water

MDPI

Article

Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow

Jiyeong Hong [1], Seoro Lee [1], Joo Hyun Bae [2], Jimin Lee [1], Woon Ji Park [1], Dongjun Lee [1], Jonggun Kim [1] and Kyoung Jae Lim [1,*]

[1] Department of Regional Infrastructure Engineering, Kangwon National University, Chuncheon-si 24341, Korea; jiyeong.hong.1@gmail.com (J.H.); seorolee91@gmail.com (S.L.); jiminlee217@gmail.com (J.L.); pwj98@kangwon.ac.kr (W.J.P.); dj90lee@gmail.com (D.L.); kimjg23@gmail.com (J.K.)

[2] Korea Water Environment Research Institute, Chuncheon-si 24408, Korea; baegop@pusan.ac.kr

* Correspondence: kjlim@kangwon.ac.kr

Received: 3 September 2020; Accepted: 15 October 2020; Published: 20 October 2020

Abstract: Predicting dam inflow is necessary for effective water management. This study created machine learning algorithms to predict the amount of inflow into the Soyang River Dam in South Korea, using weather and dam inflow data for 40 years. A total of six algorithms were used, as follows: decision tree (DT), multilayer perceptron (MLP), random forest (RF), gradient boosting (GB), recurrent neural network–long short-term memory (RNN–LSTM), and convolutional neural network–LSTM (CNN–LSTM). Among these models, the multilayer perceptron model showed the best results in predicting dam inflow, with the Nash–Sutcliffe efficiency (NSE) value of 0.812, root mean squared errors (RMSE) of 77.218 m^3/s, mean absolute error (MAE) of 29.034 m^3/s, correlation coefficient (R) of 0.924, and determination coefficient (R^2) of 0.817. However, when the amount of dam inflow is below 100 m^3/s, the ensemble models (random forest and gradient boosting models) performed better than MLP for the prediction of dam inflow. Therefore, two combined machine learning (CombML) models (RF_MLP and GB_MLP) were developed for the prediction of the dam inflow using the ensemble methods (RF and GB) at precipitation below 16 mm, and the MLP at precipitation above 16 mm. The precipitation of 16 mm is the average daily precipitation at the inflow of 100 m^3/s or more. The results show the accuracy verification results of NSE 0.857, RMSE 68.417 m^3/s, MAE 18.063 m^3/s, R 0.927, and R^2 0.859 in RF_MLP, and NSE 0.829, RMSE 73.918 m^3/s, MAE 18.093 m^3/s, R 0.912, and R^2 0.831 in GB_MLP, which infers that the combination of the models predicts the dam inflow the most accurately. CombML algorithms showed that it is possible to predict inflow through inflow learning, considering flow characteristics such as flow regimes, by combining several machine learning algorithms.

Keywords: dam inflow; decision tree; multilayer perceptron; random forest; gradient boosting; RNN–LSTM; CNN–LSTM

1. Introduction

Global warming has led to concerns over climate change and caused the complexity of the hydrologic cycle, resulting in greater uncertainty in the management of water resources [1]. Especially in Korea, it is important to establish a plan for water resource management through efficient water management, because of the high coefficient of flow fluctuation and the steep geographical characteristics [2]. Korea has continued to make programs for watershed planning and sustainable water resource management by improving the operational efficiency of hydraulic structures such as

reservoirs and multi-purpose dams [3]. However, changes in dam inflow patterns, caused by climate change, are causing difficulties in using stable water resources and establishing supply plans [4]. Therefore, it is essential to predict the dam inflow in order to establish a dam operation plan for future climate change.

Generally, various hydrological models, such as the Soil and Water Assessment Tool (SWAT), the Hydrological Simulation Program—Fortran (HSPF), and the watershed-scale Long-Term Hydrologic Impact Assessment Model (watershed-scale L-THIA), have been developed and utilized to predict river discharge and dam inflow [5–7]. The hydrological model is one of the most popular methods to predict water cycle components by reflecting the physical mechanisms of the hydrological process in the watersheds. However, due to the complicated mechanism of hydrological modeling, these models require detailed data, complex factors, and a longer computation time [8]. Also, hydrological modeling requires professional skill to run and calibrate the modeling systems, and the parameters in the hydrological model and the uncertainties arising in the course of the physical process have limitations to take into account future trends of temporal and spatial variability for the accurate prediction of the dam inflow [9]. In the case of Soyang River Dam, it is difficult to perform the proper simulation; due to the watershed covering North Korean territory, the hydrological models, such as SWAT, cannot estimate dam inflow with great accuracy, due to the scarcity of the data [10]. For these problems, the estimation models using simple data, such as statistical models or machine learning models, can be a solution.

Statistical models and the artificial intelligence models (AI models) are also used for the prediction of river discharge and dam inflow, by analyzing the correlation between the components. In particular, many typical time series models are used for hydrologic predictions, such as autoregressive–moving-average (ARMA) and autoregressive integrated moving average (ARIMA), based on regression [11]. The effects of climate change on streamflow in a glacier mountain catchment were analyzed using an ARMA model [12]. The ARIMA model was used to forecast streamflow and hydrological drought trend in Cyprus [13]. The artificial neural network (ANN) has been used to develop a model that predicts the flow of reservoirs up to six months in advance by learning, verifying, and evaluating the inflow datasets of Egypt's Aswan High Dam over 130 years [14]. Also, the comparison between ARMA, ARIMA, and Artificial Neural Network (ANN) has been performed by several researchers [15]. Machine learning techniques were used to manage the Rawal Dam, and the results showed that the J48 tree classification technique gave the best results in the management of the discharge, the improved support vector machine (SVM) method predicts the most accurate critical parameters for water level estimation, and regression techniques were found to be the most accurate in estimating water storage capacity [16]. Also, the use of the ANN-based deep neural network (DNN) is increasing, due to the academic and practical advantages for research. In addition, the long short-term memory (LSTM) neural network, one of DNN's state-of-the-art applications, has been successfully applied in various fields [17,18]. The use of convolutional neural network (CNN)–LSTM, which combines CNN with LSTM, has also been proven to be excellent in forecasting particulate matter 2.5 ($PM_{2.5}$) [19]. Although algorithms are used in various fields (e.g., atmosphere science, hydrology, etc.), there is a limit to finding optimization values through individual algorithms. To improve these problems, a combination of the algorithms has been used for air pollutant forecasts in Athens [20].

For streamflow estimation, lots of researches using machine learning algorithms have been carried out. For example, machine learning models simulated and forecasted the streamflows, with the comparison of with and without baseflow separation, and the results showed that the base flow separation improved the model accuracy [21]. The comparison of a Gaussian linear regression model (GLM), Gaussian generalized additive models (GAMs), multivariate adaptive regression splines (MARS), ANN, random forest (RF), and M5 models was analyzed to highlight the strengths and limitations of each of the models, and the results showed that GAM showed high Nash–Sutcliffe efficiency (NSE) performance, but showed a rapid increase of uncertainty with high temperatures [22]. Lead time daily streamflow was forecasted using ANN and LSTM, and the results showed that peak

flows were predicted more accurately than the low and normal flows [23]. Since the single algorithms cannot be perfect for streamflow estimation, as the weather and flow regimes vary greatly in Korea, new algorithms considering these various characteristics need to be selected and developed.

To reflect the characteristics of the catchments in South Korea, the dynamic flow regimes due to the temporal distribution of rainfall should be considered. Therefore, this study aims to (1) evaluate the performance of the algorithm to predict the amount of inflow of the Soyang River Dam, and (2) develop and evaluate the combined machine learning algorithms, with the consideration of flow duration.

2. Methodology

2.1. Study Area

The Soyang River Dam is located in the Han River Basin in South Korea (Figure 1). The Soyang River Dam is a multipurpose dam for water supply and power generation, with an effective storage capacity of 2,900,000,000 tons. The hydrologic catchment area of the dam is 2637 km^2, of predominantly forest area (89.5%), and the remaining land uses are agricultural land (5.7%), water (2.4%), and other land uses (2.4%), which covers administrative districts (Chuncheon, Inje, Yanggu, and Goseong) and some of the North Korean territory. According to the precipitation data collected at the Chuncheon weather station, annual rainfall from 1980 to 2019 is, on average, 1321 mm, and has varied between 677 mm and 2069 mm with an increasing pattern, as shown in Figure 2.

