PREDICTING THE RESERVOIR INFLOW OF BHUMIBOL DAM USING XGBOOST MACHINE LEARNING ALGORITHM

PHEERANAT DORNPUNYA

THA 2022 INTERNATIONAL CONFERENCE ON MOVING TOWARDS SUSTAINABLE WATER AND CLIMATE CHANGE MANAGEMENT AFTER COVID-19

26-28 JANUARY 2022

ADVISOR: ASSOC. PROF. DR. AREEYA RITTIMA

DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING, FACULTY OF ENGINEERING

MAHIDOL UNIVERSITY

Email: pheeranat.official@gmail.com

Statement of the problems and objectives

Figure 3 Bhumibol Dam, Tak of Thailand

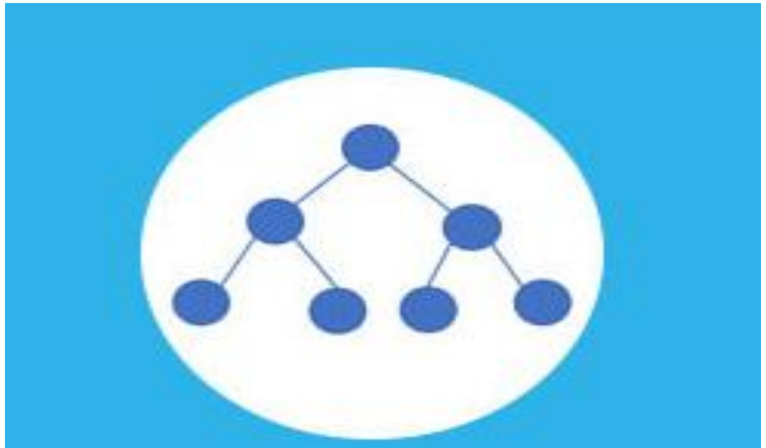


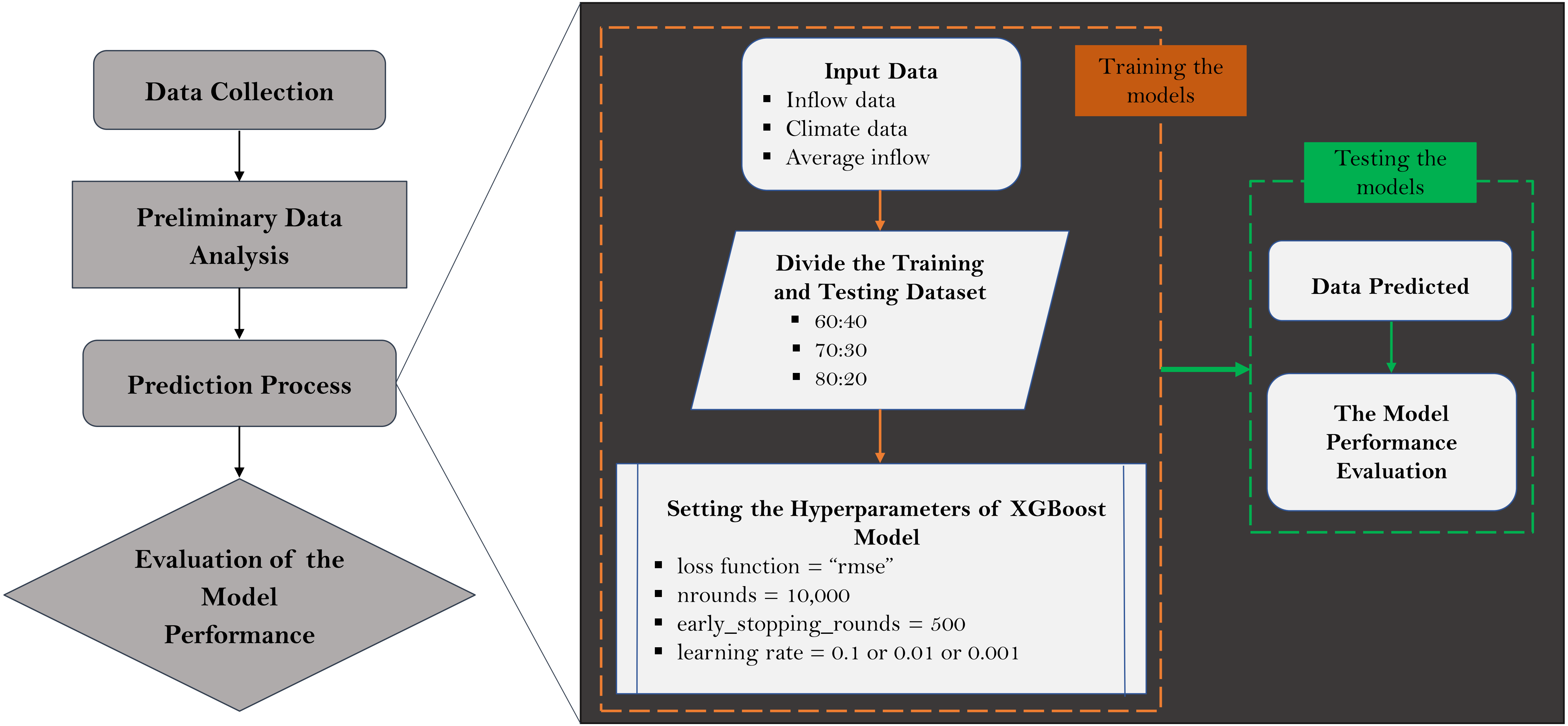
Figure 4 Schematic of XGBoost trees

- Thailand faces hydrological problems every year that affect agricultural, community, and industrial areas, especially in the central and northern regions. It also affects the water management of major dams in the country, such as the Bhumibol Dam.

- This has significant implications for the operational actors to revise the strategic plan based upon the data-driven decision-support tools to reduce disaster risks and losses. The accurate and reliable hydrological prediction plays vital role in the decision-making process specifically for real time operation of dam-reservoir system. Machine Learning (ML) which is the advanced area of Artificial Intelligence (AI), has been extensively used to improve predictive accuracy and understand hydrological uncertainty and provide the multiple lead times. It has proved a great success in predicting hydrological data such as rainfall, reservoir inflow, and river flow, etc.

- Therefore, to solve the problems, this research aims to develop the prediction models of the reservoir inflow of the Bhumibol Dam using XGBoost algorithm, which is a Machine Learning (ML) technique, and compare the performance of daily and monthly prediction models.

Methodology



Methodology

Data Collection

Table 1 Data collection for this study

Data Category	Source
Reservoir Data	
▪ Reservoir inflow data	EGAT[1/]
Hydrological and Climate Data	
▪ Climate & Rainfall data	TMD[2/] /Web-Based Data Sources[3/]

Table 2 The climate station sites nearby the Bhumibol Dam

Code	Weather Observing Station Name	Geography Coordinate	
		Latitude	Longitude
0002	Tak	16.880000	99.140000
0006	Bhumibol Dam	17.243611	99.002222
0007	Mea Sot	16.700000	98.541944
0015	Si Samrong	17.486389	99.526667
0017	Doi Musir	16.700000	98.935278
0019	Thoen	17.636667	99.245556

