**Advances in Computational Design**, *Vol. 7, No. 2 (2022) 113-128*
https://doi.org/10.12989/acd.2022.7.2.113 

# Water consumption prediction based on machine learning methods and public data

Witwisit Kesornsit[1a] and Yaowarat Sirisathitkul*[2]

[1]*Government Data Solution Division, Department of Data Solution, Digital Government Development Agency (Public Organization), Bangkok, Thailand*
[2]*Department of Computer Engineering and Electronics, School of Engineering and Technology, Walailak University, Nakhon Si Thammarat, Thailand*

*(Received September 16, 2021, Revised December 14, 2021, Accepted December 26, 2021)*

**Abstract.** Water consumption is strongly affected by numerous factors, such as population, climatic, geographic, and socio-economic factors. Therefore, the implementation of a reliable predictive model of water consumption pattern is challenging task. This study investigates the performance of predictive models based on multi-layer perceptron (MLP), multiple linear regression (MLR), and support vector regression (SVR). To understand the significant factors affecting water consumption, the stepwise regression (SW) procedure is used in MLR to obtain suitable variables. Then, this study also implements three predictive models based on these significant variables (e.g., SWMLR, SWMLP, and SWSVR). Annual data of water consumption in Thailand during 2006 – 2015 were compiled and categorized by provinces and distributors. By comparing the predictive performance of models with all variables, the results demonstrate that the MLP models outperformed the MLR and SVR models. As compared to the models with selected variables, the predictive capability of SWMLP was superior to SWMLR and SWSVR. Therefore, the SWMLP still provided satisfactory results with the minimum number of explanatory variables which in turn reduced the computation time and other resources required while performing the predictive task. It can be concluded that the MLP exhibited the best result and can be utilized as a reliable water demand predictive model for both of all variables and selected variables cases. These findings support important implications and serve as a feasible water consumption predictive model and can be used for water resources management to produce sufficient tap water to meet the demand in each province of Thailand.

**Keywords:** artificial neural network; machine learning; multi-layer perceptron; multiple linear regression; predictive model; stepwise regression; support vector regression; water consumption

## 1. Introduction

Over the past decades, fast-growing demand for water has been a global concern, especially in terms of household, agriculture, and industrial undertaking. Moreover, water provides a link between different elements in an ecosystem (Stampoulis *et al.* 2021). Many parts of the world are facing the risks of severe water scarcity situations (Yang *et al.* 2021). Water shortage has become a

*Corresponding author, Assistant Professor, E-mail: kinsywu@gmail.com; syaowara@mail.wu.ac.th
[a]M.Sc. Research Scholar., E-mail: witwisit.kes@gmail.com


http://www.techno-press.org/?journal=acd&subpage=7 ISSN: 2383-8477 (Print), 2466-0523 (Online)

major problem due to urbanization and industrialization (Wei *et al.* 2021). The discrepancy between water consumption and water supply has been notable particularly in the arid and semi-arid regions and in areas with expanding populations. According to the 2020 UN World Water Development Report, the quality, quantity, and accessibility of water for billions of people will be affected by climate change. Changes in water circulation patterns also endangers and severely threatens the achievement of the Sustainable Development Goals (UN Water, 2020). Therefore, it is essential to accurately estimate or predict water demand for adequate production planning especially in areas where water is seasonally limited, and climate is very diverse.

According to the National Strategy (2018–2037) of Thailand, the national development goals of the National Strategy are "a secure nation, contented people, continued economic growth, an equal society, and sustainable natural resources." The accomplishment of these goals is expected to occur by promoting environmentally sustainable growth and quality of life and improving the proficiency of government agencies (National Strategy Secretariat Office 2019). Therefore, water management is critical for policymakers in refining the provision of water supply services to ensure water availability and sustainability in the future.

The water demand management technique integrates demand and supply management such that future demand does not exceed water availability. In Addition, it can provide more feasible alternatives by enabling the integration of demand-side and supply-side management for enhanced water management (Xiao *et al.* 2021). Water consumption may be adversely affected by a large number of parameters, such as population growth and socio-economic, demographic, climatic, and geographic factors. The increasing water consumption is due to the increasing population, crop and livestock production and industrial plant (Stampoulis *et al.* 2021). To design a rule for water management, potential parameters influencing water demand should be examined in constructing a reliable water consumption predictive model. Thus, forecasting of water demand is vital for a proper management of provincial water systems. It is also essential for the operation of reservoirs provision, treatment plants, and distribution services. A reliable water consumption predictive model should also be capable of identifying the influencing factors affecting water storage and consumption reduction, water security, and urban sustainability.

Different methods and models in statistical analysis and machine learning are widely applied to forecasting of varying fields, such as science, health science, fisheries, industry, and social sciences. Machine learning has the proficiency to minimize the computational time, cost, and errors involved. Recent research shows that machine learning method within the group of data-driven is increasingly used to perform prediction as an attractive, rapid, and reliable computing tool. Machine learning techniques applied in prediction models include the support vector machine (SVM) (Agarwal *et al.* 2020), extreme learning machines (ELM) (Sattar *et al.* 2019, Abba *et al.* 2020a, Campos *et al.* 2020), stepwise regression (SW) (Ghani *et al.* 2010, Ahmad and Chen 2018), multiple linear regression (MLR) (Al-Hamad and Qamber 2019, Oyebode and Ighravwe 2019), evolutionary computation (EC) (Oyebode *et al.* 2019), k-nearest neighbor algorithms (KNN) (Benitez *et al.* 2019), artificial neural networks (ANN) (Cetinkaya and Baykan 2020, Thinakaran *et al.* 2020, Sahu *et al.* 2021), and random forest (RF) (Sirisathitkul *et al.* 2019, Pahlavan-Rad *et al.* 2020, Wongso *et al.* 2020). Stampoulis *et al.* (2021) illustrated the use of five machine learning models, i.e., linear regression, RF, KNN, support vector regression (SVR) and multi-layer perceptron (MLP) as a predictive model of precipitation rate and vegetation classification in east Africa for 2003 - 2011. Arora *et al.* (2021) examined the potential of SVM, RF, KNN, decision tree and naïve bayes in predicting prognosis of cervical cancer.

The relevant literature on water demand forecasting is summarized in many published articles

(Donkor *et al.* 2014, Ghalehkhondabi *et al.* 2017, de Souza Groppo *et al.* 2019). Existing forecasting models differ according to the variables to be forecasted, forecasting time period, and management objectives. Water demand forecasting approaches may be classified as either linear or nonlinear models (Romano and Kapelan 2014). The linear methods are not effective in the case of the nonlinearity of water demand data (Candelieri *et al.* 2015). The machine learning techniques are superior to statistical methods due to their capability to handle the nonlinearity and imprecision of available data (Ghalehkhondabi *et al.* 2017).