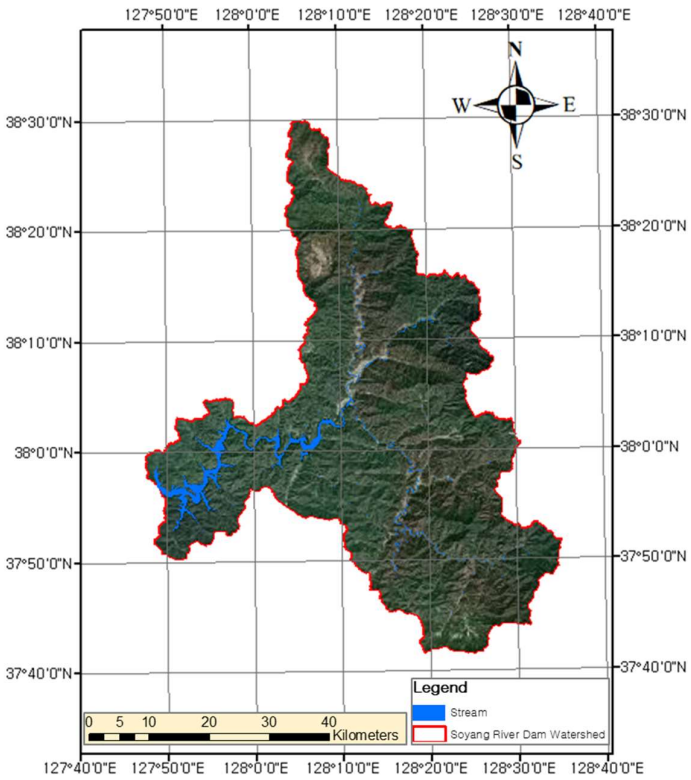


Figure 1. The study area of Soyang River Dam.

The Soyang River Dam is the main water source for the metropolitan area. The inflow from Inbuk Stream and Naerin Stream are the main tributaries to the Soyang River Dam, and the flow rate at the confluence accounts for more than 90% of the total inflow of the Soyang River Dam [24]. The discharge of Soyang River Dam, which directly affects the water environment at the downstream, is carried out during water level control before the flood, water supply in drought season, and power generation in the dam. Between 1980 and 2019, the average flow and the peak flow recorded at the Soyang River Dam were 68.6 m^3/s and 7405.6 m^3/s with a decreasing pattern, as shown in Figure 2. The increasing tendency

of precipitation and the decreasing tendency of inflow are caused by the increase in evapotranspiration due to the increasing tendency of the annual average of the maximum temperature.

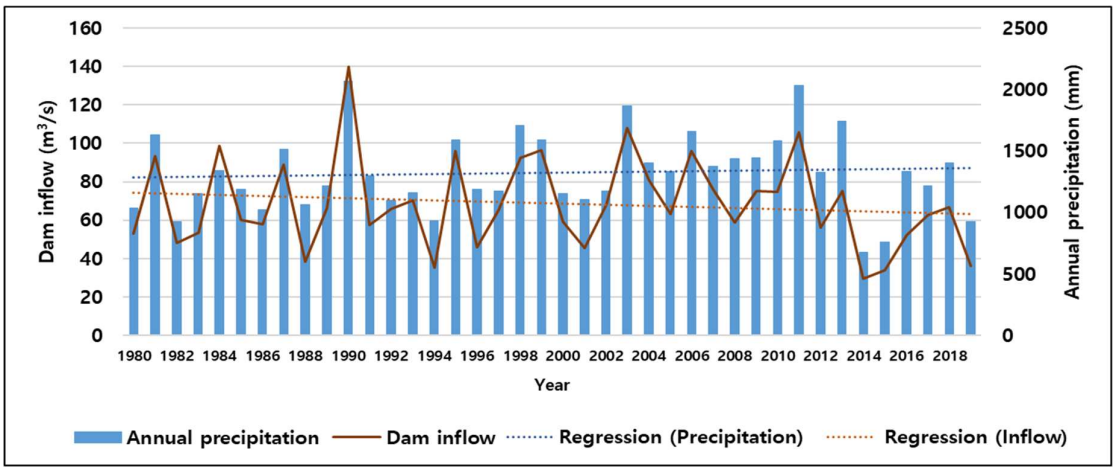


Figure 2. Annual precipitation and mean annual dam inflow during 1980–2019 of Soyang River Dam.

South Korea suffered the worst droughts from 2014 to 2015, receiving less than 43% of the annual precipitation average of the past 30 years [25]. Predicting floods and droughts can reduce damage caused by disasters, and enable efficient water management.

2.2. Data Descriptions

The machine learning prediction models were built for inflow estimation flowing into the Soyang River Dam, with the use of inflow data and weather data. The machine learning prediction models were constructed for a learning period of 40 years (1980–2019). The time series data of the weather (precipitation, maximum temperature, minimum temperature, humidity, wind speed, and solar radiation) were extracted from the Korea Meteorological Administration [26] database for the Chuncheon observation station located the nearest to the Soyang River Dam. The inflow data of the Soyang River Dam were obtained from the Korea Water Resources Corporation (K-water) [27]. The time series data used for prediction of the inflow of the dam are shown in Figure 3. Figure 3 represents 40 years of the variation in the amount of dam inflow (m^3/s), precipitation (mm), maximum temperature (°C), minimum temperature (°C), humidity (%), wind speed (m/s), and solar radiation (MJ/m^2). Figure 2 also shows that the patterns of precipitation and dam inflow are similar.

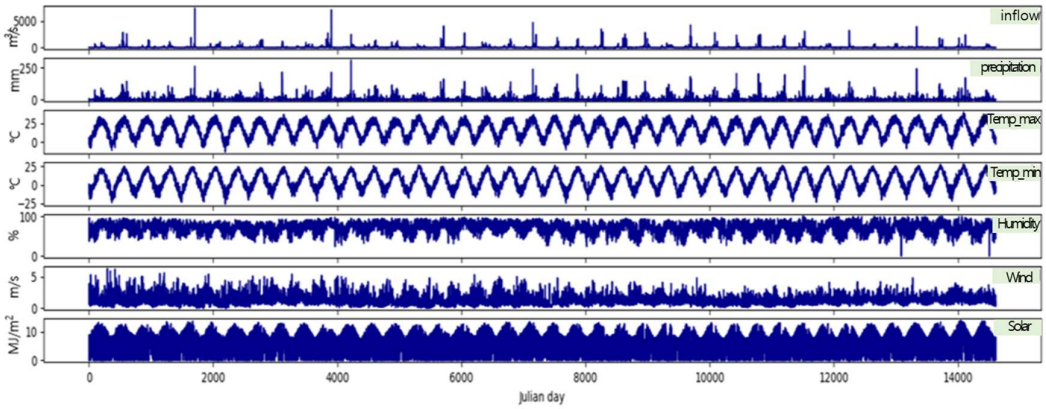


Figure 3. Time series variation of dam inflow and weather data.

In this study, the machine learning prediction model for the dam inflow of a day used the following data: weather data of the day (forecasted), weather data of one day ago, weather data of two days

ago, the inflow of one day ago, and the inflow of two days ago, for the study period of 40 years (1980–2019) (Table 1). The weather data includes precipitation, maximum temperature, minimum temperature, relative humidity, wind speed, and solar radiation. In addition, prior weather conditions in the basin have a significant impact on soil moisture in the basin [28]; therefore, weather data and flow data of one day ago and two day ago were used for machine learning, to take into account prior weather conditions.

Table 1. The input data for machine learning models.

	Input Variable	Output Variable
Weather Data of Two Days Ago	Inflow (t − 2), precipitation (t − 2), temp_max (t − 2), temp_min (t − 2), humidity (t − 2), wind (t − 2), solar (t − 2)	Inflow of the day: Inflow (t)
Weather Data of One Day Ago	Inflow (t − 1), precipitation (t − 1), temp_max (t − 1), temp_min (t − 1), humidity (t − 1), wind (t − 1), solar (t − 1)	
Weather Data of the Day (Forecasted)	precipitation (t), temp_max (t), temp_min (t), humidity (t), wind (t), solar (t)	

Note: The 'inflow', 'precipitation', 'temp_max', 'temp_min', 'humidity', 'wind, solar', '(t − 2)', '(t − 1)', and 't' are dam flow, precipitation, maximum temperature, minimum temperature, humidity, solar radiation, two days ago, one day ago, and the day, respectively.