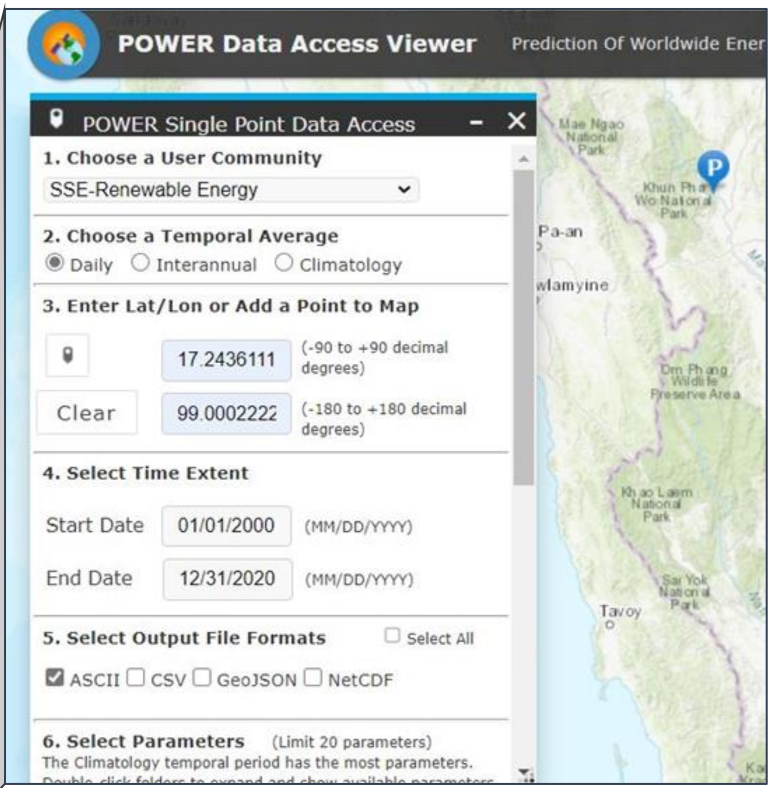


Figure 5 Web base data source of climate data by NASA

The climate data was obtained from the National Aeronautics and Space Administration (NASA) and was based on the same geographic coordinates as TMD climate stations located around the reservoir site.

Note; [1/]EGAT = Electricity Generating Authority of Thailand

[2/]TMD = Thai Meteorological Department

[3/]https://power.larc.nasa.gov/data-access-viewer/

Methodology

Data Collection

Preliminary Data Analysis

Note;

1/AH = Average Air Humidity (%)

2/AP = Average Air Pressure (hPa)

3/AT = Average Temperature (°C)

4/Prec. = Precipitation (mm/day).

Basic Statistic Analysis

- The Bhumibol Dam has the reservoir capacity of 9,662 MCM covering drainage area of 26,386 km². The basic statistics of climate data and reservoir inflow of the Bhumibol Dam collected from 2000–2020 (21 years) are summarized in Table 3.

Table 3 Descriptive statistics of climate and reservoir inflow data in the study area

Required data	Values	Time of Occurrence
Max. daily prec.	95.48	03/05/2001
Max. monthly prec.	382.72	05/2007
Max. daily evap.	37.70	17/03/2008
Max. monthly evap.	137.45	05/2013
Peak daily inflow	311.46	03/10/2009
Peak monthly inflow	2,990.21	09/2002
Avg. daily inflow	14.90	
Avg. monthly inflow	453.67	

Correlation Analysis

- The purpose of the correlation analysis was to evaluate the relationship between meteorological data and reservoir inflows utilizing daily data from 2000 to 2020. Climate variables included AH1/, AP2/, AT3/, and Prec.4/

Table 4 The correlation coefficients between the observed reservoir inflow and climate data

Data Source	Station Code	AH	AP	AT	Prec.
NASA	0002-Tak	0.5210	-01484	-0.1897	0.3648
	0006-Bhumibol Dam	0.5182	-0.1458	-0.1730	**0.3693**
	0007-Mea Sot	0.4909	-0.1787	-0.0757	0.3603
	0015-Si Samrong	0.5205	-0.1389	-0.1780	0.3628
	0017-Doi Musir	0.4909	-0.4787	-0.0757	0.3603
	0019-Thoen	0.5049	-0.1463	-0.1327	0.3550
TMD	0002-Tak	0.4015	-0.1167	-0.1145	0.2840
	0006-Bhumibol Dam	0.4016	-0.0073	-0.1032	0.2886
	0007-Mea Sot	0.4010	-0.1643	-0.0957	0.1966
	0015-Si Samrong	0.3185	-0.0208	-0.0266	0.1621
	0017-Doi Musir	0.2116	0.0028	0.0059	0.0341
	0019-Thoen	0.4604	-0.0896	-0.0911	0.1913

Methodology

Data Collection

Preliminary Data Analysis

Prediction Process

Extreme Gradient Boosting (XGBoost) Algorithm

To develop the daily and monthly prediction models of reservoir inflow of the Bhumibol Dam, the Extreme Gradient Boosting (XGBoost) which is a decision–tree–based ensemble machine learning algorithm, was used in this study.

The objective function measuring how well the model is suited with the training data, should be defined. In general, a characteristic of objective functions contains two main terms; (1) training loss function and (2) regularization term as expressed in Eq. (1)

$$\mathbf{Obj(\theta) = L(\theta) + \Omega\,(\theta)} \qquad .....(1)$$

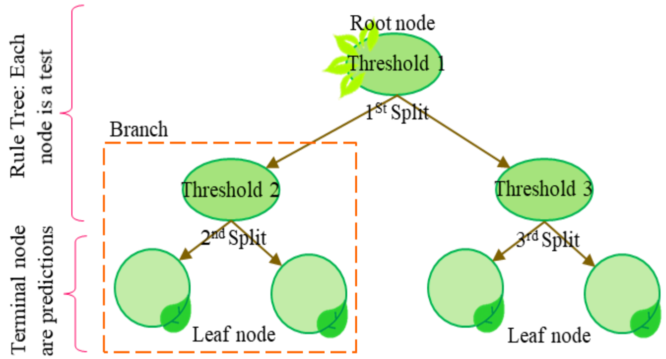


Figure 6 The decision tree components of the XGBoost

$L(\theta)$ is the training loss function which can be categorized into two types; classification and regression losses. A common type of regression loss is mean squared error as given in Eq. (2).