ANN, MLR, SW, and SVR models are the most common machine learning methods used and have been extensively applied to forecast water demand. Bata *et al.* (2020) used a self-organizing map (SOM) to investigate short-term water requirements in Southwestern Ontario, Canada. Yang *et al.* (2021) compared the performance of MLR, ANN and autoregressive state-space approach for China's Loess Plateau streamflow estimation model. Smolak *et al.* (2020) compared the performance of water consumption forecasting by using SVR and the associated machine learning techniques. Several research papers have compared the water consumption forecasting efficiency between conventional regression models and several ANN models (Adamowski and Karapataki 2010, Herrera *et al.* 2010, Abba *et al.* 2020b). Many previous studies suggested that the robust ANN technique was superior among all conventional models (Mouatadid and Adamowki 2016, Guo *et al.* 2018, Oyebode *et al.* 2019).

Moreover, various researchers have applied different combinations of machine learning-based models or hybrid approaches for prediction. A hybrid model comprises two or more methods, one of which works as the fundamental model, while the others perform preprocessing or postprocessing techniques (Zubaidi *et al.* 2020b). For example, Abba *et al.* (2020a) applied the linear regression and SW models to predict the water quality index at Kinta River Basin in Malaysia. Wu and Zhou (2010) showed that the combination of linear regression and ANN models offers more accurate forecasts than those separately obtained from each model. The hybrid model BSA-ANN was constructed by Zubaidi *et al.* (2020a) to determine the monthly water demand in the Gauteng province of the Republic of South Africa, which intensely suffered from the impact of population growth and climate change. Altunkaynak and Nigussie (2018) revealed that hybrid models based on MLP were robust and insightful. Furthermore, the performance of water prediction models has been optimized by using data preprocessing techniques (Seo *et al.* 2018, Zubaidi *et al.* 2020a). Another relevant consideration involves choosing the best scenario of model input. Zubaidi *et al.* (2018) applied a variance inflation factor (VIF) value to designate the model inputs that drive the explanatory variables.

According to the literature, it is noticeable that all studies in the context of prediction have shown the reliability of machine learning models. Many single models produce unacceptable results due to the limitation of the necessity of identifying suitable explanatory variables. Hence, to alleviate these shortcomings, it is necessary to develop a suitable selection approach for the appropriate variables and employ the combined techniques to enhance the accuracy of water consumption models.

To address abovementioned problems, this study aims to select distinctive factors related to water usage and proposes a reliable predictive model to forecast water consumption. It is focused on a medium-term prediction (1–10 years) and employs historical data on water consumption. The model performance is assessed by using the standard statistical measures. The accuracy of appropriate models is essential to forecast the future water demand. The development of highly reliable models is also expected to perform a crucial role in planning provincial and metropolitan waterworks policies.

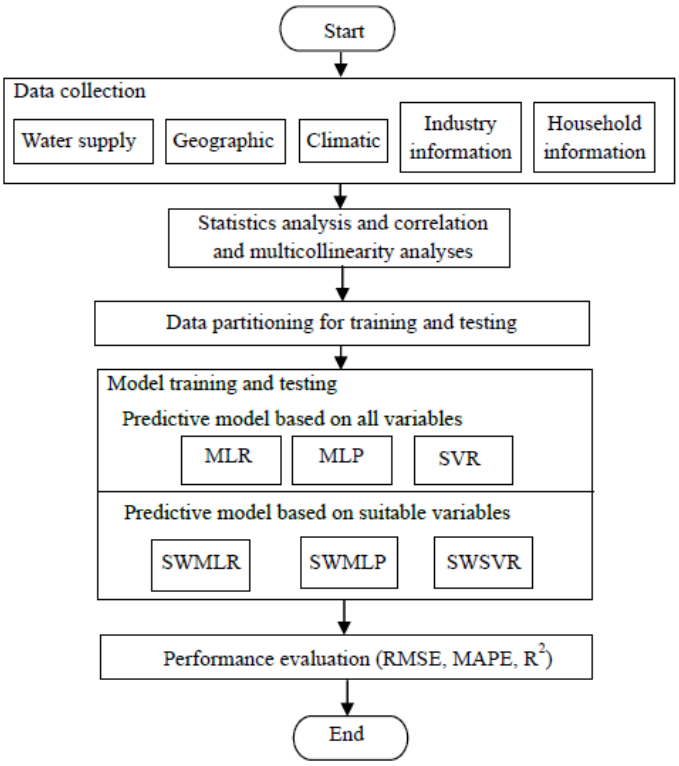


Fig. 1 Framework for the water consumption prediction

## 2. Materials and methods

The flowchart used in this analysis is depicted in Fig 1. This study starts with data collection. The missing data on factory amounts and registered capital are imputed by the exponential smoothing technique. The data is then passed to the statistics analysis and correlation and multicollinearity analyses. The developed models are trained and tested with 10 replications. A simple sampling method was used to randomize the data into training and testing dataset at a ratio of 70%:30%; with a total 750 observations, the training data corresponded to 70% of the measurement (525 of 750) and test data corresponded to 30% of the measurement (225 of 750). All six predictive models (e.g., MLR, MLP, SVR, SWMLR, SWMLP, and SWSVR) were implemented and tested using the R programming language and software environment on an Intel® Core™ i7-8550U CPU at 1.80 GHz, running a 64-bit Windows 10 operating system with 8 GB RAM. The performance metrics are used to evaluate these models. The model with the highest index is selected. A detailed of each step is given as sub-section.

### *2.1 Data collection*

This study used the annual data of water consumption amount from the metropolitan waterworks authority station and provincial waterworks authority stations during 2006 – 2015 for 75 provinces across Thailand. The measured water consumption data were collected from open

Table 1 Summary of the response and explanatory variables employed for the water consumption predictive model

| Variable type | Category | Variable name | Type | Description |
|---|---|---|---|---|
| Response | Water supply | Water amount | Numeric | - |
| Explanatory | Water consumption | Tap water type | Nominal | 0 – Metropolitan<br>1 – Provincial |
| | | Sector | Nominal | 0 – Central and East<br>1 – North<br>2 – North East<br>3 – South |
| | | Water supply user amount | Numeric | |
| | | Water sales amount | Numeric | |
| | Geographic | Population | Numeric | |
| | | Area | Numeric | |
| | | People per sq. km | Numeric | |
| | Climatic | Mean dry bulb temperature | Numeric | |
| | | Mean maximum temperature | Numeric | |
| | | Mean minimum temperature | Numeric | |
| | | Mean msl pressure | Numeric | |
| | | Mean relative humidity | Numeric | |
| | | Mean station pressure | Numeric | |
| | | Total rainfall | Numeric | |
| | | Std total rainfall | Numeric | |
| | Industry information | Factory amount | Numeric | |
| | | Registered capital | Numeric | |
| | | Worker amount | Numeric | |
| | Household information | Average monthly income | Numeric | |
| | | Average monthly expenditure | Numeric | |
| | | Average debt per household | Numeric | |
| | | Number of households | Numeric | |

data, which were assembled from publicly open sources, such as the official website of Open Government Data of Thailand (Digital Government Development Agency 2019), National Statistical Office of Thailand (National Statistical Office 2019), and Thai Meteorological Department (Thai Meteorological Department 2019).