Data preprocessing, such as to improve the quality of data and to generate comprehensible information, needs to be done for effective machine learning [29]. Data preprocessing includes a selection of input variables, standardization, noise instance removal, data dimension reduction, and multiple collimations. Several studies revealed that more data preprocessing makes better predictive performance of the model [30,31]. In this study, data preprocessing was performed by the scaling method and standardization methods, including shape scaling, normalization, and standardization using the 'StandardScaler' function, which is one of the most commonly used scalers of the 'sklearn.preprocessing' library for a preprocessing step [32,33]. This step applies to a dataset an average of 0 and variance 1, by applying linear transformations to all data.

2.3. Machine Learning Algorithms

Machine learning is an algorithm that learns from data and improves its performance as it learns. Machine learning is classified among three separate branches: supervised learning, unsupervised learning, and reinforcement learning [29,34]. In this study, supervised learning was used to predict the inflow of the Soyang River Dam. Supervised learning uses labeled data (e.g., precipitation, temperature, inflow, outflow, etc.) for training, and infers a function from the data. A total of six methods, including RNN–LSTM and CNN–LSTM, were used to build models to estimate the amount of dam inflow. Model information is given in Table 2.

Table 2. The description of machine learning models.

Machine Learning Models	Module	Function	Notation
Decision Tree	sklearn.tree	DecisionTreeRegressor	DT
Multilayer Perceptron	sklearn.neural_network	MLPRegressor	MLP
Random Forest	sklearn.ensemble	RandomForestRegressor	RF
Gradient Boosting	sklearn.ensemble	GradientBoostingRegressor	GB
RNN–LSTM	keras.models.Sequential	LSTM, Dense, Dropout	LSTM
CNN–LSTM	keras.models.Sequential	LSTM, Dense, Dropout, Conv1D, MaxPooling1D	CNN–LSTM

Decision tree, multilayer perceptron, random forest, and gradient boosting used regression functions in the scikit-learn library of Python, while RNN–LSTM and CNN–LSTM used sequential

functions of Keras modules in the TensorFlow library. "Notation" refers to the name used when briefly stated in the graph.

Decision Tree

A decision tree is a widely used model for classification and regression, and it learns as it continues to ask yes or no questions to reach a decision [35]. In the decision tree, the hyperparameter that controls model complexity is a prepruning parameter that causes the tree to stop before it is completely created. In general, the designation of either "max_depth", "max_leaf_nodes", or "min_samples_leaf" is sufficient to prevent overfitting. The decision tree in the scikit-learn is intended to provide adequate prefabrication through "min_samples_leaf". The critical hyperparameters in the decision tree (DT) regressor are the following: entropy for criterion, 1 for min_samples_leaf, 0 for min_impurity_decrease, best for splitter, 2 for min_samples_split, and 0 for random_state (Table 3).

Table 3. The critical hyperparameters in nonlinear regression machine learning algorithms.

Decision Tree Regressor		**MLP Regressor**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
Criterion	Entropy	hidden_layer_sizes	(50,50,50)
Min_Samples_Leaf	1	solver	Adam
Min_Impurity_Decrease	0	learning_rate_init	0.001
Splitter	Best	max_iter	200
Min_Samples_Split	2	momentum	0.9
Random_State	0	beta_1	0.9
		epsilon	1×10^{-8}
		activation	relu
Random Forest Regressor		**Gradient Boosting Regressor**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
n_Estimators	50	Loss	ls
Min_samples_split	2	n_estimators	100
Min_Weight_Fraction_Leaf	0	criterion	friedman_mse
Min_Impurity_Decrease	0	min_samples_leaf	1
Verbose	0	max_depth	10
Criterion	Mse	alpha	0.9
Min_Samples_Leaf	1	presort	Auto
Max_Features	Auto	tol	1×10^{-4}
Bootstrap	True	learning_rate	0.1
		subsample	1.0
		min_samples_split	2
		validation_fraction	0.1

Multilayer Perceptron

A multilayer perceptron (MLP) is one of the feed-forward neural network (FFNN) structures, consisting of a total of three layers: input layer, hidden layer, and output layer [36]. The input data is entered at the input layer, weighted to fit the set hidden layer structure, and the results are printed out at the output layer [36]. Recently, MLPs configured with more than one hidden layer have produced more accurate predictions than other machine learning techniques [37]. Table 3 provides details of the setting of hyperparameters for the MLPRegressor function. The critical hyperparameters in the MLP regressor are the following: 50 nodes for each of the three layers for hidden_layer_sizes, adam for solver, 0.001 for learning_rate_init, 200 for max_iter, 0.9 for momentum, 0.9 for beta_1, 1e−8 for epsilon, and relu for activation.

Random Forest

Random forest is a classification technique, developed by Breiman [38], that combines the bagging algorithm, the ensemble learning method, and the classification and registration tree (CART) algorithm. Random forest is also executable on large-scale data and provides high accuracy because it runs

using many variables without removing them [37,39]. In addition, compared to the artificial neural network and support vector regression, the hyperparameter is simple for detailed tuning. Of the hyperparameters set in the function "RandomForestRegressor" used in this study, the sensitive hyperparameter is the primary parameter. Since the random forest model is an ensemble model of the decision tree, the primary parameter is the tree number, "n-estimator", and the value is set to 50. The other variables are as the following: 2 for min_samples_split, 0 for min_weight_fraction_leaf, 0 for min_impurity decrease, 0 for verbose, mse for criterion, 1 for min_samples_leaf, auto for max_features, and true for bootstrap (Table 3).

Gradient Boosting

Gradient boosting is an ensemble model that learns the boosting algorithm by ensemble learning for the decision tree. In gradient boosting, the gradient reveals the weaknesses of the model that have been learned so far, while other models focus on it to boost performance. The parameters that minimize the loss function that quantifies errors in the predictive model should be found for better prediction. The advantage of gradient boosting is that the other loss functions can be used as much as possible. The character of the loss function is automatically reflected in learning through the gradient [40].

The hyperparameters of the "GradientBoostingRegressor" function set out in this study were given as the following: 1s for loss, 100 for n_estimators, friedman_mse for criterion, 1 for min_samples_leaf, 10 for max_depth, 0.9 for alpha, auto for presort, 1×10^{-4} for tol, 0.1 for learning_rate, 1.0 for subsample, 2 for min_samples_split, and 0.1 for validation_fraction (Table 3).

LSTM

LSTM is an RNN, and RNN is a type of deep learning algorithm that learns time series data repeatedly [41]. RNN is a structure in which the output data of the previous RNN, in the course of data learning, affects the output data of the current RNN. This allows for the connection of current and past learning and is useful for continuous and repetitive learning; however, the predictive performance is compromised with use of data from the past too far. LSTM is a type of RNN-based deep learning algorithm that makes it easy to predict time series data by taking into account the order or time aspects of learning, and by preventing chronic problems of weight loss in RNN [42,43]. Several recent studies have shown that LSTM transforms its structure to improve its predictive performance [15]. In this study, RNN–LSTM built the layers of "LSTM" and "dense", and put "dropout" layer in the middle to prevent overfitting (Figure 4).

CNN–LSTM

As the performance of deep learning has recently been verified throughout data science and technology, it is believed that the deep neural network (DNN), a deep learning technique to solve the problem of numerical prediction, can also contribute to improving the accuracy of the calculation of dam inflow. CNN–LSTM, which is joined by a leading algorithm with CNN, is one of the examples of LSTM transformation [22,44,45].

Information on layers added to the sequential functions of CNN–LSTM in this study is shown in Figure 4, consisting of seven layers. In the case of CNN–LSTM, CNN, which uses two-dimensional data mainly, can be used for one-dimensional time series data to extract data characteristics and analyze data prediction. Additional "Conv1D" and "MaxPooling1D" were used to construct layers for CNN–LSTM.