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - p_i)^2 \qquad .....(2)$$

The regularization term $\Omega(\theta)$ in Eq. (3) is one of the significant term that helps control the complexity of the model and avoid overfitting.

$$\Omega\,(\theta) = \gamma T + \frac{1}{2}\lambda\sum_{i=1}^{T} O_{value}^{2} \qquad .....(3)$$

Methodology

Data Collection

Preliminary Data Analysis

Prediction Process

$$\mathrm{Sim} = \frac{\sum_{i=1}^{n}(y_i - p_i)^2}{n + \lambda} \quad \text{.....(4)}$$

The loss function $L(\theta)$ indicates the scores of the tree and leaf. It is intractable to learn all the trees at once. Instead, we use an additive strategy: fix what we have learned and add one new tree at a time. Similarity score (Sim) is computed to indicate a score of each node by using Eq. (4).

$$\text{Gain value} = \mathrm{Sim}_{left} + \mathrm{Sim}_{right} + \mathrm{Sim}_{root} \quad \text{.....(5)}$$

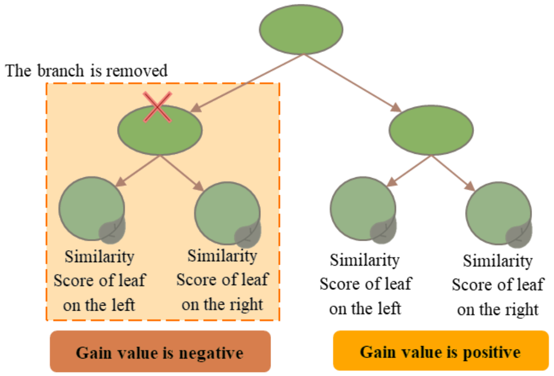


Figure 7 Steps to split the decision tree using Gain value

The Gain value is calculated to measure how good a tree structure is. The Gain value indicates whether a tree can split the leaves or not. When the gain values are negative, the branch is removed as shown in Fig.7.

$$O_{value} = \frac{\sum_{i=1}^{n}(y_i - p_i)}{n + \lambda} \quad \text{.....(6)}$$

The output values (O_{value}) are calculated by Eq. (6) for all leaves to get the final tree at the end of first model since some leaf has more than one residual.

$$p_i^t = p_i^0 + \varepsilon\left[\sum_{i=1}^{n} L\left(y_i, p_i^0 + O_{value}\right) + \frac{1}{2}\lambda O_{value}^2\right] \quad \text{.....(7)}$$

The final prediction is the additive sum of the initial predicted value (p_i^o) and objective function combining with loss function and a regularization term, as shown in Eq. (7).

Methodology

Data Collection

Preliminary Data Analysis

Prediction Process

Model Parameters Setting

Setting the model structures were performed corresponding to the model input variables selected: *climate and observed inflow* data at time step t (Inflow/Precipitation/Humidity), the *ratio of training–testing* dataset (60:40/70:30/80:20), *number of average inflow at the delayed time steps* (3 and 7) and *learning rates* (0.1/0.01/0.001). Consequently, 54 scenarios of XGBoost daily and monthly models (@3×2×3×3) were trained and evaluated to produce good prediction results as shown in Fig.8.