For the purposes of this analysis, the 22 explanatory variables are grouped into five categories, including water supply, geographic, climatic, and industrial information, as well as household information. It should be noted that water consumption data are available at every 5-year period and the explanatory variables are refined to match the time frames of response variable. The variables of the annual averaged data for the 10-year period between 2006 and 2015 from 75 provinces with 750 instances. In this study, the water amount is used as a response variable, and other variables are used as explanatory variables as shown in Table 1.

### 2.2 Statistics analysis and correlation and multicollinearity analyses

The data were processed through an initial exploratory analytical investigation on each input by checking the normal distribution, linearity, and multicollinearity between independent variables. The normal distribution of all independent variables must be investigated by kurtosis and skewness coefficients. The existence correlation between the two variables was investigated by applying the Pearson correlation. The value of VIF indicated the presence of multicollinearity between independent variables in the model. This implies that these variable additions have no impact on the prediction accuracy. Thus, suitable explanatory variables were selected by applying the VIF value. Ghani and Ahmad (2010) suggested that the values of VIF greater than 5 or 10 could be used to detect multicollinearity between independent variables. Based on practical consideration in our study, the VIF cutoff value should be greater than 10. Therefore, some of the independent variables violating this should be removed from the model. Out of 22 explanatory variables listed in Table 1, three variables namely: "Water supply user amount", "People per sq. km" and "Number of households" are left out.

### 2.3 Models development based on all variables

#### 2.3.1 Multiple Linear Regression (MLR)

The MLR is a regression technique for analyzing the relationship between a single response variable and two or more explanatory variables. The mathematical formula is linear equations of the form:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \quad (1)$$

where $y$ is the response variable; $x_1, x_2, \ldots x_n$ are the $n$ known explanatory variables; and $b_0, b_1, b_2, \ldots b_n$ are estimated parameter coefficients of intercept and slopes, respectively (Davis, 2011).

In this study, the MLR is used as a benchmark model. The MLR is used to evaluate the relationship among all 19 explanatory variables and the response variable. The coefficients and intercept values in the MLR equation are estimated by using an ordinary least squares (OLS) method.

#### 2.3.2 Multi-Layer Perceptron (MLP)

The MLP is an ANN, structured as a multi-layer design based on the human nervous system. The MLP is the most common form of ANN, which has an input layer with one or more hidden layer connected to one output layer. The input variable (i.e., independent variable) can be either categorical or numeric types; however, the dependent variable must be numeric (Anand and Suganthi 2018, Farizawani *et al.* 2020).

For this study, MLP is constructed with 19 nodes in the input layer, 14 nodes in three hidden layers (i.e., 8, 4, 2), and 1 node in the output layer. Each input variable was normalized to a [0, 1] interval, and the logistic activation function was utilized to define the output for each neuron in the hidden layer. The MLP was trained with the training dataset with a learning rate of 0.001, and the stopping criterion was either encountering the training goal or reaching the limit of 1,000 epochs.

#### 2.3.3 Support Vector Regression (SVR)

The SVR is developed from SVM to provide acceptable solutions to nonlinear problems and

the provided dataset (Zhong *et al.* 2019, Smolak *et al.* 2020). For the multi-dimensional datasets, the multivariate regression was used as follows:

$$f(x) = \omega^T \boldsymbol{x} + b \tag{2}$$

where $\omega \in \boldsymbol{R}^n$ is a weighted feature vector, $b$ is the intercept and x is the input vector in $\boldsymbol{R}^n$.

The SVR formulates this function as an optimization problem while minimizing the prediction error (Smolak *et al.* 2020). In this study, the SVR was constructed with 19 input variables and performed an epsilon regression and tuned method to train models with epsilon $(\epsilon)$ in $0, 0.1, 0.2, \ldots, 1$ and a cost parameter with $2^2, 2^3, \ldots, 2^9$.

### *2.4 Model development based on suitable variables*

#### *2.4.1 Stepwise Multiple Linear Regression (SWMLR)*

The SW is a statistical regression method designed to select the suitable explanatory variables using an automatic procedure. In each step, the addition or subtraction of explanatory variables from the model are through a series of tests (e.g. F-tests, t-tests) to find a set of independent variables that significantly influence the dependent variable.

For this study, the SWMLR was established by using the MLR with the OLS parameter estimation and stepwise procedure. All 19 explanatory variables were introduced into or removed individually from the stepwise regression equation. A series of AIC were performed for the selected explanatory variables in which the significance levels for F-to-enter and-to-remove were set at 0.05 and 0.05, respectively. The process was iterated until there were no more significant explanatory variables. A total of eight explanatory variables are selected out of 19 variables. SWMLR has also been used to detect the significant explanatory variables as demonstrated by Ahmad and Chen (2018) and Yang *et al.* (2021).

#### *2.4.2 Stepwise Multi-Layer Perceptron (SWMLP)*

The SWMLP employed the eight selected suitable explanatory variables from the SWMLR as input variables. It is constructed with eight nodes in the input layer, six nodes in two hidden layers (i.e., 4 and 2), and one node in the output layer. The learning rating and the stopping criterion are equal to those of the MLP.

#### *2.4.3 Stepwise Support Vector Regression (SWSVR)*

The SWSVR utilized the eight input variables to construct a function estimator in the predictive model with the same condition, as mentioned in the SVR.

### *2.5 Model evaluation metrics*

Machine learning provides several approaches for partitioning experimental data, such as training/testing partitioning and cross-validation (Liu and Cocea 2017). In this study, 70% of the completed data was utilized for training, while 30% was designated for testing the models, as previously performed by (Ahmad and Chen 2018, Oyebode and Ighravwe 2019). The accuracy of each model in the training/testing stage is evaluated by calculating the difference between the predicted and real values in the holdout sample. The determination of the best model is based on these measures, with the minimum mean error used to obtain high accuracy of the future predictions. In this study, three sets of statistical measures were employed to investigate the

accuracy of the predictive model, including root mean-square error (RMSE), mean absolute percent error (MAPE), and coefficient of determination ($R^2$), as suggested in (Oyebode and Ighravwe 2019, Abba *et al.* 2020a, Sahu *et al.* 2021). The best model was elected based on the lowest MAPE (Sahu *et al.* 2021), lowest RMSE, and highest $R^2$ (Agarwal *et al.* 2020, Yang *et al.* 2021). The details of these indices are expressed below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}} \tag{3}$$

$$MAPE = 100\frac{1}{N}\sum_{i=1}^{N}\frac{|O_i - P_i|}{O_i} \tag{4}$$

$$R^2 = \left[\frac{\sum_{i=1}^{N}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2 \times \sum_{i=1}^{N}(P_i - \bar{P})^2}}\right] \tag{5}$$

where $N$ is the number of measurements, $O_i$ is the amount of water observed at sample $i$ and $P_i$ is the amount of water predicted at sample $i$, $\bar{O}$ is an observed average of the amount of water and $\bar{P}$ is a predicted average of the amount of water.