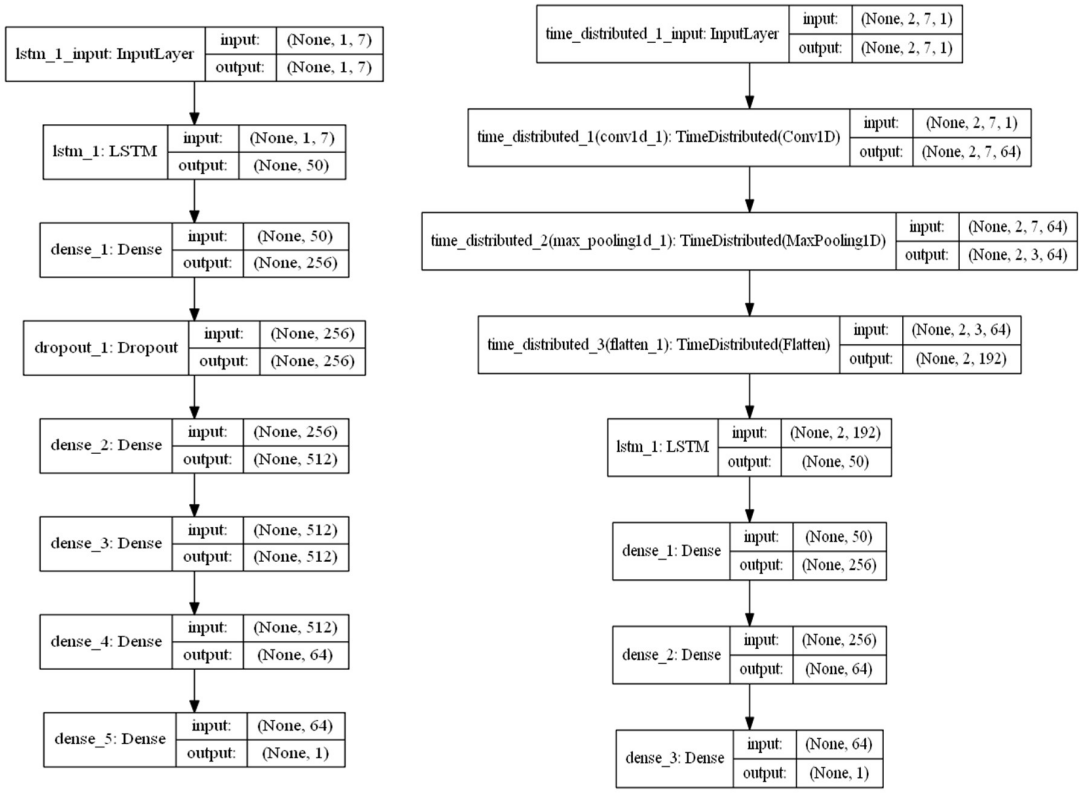


Figure 4. Illustration of the proposed (**a**) recurrent neural network–long short-term memory (RNN–LSTM); and (**b**) convolutional neural network–long short-term memory CNN–LSTM network for dam inflow prediction.

2.4. Development of Combined Machine Learning Algorithms (CombML)

As indicated in the objective of this study, we predicted the inflow of Soyang River Dam using the past weather data, inflow, and the weather forecast, accurately. The results of the study showed that the ensemble models (the random forest and gradient boosting) performed well in under 100 m^3/s of the inflow. On the other hand, MLP has merits when predicting the inflow of over 100 m^3/s.

Therefore, a new model, combining MLP and ensemble models, was created to predict the dam inflow; however, it is impossible to predict the inflow of the next day. Therefore, the forecasted precipitation data, which were shown to have the highest correlation with the dam inflow by the heatmap analyzing, were used as a standard for the new ensemble model. The reference point was set by averaging rainfall on days with dam inflow greater than 100 m^3/s. The average precipitation of the filtered dam inflow was 16 mm.

2.5. Model Training Test

In this study, the inflow of Soyang River Dam was predicted using weather data and inflow data during the period 1980 to 1919. The model training period was from 1980 to 2016, and the test period was from 2017 to 2019.

To assess the performance of each machine learning model, Nash–Sutcliffe efficiency (NSE), root mean squared errors (RMSE), the mean absolute error (MAE), correlation coefficient (R), and determination coefficient (R^2) were used. Numerous studies indicated the appropriateness of these measures to assess the accuracy of hydrological models [46–48]. NSE, RMSE, MAE, R, and R^2 for evaluation of the model accuracy can be calculated from Equations (1)–(5).

$$NSE = 1 - \frac{\sum (O_t - M_t)^2}{\sum (O_t - \overline{O_t})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum (O_t - M_t)^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n} \sum \left| M_t - O_t \right| \quad (3)$$

$$R = \frac{\sum \left(O_t - \overline{O_t} \right) \left(M_t - \overline{M_t} \right)}{\sqrt{\sum (O_t - \overline{O_t})^2 \sum (M_t - \overline{M_t})^2}} \quad (4)$$

$$R^2 = \frac{\left[\sum \left(O_t - \overline{O_t} \right) \left(M_t - \overline{M_t} \right) \right]^2}{\sum (O_t - \overline{O_t})^2 \sum (M_t - \overline{M_t})^2} \quad (5)$$

where O_t is the actual value of t, $\overline{O_t}$ is the mean of the actual value, M_t is the estimated value of t, $\overline{M_t}$ is the mean of the estimated value, and n is the total number of times.

RMSE is the standard deviation of the residuals, and MAE is the mean of the absolute values of the errors. Therefore, the closer the verification values are to zero, the more similar the observed and the model values are.

R, the correlation coefficient, represents the magnitude of the correlation; R values are +1 if the observed and simulated values are the same, 0 if they are completely different, and −1 if they are completely the same in the opposite direction. The R^2 compares the propensity of the observed to the simulated value.

3. Results and Discussion

3.1. Impact Factor Analysis

In this study, heatmap analysis was used to evaluate the correlation of the data used (Figure 5). The evaluation showed the correlation coefficient of 0.59 as the highest correlation between precipitation and dam inflow, followed by minimum temperature, humidity, maximum temperature, wind, and solar, which infers that the characteristics of precipitation have the most robust effects on the dam inflow.

Figure 5. Heat map to analyze correlation coefficients of model input data in Soyang River Dam.

Figure 6 shows the feature importance of each input data used in the classification of the decision tree model. The feature importances of precipitation, maximum temperature, minimum temperature, humidity, wind speed, and solar radiation, which are the weather data used to predict the inflow of dams, are 0.549, 0.097, 0.150, 0.073, 0.078, and 0.054, respectively. Regarding the feature importance on the Soyang River Dam inflow prediction, precipitation indicates the highest importance for predicting dam inflow, whereas the other input data indicate less importance for predicting dam inflow.

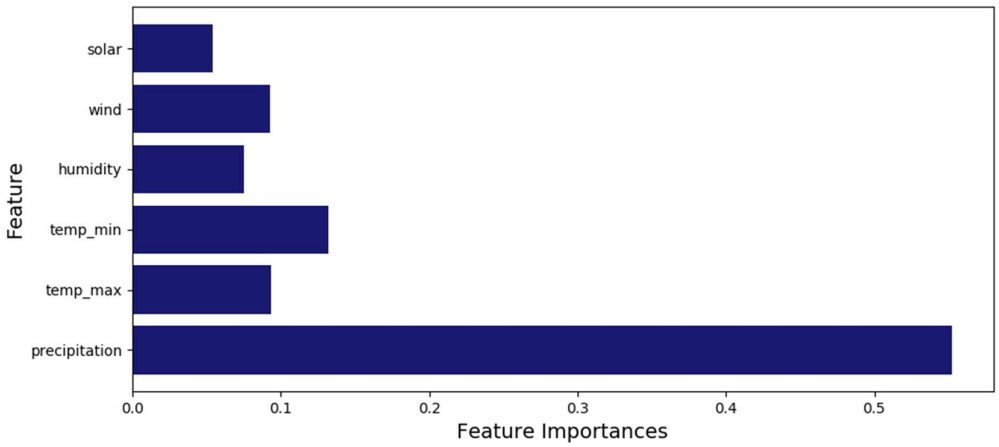


Figure 6. Feature importances in the decision tree model.