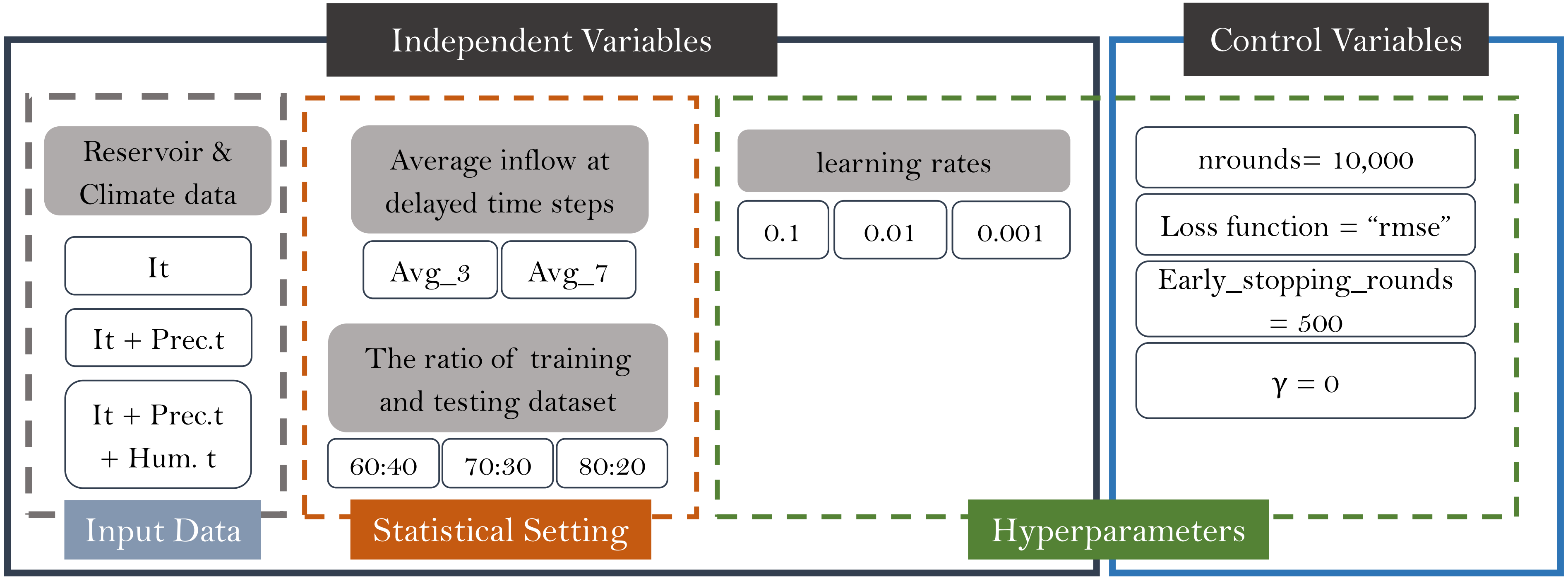
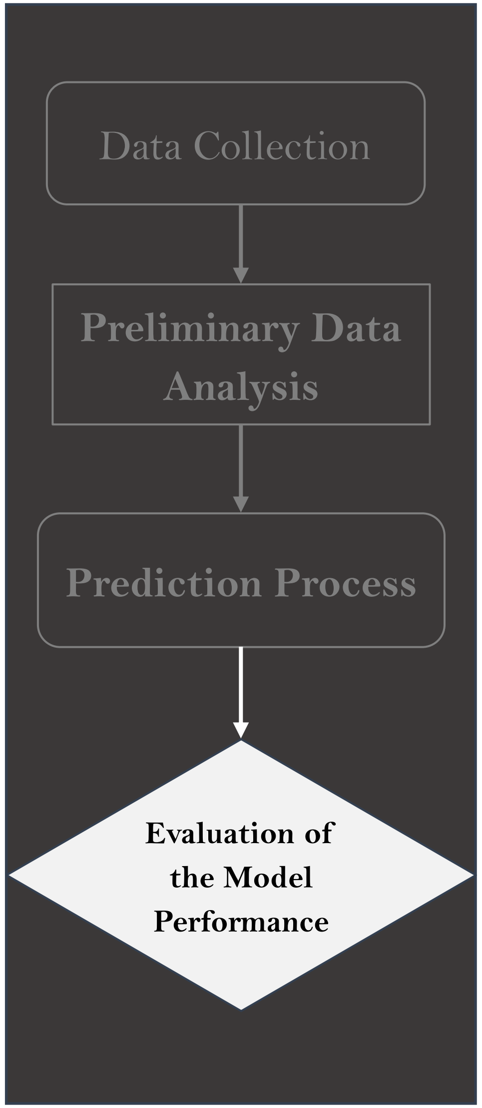


Figure 8 Input variables and model parameters for developing the reservoir inflow prediction models