## 3. Results and discussions

### *3.1 Statistics analysis and correlation and multicollinearity analyses*

All the data were analyzed using the R programming langue software package version 4.0.3 and the associated framework. The statistical descriptions of numeric variables during 2006 – 2015 are presented in Table 2. The water consumption variables exhibited variations in extended ranges, such as the water amount varying from $5.1525\times10^5$ to $1.8351\times10^9$ $m^3$ and water supply user amount varying from 1,987 to $1.5584\times10^6$ per person. The variation patterns of the water consumption parameters confirm the complex, nonlinear modelling process. The standard deviation values are close to the mean, and most skewness coefficient values were also low for the dataset. This indicates that the model fits the data well. According to the kurtosis coefficient values, most variables were normally distributed, with the exception of water supply user amount, people per sq. km, and number of households.

Table 2 Basic descriptive statistics of water consumption numeric variables

| Category | Variables | Unit | Minimum | Maximum | Mean | Standard deviation | Kurtosis coefficient | Skewness coefficient |
|---|---|---|---|---|---|---|---|---|
| Water consumption | Water amount | $m^3$ | 5.1525E +05 | 1.8351E +09 | 8.8497E +07 | 3.4238E +08 | 19.973 | 4.657 |
| Water supply | Water supply user amount | Person | 1987.000 | 1.5584E +06 | 68642. 365 | 1.6812E +05 | 53.851 | 7.0479 |

Table 2 Continued

| Category | Category | Category | Category | Category | Category | Category | Category | Category |
|---|---|---|---|---|---|---|---|---|
| Water supply | Water sales amount | $m^3$ | 358675.000 | 1.383E+09 | 6.3554E+07 | 2.4760E+08 | 20.227 | 4.678 |
| Geographic | Population | Person | 193305.000 | 5666264.000 | 874647.107 | 736015.713 | 21.894 | 3.904 |
| | Area | $km^2$ | 416.710 | 20493.960 | 6731.918 | 4687.950 | 0.712 | 1.008 |
| | People per sq. km | Person/$km^2$ | 22.000 | 3612.000 | 247.960 | 487.265 | 30.377 | 5.223 |
| Climatic | Mean dry bulb temperature | °C | 23.725 | 29.241 | 27.211 | 0.975 | 1.730 | -1.027 |
| | Mean maximum temperature | °C | 29.132 | 33.957 | 32.766 | 0.813 | 3.511 | -1.166 |
| | Mean minimum temperature | °C | 19.018 | 25.332 | 22.847 | 1.320 | 0.471 | -0.633 |
| | Mean msl pressure | Mb | 1008.208 | 1010.813 | 1009.331 | 0.413 | 1.825 | 0.454 |
| | Mean relative humidity | % | 69.526 | 83.699 | 75.765 | 3.299 | -0.372 | 0.689 |
| | Mean station pressure | mb | 938.916 | 1009.251 | 1001.009 | 10.448 | 14.673 | -3.112 |
| | Total rainfall | mm | 81.997 | 403.816 | 128.924 | 51.578 | 10.600 | 2.846 |
| | Std total rainfall | - | 62.571 | 366.172 | 102.664 | 46.469 | 14.736 | 3.519 |
| Industry information | Factory amount | - | 262.000 | 19664.000 | 1858.004 | 2470.558 | 25.486 | 4.464 |
| | Registered capital | THB | 1623.000 | 1083873.409 | 70229.953 | 150161.765 | 17.964 | 3.887 |
| | Worker amount | - | 2677.000 | 575317.000 | 52927.813 | 102526.394 | 12.307 | 3.400 |
| Household information | Average monthly income | THB | 6544.000 | 49190.800 | 20223.499 | 6543.656 | 2.028 | 1.069 |
| | Average monthly expenditure | THB | 6332.000 | 35023.700 | 17634.290 | 4785.037 | 1.065 | 0.947 |
| | Average debt per household | THB | 5778.000 | 386957.400 | 129981.960 | 51611.974 | 1.575 | 0.606 |
| | number of households | - | 54.647 | 2913.929 | 267.646 | 273.381 | 40.322 | 5.516 |