3.2. Prediction Results Using Machine Learning Algorithms

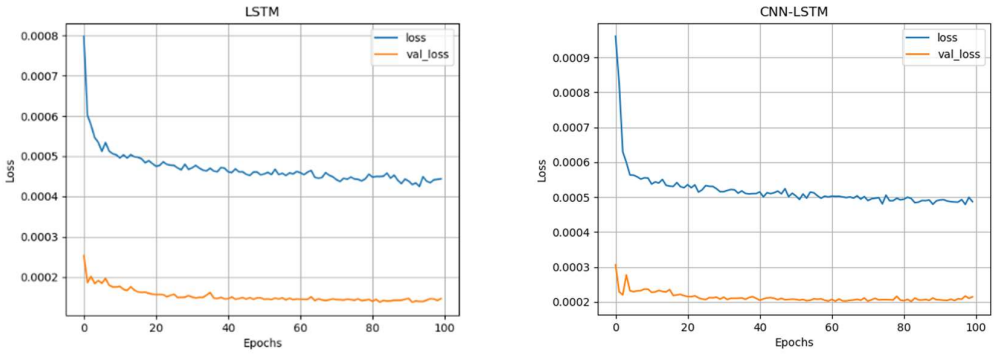
Table 4 shows the prediction accuracy results (NSE, RMSE, MAE, R, and R^2) of six machine learning models, by comparing the predicted dam inflow to the observed dam inflow. The result from the MLP model showed high prediction accuracy, with NSE of 0.812, RMSE of 77.218 m^3/s, MAE of 29.034 m^3/s, R of 0.924, and R^2 of 0.817. MLP is the most widely used and powerful model of supervised learning as a predictive algorithm [29,49]. The prediction of the dam inflow using MLP was more accurate than using deep learning, which had more layers. Some previous studies have shown that a simpler neural network model, such as MLP, can perform much better than more complex models, such as deep-learning stacked autoencoder (SAE) [50]. Due to the characteristics of the data, performing the formation of deep layers leads to overfitting, which, in turn, prevents the actual prediction from reducing the loss.

Table 4. Prediction accuracy results of six machine learning models.

Method	NSE	RMSE (m^3/s)	MAE (m^3/s)	R	R^2
Decision Tree	0.589	114.04	27.333	0.775	0.601
MLP	0.812	77.218	29.034	0.904	0.817
Random Forest	0.745	89.73	20.372	0.867	0.753
Gradient Boosting	0.718	94.486	20.522	0.848	0.718
LSTM	0.429	134.329	26.332	0.675	0.455
CNN–LSTM	0.455	131.243	35.921	0.694	0.482

The second-best predicted models are the ensemble models, the random forest, and the gradient boosting models. The results from the random forest and the gradient boosting model showed prediction accuracy with NSE of 0.745 and 0.718, respectively, and R^2 of 0.753 and 0.718, respectively, which indicates that the models are the most effective, except MLP, in predicting dam inflow. On the other hand, CNN–LSTM (LSTM combined with CNN) has prediction accuracy, with NSE of 0.455. It is slightly better than the RNN–LSTM, with NSE of 0.429, but less predictable than other machine learning.

Figure 7 shows the loss variation calculated by a mean squared error of the learning and validation material, given 100 epochs in deep learning RNN–LSTM and CNN–LSTM, and shows that verification loss falls to 0.0001. Figure 7 clearly indicates that the data can be trained enough, with about 60 epochs.



(**a**) (**b**)

Figure 7. Comparison of training loss and validation loss of (**a**) long short-term memory (LSTM) and (**b**) CNN–LSTM model.

Due to the limitation of long-term simultaneous analysis by one graph, the results from six models were analyzed separately in Figure 8. On 5 July 2016, the 187th day of Julian Day, the peak inflow of 3918.5 m^3/s was recorded, and on 29 August 2018, the 972nd day, the second-highest inflow of 2380.8 m^3/s was recorded, between 2016 and 2019. The prediction models for dam inflow failed to predict inflow accurately when an excessive flow had flowed into the dam during the evaluation period, due to the scarcity of training data for the intensive flood. On the other hand, for dam inflow of 500 m^3/s or less, it can be seen that the trend is generally well matched.

Figure 8 shows that the prediction by (b) MLP of the nonlinear regression models and (e) LSTM of deep learning models best represents the first peak flow of the time series variation of the observed values; and the predictive changes by the (a) decision tree and (d) gradient boosting models can be seen to be overestimated, relative to the observed values. Among the other techniques, the MLP model performed the best, representing the time series variation of observed values comprehensively.

Due to the difficulties of analyzing residuals from the observation in time series analysis of dam inflows, Figure 9 illustrates X–Y plots of the values predicted for flow rates less than 100 m^3/s, compared to the observed values by models. Of the daily inflow of Soyang River Dam, 87.15% is less than 100 m^3/s, and most of the observed inflow data are included in this range, unless it is heavy rainfall. Figure 10 showed that DT, RF, and GB predict the dam inflow appropriately; however, MLP and CNN–LSTM tend to overestimate the dam inflow, and LSTM performed the least accurately among the algorithms. Figure 10 illustrates the dam inflow predicted for flow rates above 100 m^3/s. Figure 10 shows that the distribution was significantly different from the observed values, compared to the distributions below 100 m^3/s with these and predicted results by (c) random forest and (d) gradient boosting models in Figure 9, and the distribution of less residual difference with the observed values. In other words, it is seen that the amount of dam inflow below 100 m^3/s is well predicted by the ensemble models.

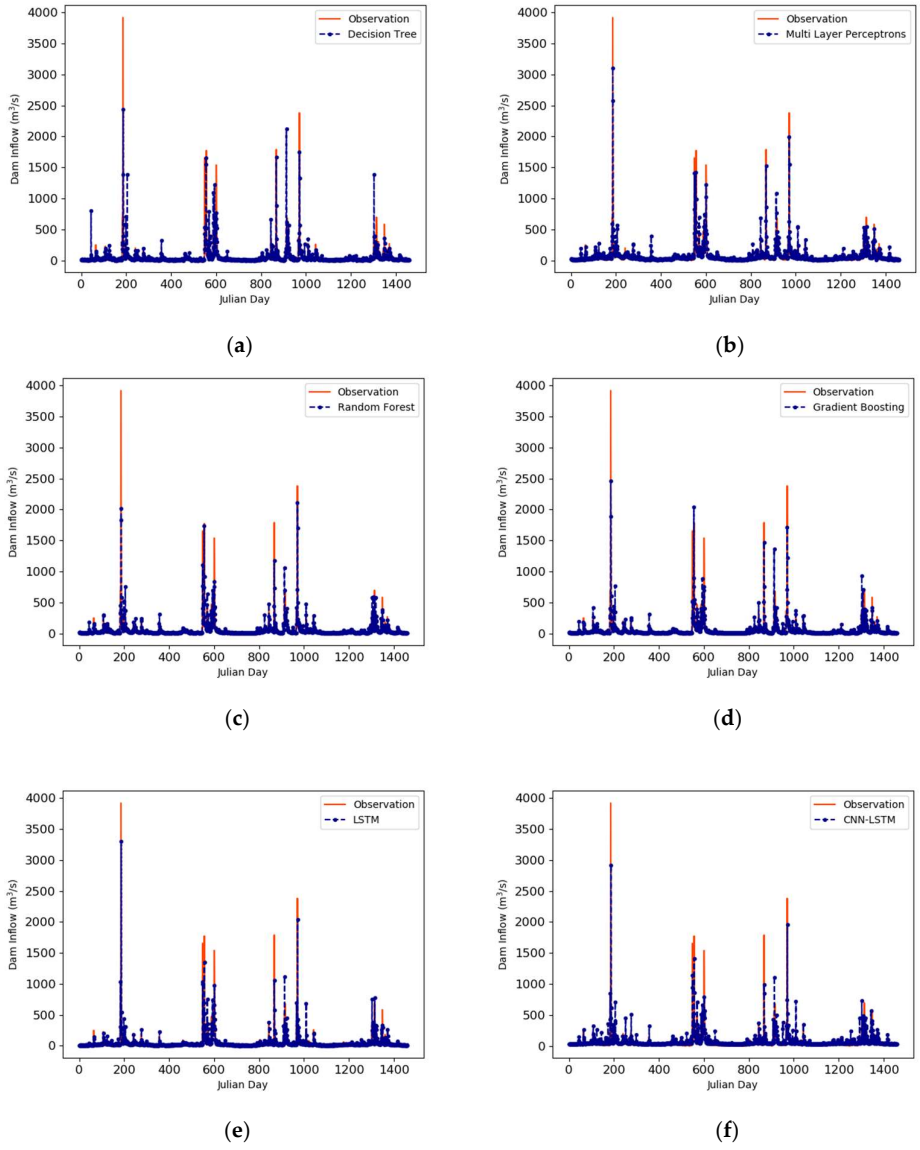


Figure 8. The comparisons of forecasting results using machine learning of (**a**) decision tree, (**b**) multilayer perceptrons, (**c**) random forest, (**d**) gradient, (**e**) LSTM, and (**f**) CNN–LSTM.