Methodology




```r
Proportion6040_BB1 = 0.6
Train6040_BB1 = Input6040_BB1[1:round(Proportion6040_BB1*N1),]
head(Train6040_BB1)

Test6040_BB1 = Input6040_BB1[(round(Proportion6040_BB1*N1)+1):N1,]
head(Test6040_BB1)

Target6040_BB1 = Train6040_BB1$Inflow_1

Train6040_BB1Matrix = sparse.model.matrix(
  ~lag_1
  + avg_3
  , data = Train6040_BB1
  , sparse=FALSE, sci=FALSE)
Train6040_BB1DMatrix <- xgb.DMatrix(data = Train6040_BB1Matrix, label = Target6040_BB1)

Learning_Rate6040_BB1 = 0.1
params6040_BB1 = list(booster = "gbtree"
                      , object = "reg:linear"
                      , eta = Learning_Rate6040_  1
                      , gamma = 0
                      , max.depth = 6
)

xgb.tab <- xgb.cv(data = Train6040_BB1DMatrix
                  , param = params6040_BB1
                  , maximize = FALSE
                  , evaluation = "rmse"
                  , nrounds = 10000
                  , nthreads = 10
                  , nfold = 2
                  , set.seed(1)
                  , early_stopping_rounds = 500
                  )
```

```
> print(xgb.tab)
##### xgb.cv 2-folds
    iter train_rmse_mean train_rmse_std test_rmse_mean test_rmse_std
       1       28.009162      0.2063340       28.10540     0.1840155
       2       25.792614      0.2147640       25.97055     0.1573265
       3       23.831938      0.2241495       24.09864     0.1371535
       4       22.091543      0.2413625       22.48938     0.1302040
       5       20.544870      0.2519740       21.10766     0.1214000
---
     525        2.573136      0.1419780       15.28667     0.0616295
     526        2.572287      0.1413060       15.28725     0.0610510
     527        2.566983      0.1397850       15.28701     0.0625700
     528        2.561488      0.1376165       15.28851     0.0613340
     529        2.557986      0.1384890       15.28886     0.0613695
Best iteration:
 iter train_rmse_mean train_rmse_std test_rmse_mean test_rmse_std
```

Figure 9 RStudio; an open–source software library for R programming

To evaluate the prediction model performance, the statistical methods; Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Coefficient of Determination (R^2), Coefficient of Correlation (R), and Nash–Sutcliffe Efficiency (NSE) were used to indicate the perfect match between the predicted values (P_i) and observation values (O_i).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(O_i - p_i)^2}{n}} \quad \text{.....(8)}$$

$$MSE = \frac{\sum_{i=1}^{n}(O_i - p_i)^2}{n} \quad \text{.....(9)}$$

$$R^2 = \left[\frac{\left(\sum_{i=1}^{n}(O_i - \overline{O}) \cdot (p_i - \overline{p})\right)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2 \cdot \sum_{i=1}^{n}(p_i - \overline{p})^2}\right] \quad \text{.....(10)}$$

$$R = \frac{\sum_{i=1}^{n}(O_i - \overline{O}) \cdot (p_i - \overline{p})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2 \cdot \sum_{i=1}^{n}(p_i - \overline{p})^2}} \quad \text{.....(11)}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(O_i - p_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}\right] \quad \text{.....(12)}$$

Results & Discussions

Table 5 The predictive performance of the reservoir inflow prediction models of Bhumibol Dam during 2000–2020

Model setting	Model Inputs	Daily prediction model			Monthly prediction model		
Training: Testing Ratio	–	60:40	70:30	80:20	60:40	70:30	80:20
Inputs	Avg. Inflow t–1 to t–3 (Avg_3)	✓	✓	✓	–	–	✓
	Avg. Inflow t–1 to t–7 (Avg.7)	–	–	–	✓	✓	–
	Inflow t (It)	✓	✓	✓	✓	✓	✓
	Prec.t	–	–	–	✓	✓	✓
	Hum. t				✓	✓	✓
Learning rate	–	0.1	0.1	0.001	0.001	0.01	0.01
Training dataset	RMSE	7.9321	8.0515	7.3350	521.7199	499.3052	466.1194
	MSE	62.9187	64.8271	53.8019	272,191.6256	249,305.7171	217,267.3128
	R^2	0.9219	0.9198	0.9223	0.4119	0.4254	0.4523
	R	0.9602	0.9591	0.9604	0.6418	0.6522	0.6725
	NSE	0.9089	0.8980	0.9074	0.3805	0.3814	0.4112
Testing dataset	RMSE	5.6560	5.8255	6.5457	299.2648	263.0373	256.5848
	MSE	31.9904	33.9367	42.8461	89,559.448	69,188.5968	65,835.7496
	R^2	0.8854	0.8775	0.8661	0.6366	0.6621	0.6788
	R	0.9410	0.9367	0.9306	0.7979	0.8137	0.8239
	NSE	0.8619	0.8429	0.8307	0.4612	0.5975	0.6746

