and cost effective approach to water quality index class prediction. *Journal of Hydrology*, *575*(October 2018), 148–165. https://doi.org/10.1016/j.jhydrol.2019.05.016

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, https://doi.org/10.1016/j.asej.2020.11.011

Jumin, E., Zaini, N., Ahmed, A. N., Abdullah, S., Ismail, M., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2020). Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 713–725. https://doi.org/10.1080/19942060.2020.1758792

Kamel, A. H., Afan, H. A., Sherif, M., Ahmed, A. N., & El-Shafie, A. (2021). RBFNN versus GRNN modeling approach for sub-surface evaporation rate prediction in arid region. *Sustainable Computing: Informatics and Systems*, *30*, 100514. https://doi.org/10.1016/j.suscom.2021.100514

Khan, F. A., Ali Khan, T. M., Ahmed, A. N., Afan, H. A., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2020). Complex extreme sea levels prediction analysis: Karachi coast case study. *Entropy*, *22*(5), 549. https://doi.org/10.3390/E22050549

Kumar, R., Singh, M. P., Roy, B., & Shahid, A. H. (2021). A comparative assessment of metaheuristic optimized extreme learning machine and deep neural network in multi-step-ahead long-term rainfall prediction for all-Indian regions. *Water Resources Management*, *35*(6), 1927–1960. https://doi.org/10.1007/S11269-021-02822-6

Lai, V., Ahmed, A. N., Malek, M. A., Abdulmohsin Afan, H., Ibrahim, R. K., El-Shafie, A., & El-Shafie, A. (2019). Modeling the nonlinearity of sea level oscillations in the Malaysian coastal areas using machine learning algorithms. *Sustainability*, *11*(17), 4643. https://doi.org/10.3390/su11174643

Lio, W. H., Li, A., & Meng, F. (2021). Real-time rotor effective wind speed estimation using Gaussian process regression and Kalman filtering. *Renewable Energy*, *169*, 670–686. https://doi.org/10.1016/j.renene.2021.01.040

Melville, S. (2014). Xgboost: Extreme gradient boosting. *R Lecture*, *2016*, 1–84. https://doi.org/10.1145/2939672.2939785

Mohamad, M. F., Samion, M. K. H., & Hamzah, S. B. (2014). Physical modelling for flood evaluation of Selangor river under tidal influence. *Power*, *1050*, 1T–3.

Mohammadi, B., Linh, N. T. T., Pham, Q. B., Ahmed, A. N., Vojteková, J., Guan, Y., Abba, S. I., & El-Shafie, A. (2020). Adaptive neuro-fuzzy inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series. *Hydrological Sciences Journal*, *0*(0), 1. https://doi.org/10.1080/02626667.2020.1758703

Mohammed, R. O., & Cawley, G. C. (2017). Over-fitting in model selection with Gaussian process regression. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 192–205). https://doi.org/10.1007/978-3-319-62416-7_14

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood prediction using machine learning models: Literature review. *Water*, *10*(11), 1536. https://doi.org/10.3390/w10111536

Najah, A., El-Shafie, A., & Karim, O. (2011). *Prediction of water quality parameters using artificial intelligence: Case study-Johor River Basin*. LAP Lambert Academic Publishing.

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. *Hydrology and Earth System Sciences*, *15*(8), 2693–2708. https://doi.org/10.5194/hess-15-2693-2011

Najah, A., El-Shafie, A. H., Karim, O. A., Jaafar, O., & El-Shafie, A. H. (2011). An application of different artificial intelligences techniques for water quality prediction. *International Journal of Physical Sciences*, *6*(22), 5298–5308. https://doi.org/10.5897/IJPS11.1180

Najah, A., Lam, T. V., Duy, N., Thieu, N. V., & Kisi, O. (2021). A comprehensive comparison of recent developed meta-heuristic algorithms for streamflow time series forecasting problem. *Applied Soft Computing*, *105*, 107282. https://doi.org/10.1016/j.asoc.2021.107282

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. *Journal of Hydrology*, *578*, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Najah Ahmed, A., El-Shafie, A. A., Karim, O. A., & El-Shafie, A. A. (2012). An augmented wavelet de-noising technique with neuro-fuzzy inference system for water quality prediction. *International Journal of Innovative Computing, Information and Control*, *8*(10 B), 7055–7082.

Napi, N. N. L. M., Abdullah, S., Mansor, A. A., Ahmed, A. N., Mansor, W. N. W., & Ismail, M. (2021). Input variable selection for hourly ozone (O_3) concentration prediction in urban area. *ARPN Journal of Engineering and Applied Sciences*, *16*(2), 157–164.

Napi, N. N. L. M., Mohamed, M. S. N., Abdullah, S., Mansor, A. A., Ahmed, A. N., & Ismail, M. (2020). Multiple linear regression (MLR) and principal component regression (PCR) for Ozone (O_3) concentrations prediction. *IOP Conference Series: Earth and Environmental Science*, *616*(1), 012004. https://doi.org/10.1088/1755-1315/616/1/012004

Nur Adli Zakaria, M., Abdul Malek, M., Zolkepli, M., & Najah Ahmed, A. (2021). Application of artificial intelligence algorithms for hourly river level forecast: A case study of Muda River, Malaysia. *Alexandria Engineering Journal*, *60*(4), 4015–4028. https://doi.org/10.1016/j.aej.2021.02.046

Olivia Muslim, T., Najah Ahmed, A., Malek, M. A., El-Shafie, A., & EL-Shafie, A. (2018). Investigating the impact of wind on sea level rise using multilayer perceptron neural network (MLP-NN) at coastal area, Sabah. *International Journal of Civil Engineering and Technology*, *9*(12), 646–656.