In order to predict the amount of dam inflow during heavy rain, the significant changes in the dam inflow should be predicted at the heavy rainfall events. Figure 10 illustrates the comparison of simulated dam inflow and observed dam inflow only if the observation of dam inflow is greater than 100 m^3/s. Compared to the model performances when the dam inflow is under 100 m^3/s, all of the six models had worse results. Among the models, MLP showed the best matching prediction with the observed dam inflow; meanwhile CNN–LSTM model captured well the magnitude of peaks under 700 m^3/s, however, the model failed to capture the pattern of peaks over 700 m^3/s. The poor accuracy of the prediction for the peaks can be caused by the scarcity of the training data, since the heavy rainfall events occur rarely.

Although the frequency is low, flood prediction is necessary to prepare for the flood. In Table 5, the dam inflow observations with over 1000 m^3/s are shown with the results of machine-learning forecasts. The analysis reveals that the decision tree has a tendency to over- or underestimate the peak flow, and the ensemble models are underestimating the peak flows during the period of 2017. Although MLP failed to predict the exact values, it appears that the predicted values are close to approximation on average (Table 5).

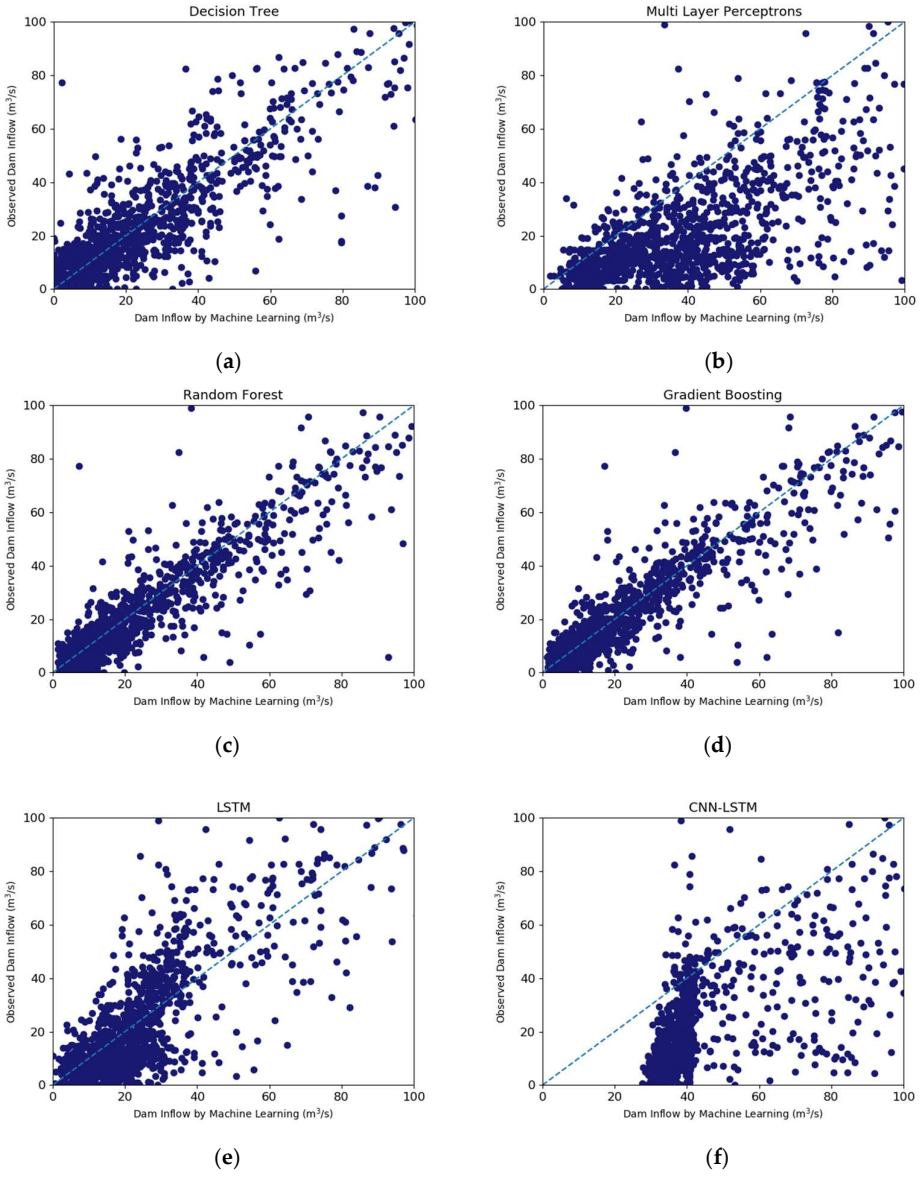


Figure 9. Comparison of dam inflow predicted by (**a**) decision tree, (**b**) multilayer perceptrons, (**c**) random forest, (**d**) gradient boosting, (**e**) LSTM, and (**f**) CNN–LSTM, and observed dam inflow below 100 m^3/s with test data.

Table 5. Prediction values of machine learning models for observed cases of over 1000 m^3/s of Soyang River Dam inflow.

Date	Observation	DT	MLP	RF	GB	LSTM	CNN-LSTM
5 July 2016	3918.50	1383.00	3106.83	2018.88	2457.94	291.74	456.17
6 July 2016	1716.20	2443.00	2581.32	1828.49	1886.02	3302.50	2911.93
3 July 2017	1652.90	432.40	1410.34	1106.03	516.64	1036.20	1140.11
11 July 2017	1773.30	645.50	993.55	918.98	491.15	1349.80	1410.94
24 August 2017	1538.50	676.10	1217.89	755.04	750.64	655.35	492.96
25 August 2017	1181.70	771.50	1022.35	830.98	690.27	975.65	790.40
18 May 2018	1788.60	1669.40	1522.95	1174.48	1471.49	1061.63	985.24
29 August 2018	2380.80	1745.50	1989.37	2108.31	1712.53	400.91	740.87
30 August 2018	1124.50	1327.60	1553.35	1703.71	1224.52	2036.86	1959.48
Average	1897.22	1232.67	1710.88	1382.77	1244.58	1234.52	1299.39

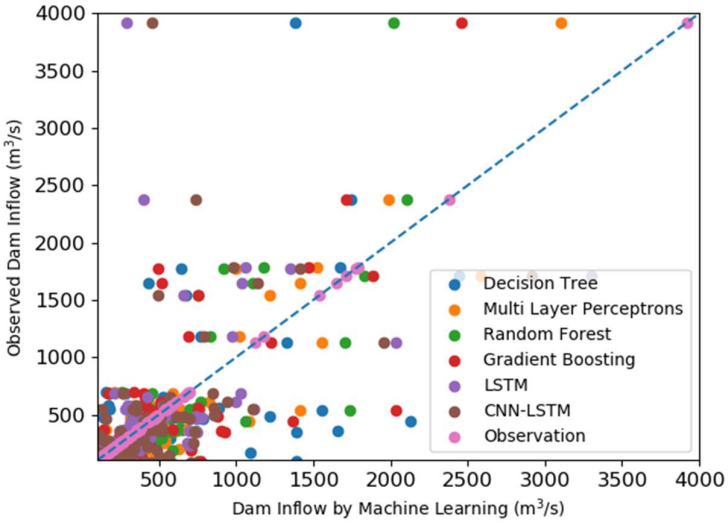


Figure 10. Comparison of dam inflow predicted by machine learning models and observed dam inflow, when the observed dam inflow is over 100 m^3/s.