- The best daily prediction model was the observed inflow at time step t, and average inflow at the delayed time steps t–1 to t–3, the ratio of training and testing dataset 60:40 and 0.1 of learning rate

- The best monthly prediction model was the observed inflow at time step t, and average inflow at the delayed time steps t–1 to t–3, precipitation and humidity data, the ratio of training and testing dataset 80:20 and 0.001 of learning rate

Results & Discussions

The best predictive performance of daily inflow prediction model of Bhumibol Dam during 2000–2020

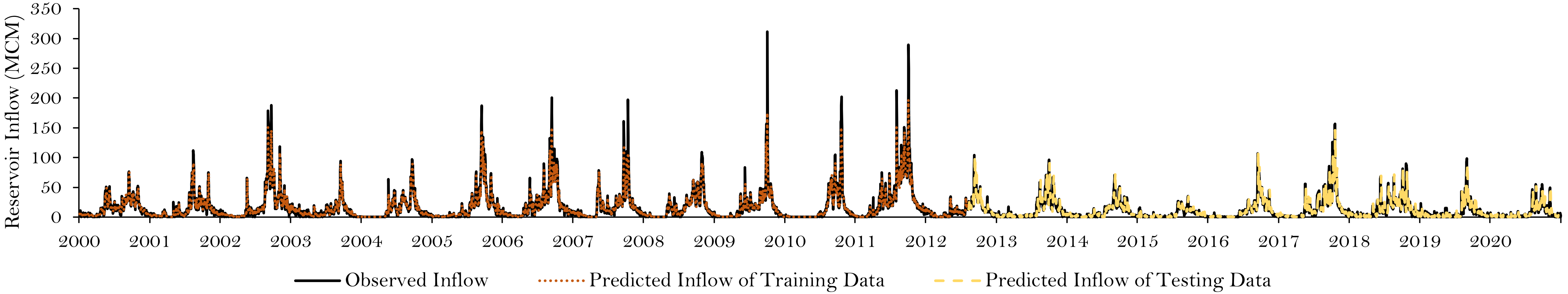


Figure 10 The qualitative comparison between observed and predicted inflows of the best daily model

Table 6 Comparison of predicted inflows obtained from the best daily prediction models and observed inflows of Bhumibol Dam

Model type	Daily Reservoir Inflow					
Model parameters	Training–Testing Ratio: 60:40					
	Inputs: Avg. Inflow t–1 to t–3					
	Learning Rate: 0.1					
Predictive performance	Average inflow (MCM/day)			Peak inflow (MCM/day)		
	Observed	Predicted	Δ (%)	Observed	Predicted	Δ (%)
Training data set	17.52	16.71	-0.81 (-4.62)	311.46	197.05	-114.41 (-36.73)
Testing data set	10.99	11.02	+0.03 (+0.27)	156.57	145.71	-10.86 (-6.93)

The figure 10 shows the qualitative comparison between observed and predicted inflows of the best daily model, it is obvious that the predicted inflows from training data are similar to the observed ones. However, under-estimated predictive results were found for the daily and monthly prediction models when the peak inflows were considerably investigated.

Results & Discussions

The best predictive performance of monthly inflow prediction model of Bhumibol Dam during 2000–2020