| | Water amount | Tab water type | Sector | Water supply user amount | Water sale amount | Population | Area | People per sq. km | Mean dry bulb temperature | Mean maximum temperature | Mean minimum temperature | Mean msl pressure | Mean relative humidity | Mean station pressure | Total rainfall | Std total rainfall | Factory amount | Registered capital | Worker amount | Average monthly income | Average monthly expenditure | Average debt per household | Number of household |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Water amount | 1.0000 | | | | | | | | | | | | | | | | | | | | | | |
| Tab water type | 0.1476 | 1.0000 | | | | | | | | | | | | | | | | | | | | | |
| Sector | 0.1068 | 0.1411 | 1.0000 | | | | | | | | | | | | | | | | | | | | |
| Water supply user amount | 0.1377 | -0.7212 | -0.2476 | 1.0000 | | | | | | | | | | | | | | | | | | | |
| Water sale amount | 0.9953 | 0.1427 | 0.1102 | 0.1446 | 1.0000 | | | | | | | | | | | | | | | | | | |
| Population | 0.1774 | -0.5231 | -0.0752 | 0.8299 | 0.1843 | 1.0000 | | | | | | | | | | | | | | | | | |
| Area | 0.0468 | 0.2459 | -0.0016 | -0.1159 | 0.0468 | 0.3177 | 1.0000 | | | | | | | | | | | | | | | | |
| People per sq. km | 0.0335 | -0.8724 | -0.1955 | 0.8927 | 0.0359 | 0.6694 | -0.3388 | 1.0000 | | | | | | | | | | | | | | | |
| Mean dry bulb temperature | 0.1315 | -0.3756 | -0.0763 | 0.3315 | 0.1388 | 0.0402 | -0.5630 | 0.4557 | 1.0000 | | | | | | | | | | | | | | |
| Mean maximum temperature | -0.0739 | -0.1338 | -0.0225 | 0.0734 | -0.0749 | -0.0812 | -0.2748 | 0.1542 | 0.6731 | 1.0000 | | | | | | | | | | | | | |
| Mean minimum temperature | 0.1590 | -0.3897 | -0.0770 | 0.3275 | 0.1708 | 0.0242 | -0.6051 | 0.4516 | 0.8982 | 0.3679 | 1.0000 | | | | | | | | | | | | |
| Mean msl pressure | 0.1759 | 0.0752 | 0.1681 | -0.0213 | 0.2003 | -0.0052 | -0.0221 | -0.0642 | 0.0014 | -0.1832 | 0.2664 | 1.0000 | | | | | | | | | | | |
| Mean relative humidity | -0.0685 | 0.2223 | -0.2228 | -0.2426 | -0.0720 | -0.3008 | -0.2191 | -0.2198 | -0.0510 | -0.3000 | 0.2412 | 0.4865 | 1.0000 | | | | | | | | | | |
| Mean station pressure | -0.0386 | -0.1373 | -0.0398 | 0.0683 | -0.0340 | -0.2287 | -0.6199 | 0.1884 | 0.7780 | 0.5439 | 0.8080 | 0.3118 | 0.3710 | 1.0000 | | | | | | | | | |
| Total rainfall | -0.1330 | -0.0113 | -0.1058 | -0.0373 | -0.1409 | -0.1288 | -0.1614 | -0.0126 | -0.0649 | -0.3480 | 0.1572 | 0.3641 | 0.6327 | 0.2150 | 1.0000 | | | | | | | | |
| Std total rainfall | -0.1613 | 0.0232 | -0.0238 | -0.0575 | -0.1701 | -0.1429 | -0.1305 | -0.0605 | -0.0758 | -0.3491 | 0.1104 | 0.2417 | 0.3987 | 0.1064 | 0.7945 | 1.0000 | | | | | | | |
| Factory amount | 0.2318 | -0.6372 | -0.0648 | 0.5790 | 0.2362 | 0.8928 | 0.0735 | 0.7684 | 0.2356 | 0.0524 | 0.1987 | -0.0666 | -0.3440 | -0.0604 | -0.1452 | -0.1358 | 1.0000 | | | | | | |
| Registered capital | 0.3947 | -0.3273 | 0.0283 | 0.3636 | 0.4173 | 0.2558 | -0.1880 | 0.3263 | 0.3770 | 0.1486 | 0.4070 | 0.0617 | -0.0996 | 0.1590 | -0.1495 | -0.1756 | 0.4404 | 1.0000 | | | | | |
| Worker amount | 0.3902 | -0.6672 | -0.1029 | 0.7580 | 0.3985 | 0.5912 | -0.2314 | 0.7355 | 0.4622 | 0.1955 | 0.4459 | -0.0091 | -0.2345 | 0.1614 | -0.1386 | -0.1498 | 0.5257 | 0.6838 | 1.0000 | | | | |
| Average monthly income | 0.2232 | -0.4422 | -0.2451 | 0.5385 | 0.2204 | 0.3049 | -0.3091 | 0.5575 | 0.4730 | 0.1702 | 0.5327 | 0.1525 | 0.1817 | 0.4014 | 0.1364 | 0.0170 | 0.4098 | 0.3551 | 0.4901 | 1.0000 | | | |
| Average monthly expenditure | 0.3013 | -0.4821 | -0.1780 | 0.5238 | 0.3061 | 0.2634 | -0.3626 | 0.5851 | 0.5351 | 0.1495 | 0.6181 | 0.2059 | 0.1259 | 0.4412 | 0.0740 | -0.0468 | 0.3854 | 0.4045 | 0.5349 | 0.8319 | 1.0000 | | |
| Average debt per household | 0.1309 | -0.2409 | -0.0521 | 0.2453 | 0.1288 | 0.1806 | -0.0540 | 0.2365 | 0.2287 | 0.1865 | 0.1747 | -0.0683 | -0.1254 | 0.1199 | -0.1069 | -0.1196 | 0.1917 | 0.1353 | 0.1789 | 0.5525 | 0.4921 | 1.0000 | |
| Number of household | 0.1369 | -0.5813 | -0.1641 | 0.5986 | 0.1415 | 0.9637 | 0.2059 | 0.7367 | 0.1093 | -0.0353 | 0.0880 | -0.0400 | -0.2949 | -0.1645 | -0.1093 | -0.1169 | 0.5092 | 0.2807 | 0.6472 | 0.3746 | 0.3222 | 0.1897 | 1.0000 |

Fig. 2 Correlation matrix between the water consumption variables

The strength and direction of the association between all water consumption variables via the correlation matrix are further analyzed and shown in Fig. 2. The Pearson correlation ($r$) is used to analyze the correlation between variables x and y in the predictive model. It is crucial to investigate the correlation coefficient between the water consumption variables because it illustrates the linear relation between the dependent and independent variables.

Fig. 2 shows a quite weak correlation of the water amount with area, people per sq. km, mean maximum temperature, mean relative humidity, and mean station pressure ($r$ = 0.0468, $r$ = 0.0335, $r$ = -0.0739, $r$ = -0.0685, $r$ = -0.0386, respectively). Positive correlation coefficients were obtained for water supply user amount and population and people per sq. km ($r$ = 0.8299, $r$ = 0.8927, respectively). There was strong evidence of a positive correlation between the population and factory amount and number of households ($r$ = 0.8928, $r$ = 0.9637, respectively). In addition, there was a substantial positive correlation between the mean minimum temperature and mean station pressure and mean dry bulb temperature ($r$ = 0.8080, $r$ = 0.8982, respectively). In contrast, there was a negative correlation between the people per sq. km and tab water type, $r$ = -0.8724. For the multicollinearity analysis, the VIF was computed. The dependent variables with a high correlation coefficient of the VIF over 10 comprised population, people per sq. km, and mean minimum temperature. Only these three variables were removed from the explanatory variables. As a result, the 19 explanatory variables were entered into the MLR, MLP, and SVR models.

It is challenging to predict the water consumption with many factors, thus reducing the complexity of input variables instead of large amounts of factors simplifies the analysis (Wei *et al.* 2021). In this study, the SW procedure was used in the MLR to select the appropriate explanatory variables, as suggested by Ahmad and Chen (2018) and Yang *et al.* (2021). Therefore, eight suitable variables were obtained: water sales amount, mean dry bulb temperature, mean maximum temperature, mean msl pressure, mean relative humidity, total rainfall, factory amount, and average monthly income. The selected significant variables by SW are in the water supply, climatic and industry information categories. As stated by Xiao *et al.* (2021), most water usages come from the industrial sector. Subsequently, these variables were used as inputs for the SWMLP and SWSVR models.