3.3. Prediction Results Using CombML

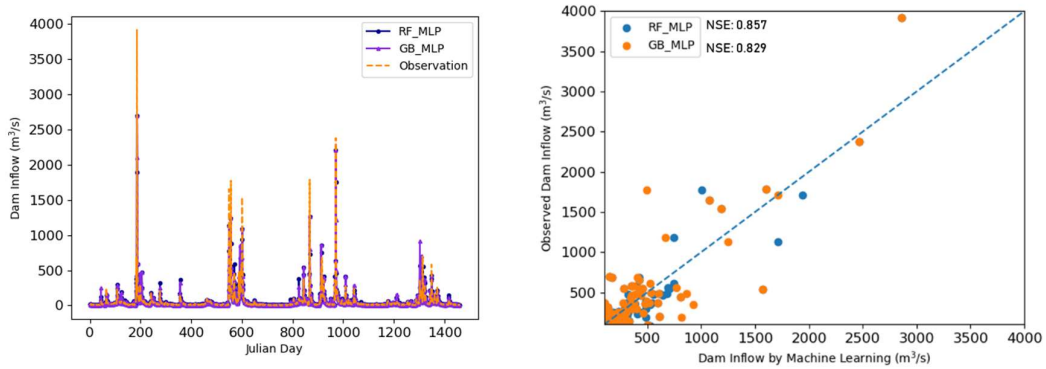
As indicated in the objective of this study, we predicted the inflow of Soyang River Dam using the past weather data, inflow, and the weather forecast accurately. The results of the study showed that the ensemble models (the random forest and gradient boosting) performed well in under 100 m^3/s of the inflow. On the other hand, MLP has merits when predicting the inflow of over 100 m^3/s.

Therefore, a new model combining MLP and ensemble models was created to predict the dam inflow. However, it is impossible to predict the inflow of the next day. Therefore, the forecasted precipitation data, which were shown to have the highest correlation with the dam inflow by the heatmap analyzing, were used as a standard for the new ensemble model. The reference point was set by averaging rainfall on days with dam inflow greater than 100 m^3/s. The average precipitation of the filtered dam inflow was 16 mm. Hence, the MLP was used when the daily precipitation is more than 16 mm, and the ensemble models (random forest and gradient boosting) were used when the daily precipitation is less than 16 mm, to predict the dam inflow.

The results are shown in Figure 11 and Table 6. The random forest–MLP combined model (RF_MLP) has the best results, with NSE of 0.857, RMSE of 68.417 m^3/s, MAE of 18.063 m^3/s, R of 0.927, and R^2 of 0.859. The gradient boosting–MLP combined model (GB_MLP) has the results of NSE of 0.829, RMSE of 73.918 m^3/s, MAE of 18.093 m^3/s, R of 0.912, and R^2 of 0.831. In the previously conducted streamflow estimation using single machine learning algorithms [22,23], and the prediction results of single algorithms conducted in this study, there were comparatively higher uncertainties in a certain situation. For the inflow of the Soyang River Dam, since the flow duration were the variables that affected the prediction accuracy of the algorithms, the prediction accuracy has been improved for either peak flow or the normal and low flow using the CombML algorithms. The accuracy improvement illustrates that it is necessary to construct the dam inflow prediction system with the interval division prediction using the combined model.

Table 6. Prediction accuracy results of combined machine learning (CombML) models.

	NSE	RMSE (m^3/s)	MAE (m^3/s)	R	R^2
RF_MLP	0.857	68.417	18.063	0.927	0.859
GB_MLP	0.829	73.918	18.093	0.912	0.831



(**a**) (**b**)

Figure 11. Comparison of predictive results by input data in multilayer perceptrons model: (**a**) dam inflow prediction with observed dam inflow below 100 m^3/s; (**b**) dam inflow prediction with observed dam inflow over 100 m^3/s.

4. Conclusions

This study evaluated the performance of the algorithms to predict the amount of inflow of the Soyang River Dam, and developed and evaluated the combined machine learning algorithms with the consideration of flow duration. As a result of the comparative analysis of inflow prediction through various algorithms, MLP was proven to be the best algorithms for flow prediction. However, even though MLP was the best algorithm for flow prediction, in terms of model performance evaluation, there was a limitation of flow prediction at the entire flow duration, which means that a single use of algorithm could not perfectly consider flow regimes. To improve this, CombML algorithms were developed, and the results show that it is possible to predict inflow through inflow learning, considering flow characteristics, such as flow in Korea.

The CombML was developed and evaluated to take account of flow regimes for the prediction of the single algorithms. The CombML models, the random forest–multilayer perceptron model (RF_MLP), and the gradient boosting–multilayer perceptron model (GB_MLP), increased the model accuracy for the prediction of the dam inflow, whereas each of the single models performed partially satisfactory. The random forest–multilayer perceptron model (RF_MLP) had the results of NSE of 0.857, RMSE of 68.417 m^3/s, MAE of 18.063 m^3/s, R of 0.927, and R^2 of 0.859. The gradient boosting–multilayer perceptron combined model (GB_MLP) had the results of NSE of 0.829, RMSE of 73.918 m^3/s, MAE of 18.093 m^3/s, R of 0.912, and R^2 of 0.831. The weakness of MLP model analysis for prediction of the dam inflow was mitigated by the improvement of the prediction of the flood runoff, by combining MLP and ensemble models. The experimental results clearly indicate that the CombML improves on the limitations of flow regime and rainfall on inflow prediction from using a single algorithm.

Although the research area was focused on the Soyang River Dam watershed, the application of the algorithm can be expected, because most of the dam watersheds in Korea are covered by forest, like the Soyang River Dam, and the referencing point is based on precipitation rather than flow rate. Therefore, it is expected that CombML, which takes into account flow regimes, can be used to establish basic data for the establishment of plans for water supply by controlling dam water level during flooding and securing dam water storage capacity during drought. Also, the global application of this method for other water sources (e.g., dam, river, stream, etc.) could be possible by analyzing the correlation of the data and combining the model, since it requires only the weather and flow data.

For further research, the development of dam inflow prediction technology considering not only natural factors, but also artificial factors, such as water use, land use changes, hydraulic structures, etc., in the upstream watershed is necessary.