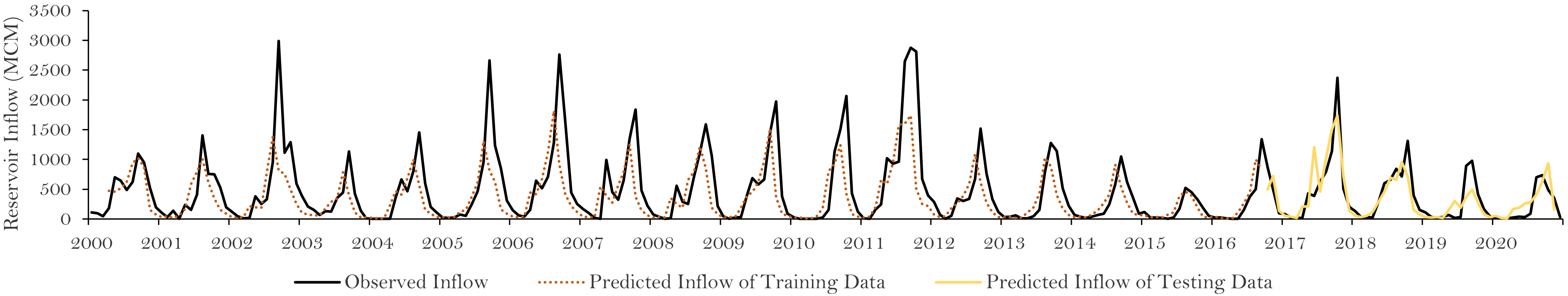


Figure 11 The qualitative comparison between observed and predicted inflows of the best monthly model

Table 7 Comparison of predicted inflows obtained from the best monthly prediction models and observed inflows of Bhumibol Dam

Model type	Monthly Reservoir Inflow					
Model parameters	Training–Testing Ratio: 80:20					
	Inputs: Inflow t, Avg. Inflow t–1 to t–3, Precipitation t and Humidity					
	Learning Rate: 0.001					
Predictive performance	Average inflow (MCM/month)			Peak inflow (MCM/month)		
	Observed	Predicted	Δ (%)	Observed	Predicted	Δ (%)
Training data set	482.45	360.10	-122.35 (-25.36)	2,990.21	1,811.99	-1,178.22 (-39.40)
Testing data set	370.31	359.75	-10.56 (-2.85)	2,373.51	1,740.76	-632.75 (-26.66)

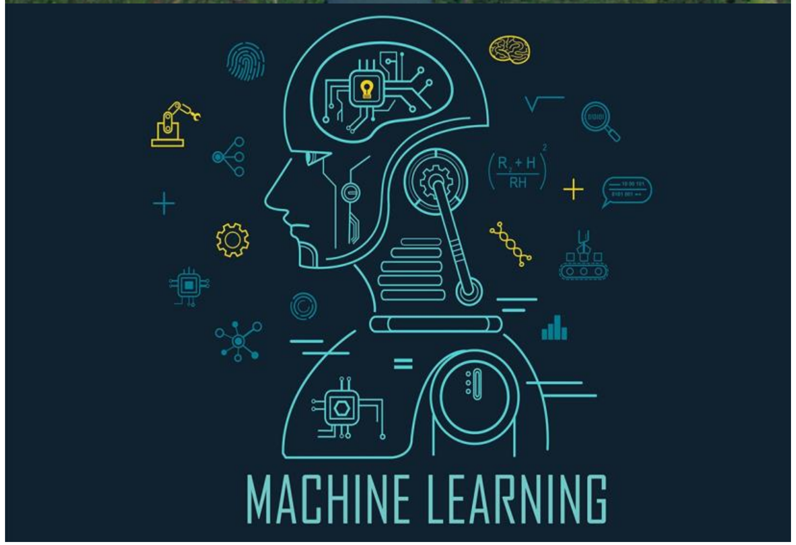
The figure 11 shows the qualitative comparison between observed and predicted inflows of the best monthly model, it is obvious that the predicted inflows from training data are similar to the observed ones. However, when it comes to the peak inflows, the predicted data cannot anticipate such high numbers.

Conclusions

- XGBoost which is a tree–based ensemble machine learning algorithm, was used to predict the daily and monthly reservoir inflows of the Bhumibol Dam, Thailand.

- The XGBoost model presented more reliable and robust prediction results especially for the daily prediction model with the highest R^2, R, NSE and small values of RMSE and MSE. It is found that the predictability of the XGBoost model to predict the daily reservoir inflow with good precision is strongly higher than the monthly inflow.

- Predicting the average values of the daily and monthly inflows gives the prediction results definitely closer to the observed inflows. However, the capability to characterize and predict the dynamics of extreme values of these two developed models is still limited. Therefore, to improve the quality of machine learning algorithm for hydrological prediction, the model parameters need to be optimized. In addition, conducting the further study using the technological advancement of machine learning is highly encouraged for the achievement of hydrological forecast on water resources management.



THANK YOU