Table 3 Performance results for the MLR, MLP, SVR, SWMLR, SWMLP, and SWSVR models

| No. of input variables | Model | Training | | | Testing | | |
|---|---|---|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ | RMSE | MAPE | $R^2$ |
| 19 | MLR | 0.0235 | 2.5097 | 0.9925 | 0.0328 | 3.2460 | 0.9296 |
| | MLP | **0.0144** | **1.7099** | **0.9979** | **0.0247** | **2.3535** | **0.9829** |
| | SVR | 0.0535 | 4.3076 | 0.9582 | 0.0688 | 4.7494 | 0.9130 |
| 8 | SWMLR | 0.0257 | 2.5097 | 0.9299 | 0.0356 | 3.2512 | **0.9924** |
| | SWMLP | **0.0132** | **1.9772** | **0.9972** | **0.0248** | **2.9330** | 0.9844 |
| | SWSVR | 0.0535 | 4.5668 | 0.9575 | 0.0826 | 4.9348 | 0.8899 |

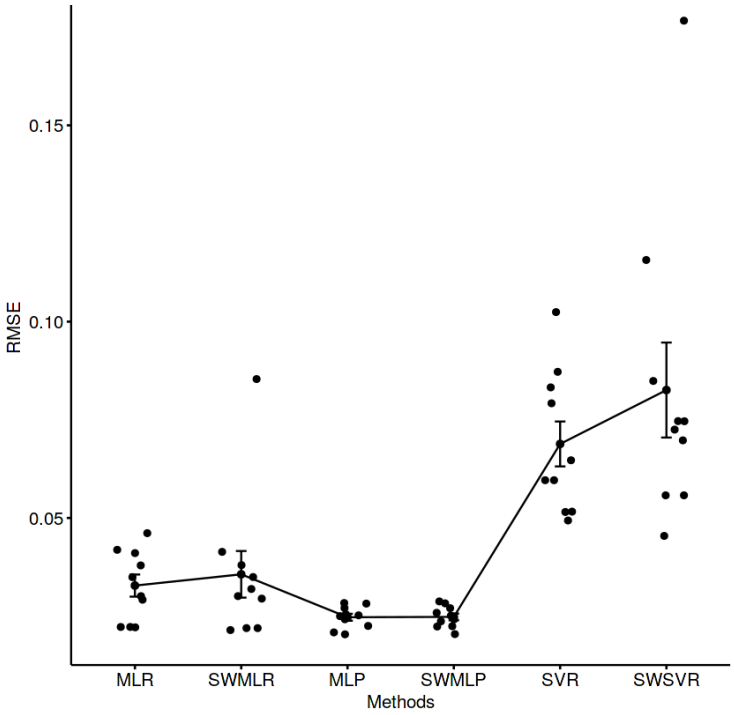


Fig. 3 Box plot of the average RMSEs of the six predictive models

### *3.2 Model comparison*

The evaluation measurements of the model in which 19 explanatory variables were entered into the MLR, MLP, and SVR models, as presented in Table 3. The SWMLR, SWMLP, and SWSVR models with the selected eight explanatory variables are also compared in the table.

According to Agarwal *et al.* (2020), the RMSE is one of the principal predictive quantitative measures for evaluating the performance of machine learning models. Hence, the minimum values of the RMSE varied in the range of 0.0132–0.0826 for the training and testing, respectively. According to the results, the MLP models exhibited the best performance, while the SWSVR models had the worst performance in the testing phase consistent with the report by Smolak *et al.* (2020). Furthermore, the quantitative results indicated that the MLP with the RMSE = 0.0247 and

Table 4 Kruskal–Wallis chi-square test statistics for the RMSE in the testing phase of the six predictive models

| Kruskal-Wallis rank sum test | | | | | |
|---|---|---|---|---|---|
| Kruskal-Wallis chi-squared ($\chi^2$) = 39.838 | | | df = 5 | p-value = **0.0000** | |
| Pairwise comparisons using Wilcoxon rank sum test | | | | | |
| | MLR | SWMLR | MLP | SWMLP | SVR |
| SWMLR | 0.7859 | - | - | - | - |
| MLP | 0.0807 | 0.0807 | - | - | - |
| SWMLP | 0.1111 | 0.1111 | 0.9705 | - | - |
| SVR | **0.0005** | **0.0032** | **0.0005** | **0.0005** | - |
| SWSVR | **0.0006** | **0.0032** | **0.0005** | **0.0005** | 0.6582 |

MAPE = 2.3535 in the testing phase exhibited superior prediction skills among the other models and then emerged as a reliable predictive model. The analysis yielded the highest accuracy of all evaluation indices for the MLP except $R^2$ for the SWMLP. Moreover, box plots are commonly used to illustrate the accuracy of the prediction by the model (Abba *et al.* 2020a). Fig. 3 presents the box plot of the RMSE in 10 replications in the testing phase of the six predictive models. The best predictive model was clearly the MLP, the values output by which did not exhibit a large variation. In the follow-up prediction patterns, the MLP model was the most competent and reliable prediction tool for all forecasting scenarios.

For the models with a fewer number of variables, the SWMLP had the highest performance in the prediction (RMSE = 0.0248, MAPE = 2.9330). The SWMLR model demonstrated the best performance in the prediction ($R^2$ = 0.9924). The predicting performance of the SWSVR was deteriorated drastically compared to the other models. For the testing phase, the RMSE of the MLP and SWMLP models was comparable (RMSE = 0.0247, RMSE = 0.0248, respectively). As noted by Olyaie *et al.* (2017), an independent variable with a low skewness coefficient is appropriate for ANN models. Because the eight variables used in the SWMLP had a low skewness coefficient, the SWMLP, thus, had a closer performance to the MLP. However, the SWMLP excluded 11 variables as they had no significance in the prediction. This exclusion resulted in comparable performance of the model to the MLP model that used 19 explanatory variables as inputs. Considering the value of the forecasting time and cost, the use of a large number of independent variables to create predictive equations tended to result in considerable model complexity, which was notably time consuming and required greater computational resources. Therefore, if the performance of the prediction was not greatly different, it can be concluded that a method using fewer independent variables was better than that with more independent variables.

The Kruskal–Wallis test, a non-parametric comparison test method, was also used to order the significance of the input variables. The p-values measure the strength of the association between dependent and independent variables. In addition, the Wilcoxon' rank sum test was used to calculate the pairwise comparisons between models. The pairwise comparison method identifies which pairs are significantly different. Comparing the six predictive models in Table 4, the outputs of the Kruskal–Wallis test indicated a statistically significant difference (p-value = 0.0000). Meanwhile, the Wilcoxon rank sum test showed that the SVR and SWSVR models were different from the other models.

## 4. Conclusions

In this study, the performance of combined models was assessed to determine annual water consumption based on previous water usage. Historical data of yearly water consumption in each province over 10 years in Thailand were utilized to construct and assess the predictive models. The outcomes showed that the SW procedure could be used to select significant variable for MLR, MLP, and SVR models to simulate the water consumption for provincial waterworks in Thailand. Moreover, the SW procedure was successfully applied to select the distinctive explanatory variables. The benefits of using a small set of input are evading the time-costly computations, easing the modeling problem. It leads to a nimble and optimized predictive model, and also the most agile and optimal learning modules. The overall outcome has demonstrated that the MLP and SWMLP proved effective and satisfactory based on the fitness function (i.e., RMSE) and, hence, served as a reliable predictive model. Such a model could be of benefit in providing a reliable water supply to meet the demand of household and industrial sectors while alleviating excessive water consumption.