Author Contributions: Conceptualization: K.J.L., J.H., J.K., and J.H.B.; methodology: J.H.B.; formal analysis: J.H., J.H.B. and W.J.P.; data curation: J.H., J.L., and D.L.; writing—original draft preparation: J.H.; writing—review and editing: S.L.; supervision: K.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Korea Environment Industry & Technology Institute(KEITI) through Aquatic Ecosystem Conservation Research Program, funded by Korea Ministry of Environment(MOE), grant number 2020003030004.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Donnelly, C.; Greuell, W.; Andersson, J.; Gerten, D.; Pisacane, G.; Roudier, P.; Ludwig, F. Impacts of climate change on European hydrology at 1.5, 2 and 3 degrees mean global warming above preindustrial level. *Clim. Chang.* **2017**, *143*, 13–26. [CrossRef]
2. Choi, I.-C.; Shin, H.-J.; Nguyen, T.T.; Tenhunen, J. Water policy reforms in South Korea: A historical review and ongoing challenges for sustainable water governance and management. *Water* **2017**, *9*, 717. [CrossRef]
3. Ahn, J.M.; Jung, K.Y.; Shin, D. Effects of coordinated operation of weirs and reservoirs on the water quality of the Geum River. *Water* **2017**, *9*, 423.
4. Park, J.Y.; Kim, S.J. Potential impacts of climate change on the reliability of water and hydropower supply from a multipurpose dam in South Korea. *JAWRA J. Am. Water Resour. Assoc.* **2014**, *50*, 1273–1288. [CrossRef]
5. Lee, J.E.; Heo, J.-H.; Lee, J.; Kim, N.W. Assessment of flood frequency alteration by dam construction via SWAT Simulation. *Water* **2017**, *9*, 264. [CrossRef]
6. Ryu, J.; Jang, W.S.; Kim, J.; Choi, J.D.; Engel, B.A.; Yang, J.E.; Lim, K.J. Development of a watershed-scale long-term hydrologic impact assessment model with the asymptotic curve number regression equation. *Water* **2016**, *8*, 153. [CrossRef]
7. Stern, M.; Flint, L.; Minear, J.; Flint, A.; Wright, S. Characterizing changes in streamflow and sediment supply in the Sacramento River Basin, California, using hydrological simulation program—FORTRAN (HSPF). *Water* **2016**, *8*, 432. [CrossRef]
8. Nyeko, M. Hydrologic modelling of data scarce basin with SWAT Model: Capabilities and limitations. *Water Resour. Manag.* **2015**, *29*, 81–94. [CrossRef]
9. Zhao, F.; Wu, Y.; Qiu, L.; Sun, Y.; Sun, L.; Li, Q.; Niu, J.; Wang, G. Parameter uncertainty analysis of the SWAT model in a mountain-loess transitional watershed on the Chinese Loess Plateau. *Water* **2018**, *10*, 690. [CrossRef]
10. Lee, G.; Lee, H.W.; Lee, Y.S.; Choi, J.H.; Yang, J.E.; Lim, K.J.; Kim, J. The effect of reduced flow on downstream water systems due to the kumgangsan dam under dry conditions. *Water* **2019**, *11*, 739. [CrossRef]
11. Valipour, M.; Banihabib, M.E.; Behbahani, S.M.R. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *J. Hydrol.* **2013**, *476*, 433–441. [CrossRef]
12. Liu, Y.; Wu, J.; Liu, Y.; Hu, B.X.; Hao, Y.; Huo, X.; Fan, Y.; Yeh, T.J.; Wang, Z.-L. Analyzing effects of climate change on streamflow in a glacier mountain catchment using an ARMA model. *Quat. Int.* **2015**, *358*, 137–145. [CrossRef]
13. Myronidis, D.; Ioannou, K.; Fotakis, D.; Dörflinger, G. Streamflow and hydrological drought trend analysis and forecasting in Cyprus. *Water Resour. Manag.* **2018**, *32*, 1759–1776. [CrossRef]
14. Rezaie-Balf, M.; Naganna, S.R.; Kisi, O.; El-Shafie, A. Enhancing streamflow forecasting using the augmenting ensemble procedure coupled machine learning models: Case study of Aswan High Dam. *Hydrol. Sci. J.* **2019**, *64*, 1629–1646. [CrossRef]
15. Balaguer, E.; Palomares, A.; Soria, E.; Martín-Guerrero, J.D. Predicting service request in support centers based on nonlinear dynamics, ARMA modeling and neural networks. *Expert Syst. Appl.* **2008**, *34*, 665–672. [CrossRef]
16. Ali, M.; Qamar, A.M.; Ali, B. Data Analysis, Discharge Classifications, and Predictions of Hydrological Parameters for the Management of Rawal Dam in Pakistan. In Proceedings of the 12th International Conference on Machine Learning and Applications, Miami, FL, USA, 4–7 December 2013; pp. 382–385.
17. Le, X.-H.; Ho, H.V.; Lee, G.; Jung, S. Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* **2019**, *11*, 1387. [CrossRef]

18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
19. Huang, C.-J.; Kuo, P.-H. A deep cnn-lstm model for particulate matter (PM2.5) forecasting in smart cities. *Sensors* **2018**, *18*, 2220. [CrossRef] [PubMed]
20. Bougoudis, I.; Demertzis, K.; Iliadis, L. HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens. *Neural Comput. Appl.* **2016**, *27*, 1191–1206. [CrossRef]
21. Tongal, H.; Booij, M.J. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *J. Hydrol.* **2018**, *564*, 266–282. [CrossRef]
22. Shortridge, J.E.; Guikema, S.D.; Zaitchik, B.F. Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrol. Earth Syst. Sci.* **2016**, *20*, 2611–2628. [CrossRef]
23. Cheng, M.; Fang, F.; Kinouchi, T.; Navon, I.; Pain, C. Long lead-time daily and monthly streamflow forecasting using machine learning methods. *J. Hydrol.* **2020**, *590*, 125376. [CrossRef]
24. Chung, S.-W.; Lee, J.-H.; Lee, H.-S.; Maeng, S.-J. Uncertainty of discharge-SS relationship used for turbid flow modeling. *J. Korea Water Resour. Assoc.* **2011**, *44*, 991–1000. [CrossRef]
25. Jung, I.; Shin, Y.; Park, J.; Kim, D. Increasing Drought Risk in Large-Dam Basins of South Korea. In Proceedings of the AGU Fall Meeting Abstracts, New Orleans, LA, USA, 11–15 December 2017.
26. Korea Meteorological Administration (KMA). Available online: http://kma.go.kr/home/index.jsp (accessed on 12 February 2020).
27. Water Resources Management Information System (WAMIS). Available online: http://www.wamis.go.kr/main.aspx. (accessed on 12 February 2020).
28. Woo, W.; Moon, J.; Kim, N.W.; Choi, J.; Kim, K.; Park, Y.S.; Jang, W.S.; Lim, K.J. Evaluation of SATEEC daily R module using daily rainfall. *J. Korean Soc. Water Qual.* **2010**, *26*, 841–849.
29. Bae, J.H.; Han, J.; Lee, D.; Yang, J.E.; Kim, J.; Lim, K.J.; Neff, J.C.; Jang, W.S. Evaluation of sediment trapping efficiency of vegetative filter strips using machine learning models. *Sustainability* **2019**, *11*, 7212. [CrossRef]
30. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Data preprocessing for supervised leaning. *Int. J. Comput. Sci.* **2006**, *1*, 111–117.
31. Teng, C.-M. Correcting Noisy Data. In Proceedings of the 16th Inetrantional Conference on Machine Learning, Bled, Slovenia, 27–30 June 1999; pp. 239–248.
32. Scikit-Learn. RandomForestRegressor. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on 2 December 2019).
33. Thara, D.; PremaSudha, B.; Xiong, F. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognit. Lett.* **2019**, *128*, 544–550.
34. Alpaydin, E. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2014.
35. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
36. Azar, A.T.; El-Said, S.A. Probabilistic neural network for breast cancer classification. *Neural Comput. Appl.* **2013**, *23*, 1737–1751. [CrossRef]
37. Moon, J.; Park, S.; Hwang, E. A multilayer perceptron-based electric load forecasting scheme via effective recovering missing data. *KIPS Trans. Softw. Data Eng.* **2019**, *8*, 67–78.
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Panchal, G.; Ganatra, A.; Kosta, Y.; Panchal, D. Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers. *Int. J. Comput. Theory Eng.* **2011**, *3*, 332–337. [CrossRef]
40. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef] [PubMed]
41. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
42. Chen, Z.; Liu, Y.; Liu, S. Mechanical State Prediction Based on LSTM Neural Netwok. In Proceedings of the 36th Chinese Control Conference (CCC), Dalian, China, 26–28 July 2017; pp. 3876–3881.
43. Tran, Q.-K.; Song, S.-K. Water level forecasting based on deep learning: A use case of Trinity River-Texas-The United States. *J. KIISE* **2017**, *44*, 607–612. [CrossRef]
44. Fukuoka, R.; Suzuki, H.; Kitajima, T.; Kuwahara, A.; Yasuno, T. Wind Speed Prediction Model Using LSTM and 1D-CNN. *J. Signal Process.* **2018**, *22*, 207–210. [CrossRef]

45. Jung, H.C.; Sun, Y.G.; Lee, D.; Kim, S.H.; Hwang, Y.M.; Sim, I.; Oh, S.K.; Song, S.-H.; Kim, J.Y. Prediction for energy demand using 1D-CNN and bidirectional LSTM in Internet of energy. *J. IKEEE* **2019**, *23*, 134–142.
46. Legates, D.R.; McCabe, G.J., Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35*, 233–241. [CrossRef]
47. Cancelliere, A.; Di Mauro, G.; Bonaccorso, B.; Rossi, G. Drought forecasting using the standardized precipitation index. *Water Resour. Manag.* **2007**, *21*, 801–819. [CrossRef]
48. Moghimi, M.M.; Zarei, A.R. Evaluating performance and applicability of several drought indices in arid regions. *Asia-Pacific J. Atmos. Sci.* **2019**, 1–17. [CrossRef]
49. Karpagavalli, S.; Jamuna, K.; Vijaya, M. Machine learning approach for preoperative anaesthetic risk prediction. *Int. J. Recent Trends Eng.* **2009**, *1*, 19.
50. Oliveira, T.P.; Barbar, J.S.; Soares, A.S. Computer network traffic prediction: A comparison between traditional and deep learning neural networks. *Int. J. Big Data Intell.* **2016**, *3*, 28–37. [CrossRef]