Osman, A., Afan, H. A., Allawi, M. F., Jaafar, O., Noureldin, A., Hamzah, F. M., Ahmed, A. N., & El-shafie, A. (2020). Adaptive fast orthogonal search (FOS) algorithm for forecasting streamflow. *Journal of Hydrology*, *586*(February), 124896. https://doi.org/10.1016/j.jhydrol.2020.124896

Pham, Q. B., Afan, H. A., Mohammadi, B., Ahmed, A. N., Linh, N. T. T., Vo, N. D., Moazenzadeh, R., Yu, P. S., & El-Shafie, A. (2020). Hybrid model to improve the river streamflow forecasting utilizing multi-layer perceptron-based intelligent water drop optimization algorithm. *Soft Computing*, *24*(23), 18039–18056. https://doi.org/10.1007/s00500-020-05058-5

Phan, T. T. H., & Nguyen, X. H. (2020). Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. *Advances in Water Resources*, *142*, 103656. https://doi.org/10.1016/j.advwatres.2020.103656

Rasmussen, C. E. (2004). Gaussian Processes in machine learning. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3176, 63–71. https://doi.org/10.1007/978-3-540-28650-9_4

Rezaie-Balf, M., Attar, N. F., Mohammadzadeh, A., Murti, M. A., Ahmed, A. N., Fai, C. M., Nabipour, N., Alaghmand, S., & El-Shafie, A. (2020). Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: Comparative assessment of a noise suppression hybridization approach. *Journal of Cleaner Production*, *271*, 122576. https://doi.org/10.1016/j.jclepro.2020.122576

Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. (2020). Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. *Ain Shams Engineering Journal*, https://doi.org/10.1016/j.asej.2020.09.011

Rosli, N., Ambak, K., Daniel, B. D., Prasetijo, J., Tun, U., Onn, H., & Pahat, B. (2015). *Floodplains modelling due to dam break*. *1*, 1–6.

Sami, B. H. Z., Sami, B. F. Z., Fai, C. M., Essam, Y., Ahmed, A. N., & El-Shafie, A. (2021). Investigating the reliability of machine learning algorithms as a sustainable tool for total suspended solid prediction. *Ain Shams Engineering Journal*.

Sapitang, M., Ridwan, W. M., Faizal Kushiar, K., Najah Ahmed, A., & El-Shafie, A. (2020). Machine learning Application in reservoir water level forecasting for sustainable hydropower generation strategy. *Sustainability*, *12*(15), 6121. https://doi.org/10.3390/su12156121

Štefelová, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P., & Hron, K. (2021). Robust regression with compositional covariates including cellwise outliers. *Advances in Data Analysis and Classification*, 1–41. https://doi.org/10.1007/s11634-021-00436-9

Sun, A. Y., Wang, D., & Xu, X. (2014). Monthly streamflow forecasting using Gaussian process regression. *Journal of Hydrology*, *511*, 72–81. https://doi.org/10.1016/j.jhydrol.2014.01.023

Teo, A. N. F. Y., Huang, M. F. C. Y. F., Abdullah, S. D. L. S., & Shafie, A. E. (2021). Surface water quality status and prediction during movement control operation order under COVID – 19 pandemic: Case studies in Malaysia. *International Journal of Environmental Science and Technology*, 0123456789. https://doi.org/10.1007/s13762-021-03139-y

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A. N., Sammen, S. S., Kisi, O., Huang, Y. F., & El-Shafie, A. (2020). Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. *Journal of Hydrology*, *589*, 125133. https://doi.org/10.1016/j.jhydrol.2020.125133

Tikhamarine, Y., Souag-Gamane, D., Najah Ahmed, A., Kisi, O., & El-Shafie, A. (2020). Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey wolf optimization (GWO) algorithm. *Journal of Hydrology*, *582*(October 2019), 124435. https://doi.org/10.1016/j.jhydrol.2019.124435

Wilkie, D., & Galasso, C. (2021). Gaussian process regression for fatigue reliability analysis of offshore wind turbines. *Structural Safety*, *88*, 102020. https://doi.org/10.1016/j.strusafe.2020.102020

Wu, X. L., Xiang, X. H., Li, L., & Wang, C. H. (2014). Water level updating model for flow calculation of river networks. *Water Science and Engineering*, *7*(1), 60–69. https://doi.org/10.3882/j.issn.1674-2370.2014.01.007

Yafouz, A., Ahmed, A. N., Zaini, N., & El-Shafie, A. (2021). Ozone concentration forecasting based on artificial intelligence techniques: A systematic review. *Water, Air, & Soil Pollution*, *232*(2), 79. https://doi.org/10.1007/s11270-021-04989-5

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K. W. (2019). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*, *569*, 387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069

Zhang, Y., & Xu, X. (2021). Lattice misfit predictions via the Gaussian process regression for Ni-based single crystal superalloys. *Metals and Materials International*, *27*(2), 235–253. https://doi.org/10.1007/s12540-020-00883-7