The results of this study highlight the remarkable ability of machine learning models to determine relationship of water consumption to five categories, including water supply, geographic, climatic, and industrial information, as well as household information. The most selected variables by SW are in the climatic category, expressing the relationship between water consumption activities and climatic factors. It is then of great importance to systematically examine their impact on water' regime with the local climate. Future research is required to clarify these relationships for sustainable water management in water-stressed regions.

## Acknowledgement

The data used in this study was kindly provided by Department of Data Solution, Open Government Data of Thailand, National Statistical Office of Thailand and Thai Meteorological Department. The authors would like to thank the government data solution project team for their collaborative support and guidance during this study. This research was financially supported by the new strategic research project (P2P), Walailak University, Thailand.

## References

Abba, S.I., Hadi, S.J., Sammen, S.S., Salih, S.Q., Abdulkadir, R.A., Pham, Q.B. and Yaseen, Z.M. (2020a), "Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination", *J. Hydrol.*, **587**, 124974. https://doi.org/10.1016/j.jhydrol.2020.124974.

Abba S.I., Pham, Q.B., Saini, G., Linh, N.T.T., Ahmed, A.N., Mohajane, M., Khaledian, M., Abdulkadir, R.A. and Bach, Q.V. (2020b), "Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index", *Environ. Sci. Pollut. Res*., **27**, 41524-41539. https://doi.org/10.1007/s11356-020-09689-x.

Adamowski, J. and Karapataki, C. (2010), "Comparison of multivariate regression and artificial neural networks for peak urban water demand forecasting: evaluation of different ANN learning algorithms", *J. Hydrol. Eng.*, **15**(10), 729-743. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000245.

Agarwal, S., Dandge, S.S. and Chakraborty, S. (2020), "A support vector machine-based prediction model for electrochemical machining process", *Karbala Int. J. Mod. Sci*., **6**(2), 164-174.

https://doi.org/10.33640/2405-609X.1508.
Ahmad, T. and Chen, H. (2018), "Utility companies strategy for short-term energy demand forecasting using machine learning based models", *Sustain. Cities Soc.*, **39**, 401-417. https://doi.org/10.1016/j.scs.2018.03.002.
Al-Hamad, M.Y. and Qamber, I.S. (2019), "GCC electrical long-term peak load forecasting modeling using ANFIS and MLR methods", *Arab J. Basic Appl. Sci*., **26**(1), 269-282. https://doi.org/10.1080/25765299.2019.1565464.
Altunkaynak, A. and Nigussie, T.A. (2018), "Monthly water demand prediction using wavelet transform, first-order differencing and linear detrending techniques based on multilayer perceptron models", *Urban Water J.*, **15**(2), 177-181. https://doi.org/10.1080/1573062X.2018.1424219.
Anand, A. and Suganthi, L. (2018), "Hybrid GA-PSO optimization of artificial neural network for forecasting electricity demand", *Energies*, **11**(4), 728. https://doi.org/10.3390/en11040728.
Arora, M., Dhawan, S. and Singh, K. (2021), "Improved performance of machine learning algorithms for prognosis of cervical cancer", *Adv. Comput. Des*., **6**(3), 191-205. http://doi.org/10.12989/acd.2021.6.3.191.
Bata, M., Carriveau, R. and Ting, D.S.K. (2020), "Short-term water demand forecasting using hybrid supervised and unsupervised machine learning model", *Smart Water*, **5**, 2. https://doi.org/10.1186/s40713-020-00020-y.
Benitez, R., Ortiz-Caraballo, C., Preciado, J.C., Conejero, J.M., Figueroa, F.S. and Rubio-Largo, A. (2019), "A short-term databased water consumption prediction approach", *Energies*, **12**(12), 2359. https://doi.org/10.3390/en12122359.
Campos, D.S., Tadano, Y.S., Alves, T.A., Siqueira, H.V. and Marinho, M.H.N. (2020), "Unorganized machines and linear multivariate regression model applied to atmospheric pollutant forecasting", *Acta Sci. Technol.*, **42**(1), e48203. https://doi.org/10.4025/actascitechnol.v42i1.48203.
Candelieri, A., Soldi, D. and Archetti, F. (2015), "Short-term forecasting of hourly water consumption by using automatic metering readers data", *Procedia Eng*., **119**, 844-853. https://doi.org/10.1016/j.proeng.2015.08.948.
Cetinkaya, A. and Baykan, O.K. (2020), "Prediction of middle school students' programming talent using artificial neural networks", *Eng. Sci. Technol*., **23**(6), 1301-1307. https://doi.org/10.1016/j.jestch.2020.07.005.
Davis, J.H. (2011), Multiple linear regression. In: Davis, J.H. (Ed.), *Statistics for Compensation: A Practical Guide to Compensation Analysis*, John Wiley & Sons, New Jersey, U.S.A. https://doi-org /10.1002/9780470946428.ch17.
Digital Government Development Agency (2019), Open government data of Thailand. https://data.go.th.
de Souza Groppo, G., Costa, M.A. and Libanio, M. (2019), "Predicting water demand: A review of the methods employed and future possibilities", *Water Supply*, **19**(8), 2179-2198. https://doi.org/10.2166/ws.2019.122.
Donkor, E.A., Mazzuchi, T.H., Soyer, R. and Roberson, J.A. (2014), "Urban water demand forecasting: Review of methods and models", *J. Water Resour. Plan. Manag*., **140**, 146-159. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000314.
Farizawani, A.G., Puteh, M., Marina, Y. and Rivaie, A. (2020), "A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches", *J. Phys. Conf. Ser*., **1529**(2), 022040. https://doi.org/10.1088/1742-6596/1529/2/022040.
Ghalehkhondabi, I., Ardjmand, E., Young, W.A. and Weckman, G.R. (2017), "Water demand forecasting: Review of soft computing methods", *Environ. Monit. Assess*., **189**, 313. https://doi.org/10.1007/s10661-017-6030-3.
Ghani, I.M.M. and Ahmad, S. (2010), "Stepwise multiple regression method to forecast fish landing", *Procedia Soc. Behav. Sci*., **8**, 549-554. https://doi.org/10.1016/j.sbspro.2010.12.076.
Guo, G., Liu, S., Wu, Y., Li, J., Zhou, R. and Zhu, X. (2018), "Short-term water demand forecast based on deep learning method", *J. Water Resour. Plan. Manag*., **144**(12), 04018076. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000992.

Herrera, M., Torgo, L., Izquierdo, J. and Pérez-Garcìa, R. (2010), "Predictive models for forecasting hourly urban water demand", *J. Hydrol.*, **387**(1-2), 141-150. https://doi.org/10.1016/j.jhydrol.2010.04.005.

Liu, H. and Cocea, M. (2017), "Semi-random partitioning of data into training and test sets in granular computing context", *Granul. Comput*., **2**, 357-386. https://doi.org/10.1007/s41066-017-0049-2.

Mouatadid, S. and Adamowki, J. (2016), "Using extreme learning machines for short-term urban water demand forecasting", *Urban Water J*., **14**(6), 630-638. https://doi.org/10.1515/jwld-2016-0004.

National Statistical Office (2019), Statistical data. http://web.nso.go.th.

National Strategy Secretariat Office (2019), National strategy 2018-2037 (summary). https://www.moac.go.th/pyp-dwl-files-402791791893.

Olyaie, E., Abyaneh, H.Z. and Danandeh Mehr, A.D. (2017), "A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River", *Geosci. Front.*, **8**(3), 517-527. https://doi.org/10.1016/j.gsf.2016.04.007.

Oyebode, O. and Ighravwe, D.E. (2019), "Urban water demand forecasting: A comparative evaluation of conventional and soft computing techniques", *Resources*, **8**(3), 156. https://doi.org/10.3390/resources8030156.

Oyebode, O., Babatunde, D.E., Monyei, C.G. and Babatunde, O.M. (2019), "Water demand modelling using evolutionary computation techniques: Integrating water equity and justice for realization of the sustainable development goals", *Heliyon*, **5**(11), e02796. https://doi.org/10.1016/j.heliyon.2019.e02796.

Pahlavan-Rad, M.R., Dahmardeh, K., Hadizadeh, M., Keykha, G., Mohammadnia, N., Gangali, M., Keikha, M., Davatgar, N. and Brungard, C. (2020), "Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran", *Catena*, **194**, 104715. https://doi.org/10.1016/j.catena.2020.104715.

Romano, M. and Kapelan, Z. (2014), "Adaptive water demand forecasting for near real-time management of smart water distribution systems", *Environ. Model. Softw*., **60**, 265-276. https://doi.org/10.1016/j.envsoft.2014.06.016.

Sahu, K.K., Nayak, S.C. and Behera, H.S. (2021), "Multi-step-ahead exchange rate forecasting for South Asian countries using multi-verse optimized multiplicative functional link neural networks", *Karbala Int. J. Mod. Sci*., **7**(1), 48-60. https://doi.org/10.33640/2405-609X.2278.

Sattar, A.M.A., Ertugrul, O.F., Gharabaghi, B., McBean, E.A. and Cao, J. (2019), "Extreme learning machine model for water network management", *Neural Comput. Appl.*, **31**(1), 157-169. https://doi.org/10.1007/s00521-017-2987-7.

Seo, Y., Kwon, S. and Choi, Y. (2018), "Short-term water demand forecasting model combining variational mode decomposition and extreme learning machine", *Hydrology*, **5**(4), 54. https://doi.org/10.3390/hydrology5040054.

Sirisathitkul, Y., Thanathamathee, P. and Aekwarangkoon, S. (2019), "Predictive apriori algorithm in youth suicide prevention by screening depressive symptoms from patient health questionnaire-9", *TEM J.*, **8**(4), 1449-1455. https://dx.doi.org/10.18421/TEM84-49.

Smolak, K., Kasieczka, B., Fialkiewicz, W., Rohm, W., Sila-Nowicka, K. and Kopanczyk, K. (2020), "Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models", *Urban Water J*., **17**(1), 32-42. https://doi.org/10.1080/1573062X.2020.1734947.

Stampoulis, D., Damavandi, H.G., Boscovic, D. and Sabo, J. (2021), "Using satellite remote sensing and machine learning techniques towards precipitation prediction and vegetation classification", *J. Environ. Inform*., **37**(1), 1-15. https://doi.org/10.3808/jei.202000427.

Thai Meteorological Department (2019), Thailand meteorological department API: TMDAPI in Thailand climate standard normal value 1981-2010. https://data.tmd.go.th/api/index1.php.

Thinakaran, K., Rajasekar, R., Santhi, K. and Nalini, M. (2020), "Predicting the 2-dimensional airfoil by using machine learning methods", *Adv. Comput. Des*., **5**(3), 291-304. http://doi.org/10.12989/acd.2020.5.3.291.

UN Water (2020), UN World Water Development Report 2020. https://www.unwater.org/publications/world-water-development-report-2020.

Wei, H.W., Hassan, M., Che, Y., Peng, Q.K., Wang, Q., Su, Y.L. and Xie, B. (2021), "Spatio-temporal characteristics and source apportionment of water pollutants in upper reaches of Maotiao River, Southwest of China, from 2003 to 2015", *J. Environ. Inform.*, **37**(2), 93-106. http://doi.org/10.3808/jei.201900415.

Wongso, E., Nateghi, R., Zaitchik, B., Quiring, S. and Kumar, R. (2020), "A data-driven framework to characterize state-level water use in the United States", *Water Resour. Res.*, **56**(9), e2019WR024894. https://doi.org/10.1029/2019WR024894.

Wu, L. and Zhou, H. (2010), "Urban water demand forecasting based on HP filter and fuzzy neural network", *J. Hydroinformatics*, **12**(2), 172-184. https://doi.org/10.2166/hydro.2009.082.

Xiao, Y., Fang, L. and Hipel, K.W. (2021), "Conservation-targeted hydrologic-economic models for water demand management", *J. Environ. Inform*., **37**(1), 49-61. http://doi.org/10.3808/jei.201900418.

Yang, Y., Huang, T.T., Shi, Y.Z., Wendroth, O. and Liu, B.Y. (2021), "Comparing the performance of an autoregressive state-space approach to the linear regression and artificial neural network for streamflow estimation", *J. Environ. Inform*., **37**(1), 36-48. https://doi.org/10.3808/jei.202000440.

Zhong, H., Wang, J., Jia, H., Mu, Y. and Lv, S. (2019), "Vector field-based support vector regression for building energy consumption prediction", *Appl. Energy*, **242**, 403-414. https://doi.org/10.1016/j.apenergy.2019.03.078.

Zubaidi, S.L., Dooley, J., Al-Khaddar, R., Abdellatif, M., Al-Bugharbee, H. and Ortega-Martorell, S. (2018), "A novel approach for predicting monthly water demand by combining singular spectrum analysis with neural networks", *J. Hydrol*., **561**, 136-145. https://doi.org/10.1016/j.jhydrol.2018.03.047.

Zubaidi, S.L., Ortega-Martorell, S., Al-Bugharbee, H., Olier, I., Hashim, K.S., Gharghan, S.K., Kot, P. and Al-Khaddar, R. (2020a), "Urban water demand prediction for a city that suffers from climate change and population growth: Gauteng province case study", *Water*, **12**(7), 1885. https://doi.org/10.3390/w12071885.

Zubaidi, S.L., Ortega-Martorell, S., Kot, P., Al-Khaddar, R., Abdellatif, M., Gharghan, S.K., Ahmed, M.S. and Hashim, K.S. (2020b), "A method for predicting long-term municipal water demands under climate change", *Water Resour. Manag*., **34**(3), 1265-1279. https://doi.org/10.1007/s11269-020-02500-z.

*TK*