Journal of Hydrology 602 (2021) 126723

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region

Tiantian Yang [a,*], Lujun Zhang [a], Taereem Kim [a], Yang Hong [a], Di Zhang [b], Qidong Peng [b]

[a] The School of Civil Engineering and Environmental Science, University of Oklahoma, United States
[b] China Institute of Water Resources and Hydropower Research (IWHR), China

ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Vahid Nourani, Associate Editor

Keywords:
Artificial Intelligence
Data Mining
Reservoir Operation
Decision Making

ABSTRACT

In recent years, the Artificial Intelligence and Data Mining (AI&DM) models have become popular tools in assisting various aspects of reservoir operation. However, the practical uses are still rarely reported. Comparison experiment of many AI&DM models over a large number of reservoir cases is particularly valuable to help reservoir operators first examine the usefulness and transferability of different AI&DM models, and then identify the most stable and reliable AI&DM model in assist of various decision-making processes. In this study, a total of 12 AI&DM models with different parameterizations and simulation scenarios are comprehensively tested out and compared in simulating the controlled reservoir outflows of 33 reservoir cases over the Upper Colorado Region, United States. Results show that the Random Forecast and the Long-Short-Term-Memory model could consistently derive the best statistical performance than other models under the baseline simulation scenario. The employed AI&DM models could obtain satisfactory statistical interquartile ranges (25–75%) between [0.6–0.9], [0.3–0.8], and [0.2–0.8], for CORR, NSE, and KGE measurements, respectively, and [1.5–6.5], [−15 to 20], and [0.5–8.5] for the normalized RMSE, PBIAS and RSR measurements, respectively. Results also show Multi-Layer Perceptron model and Extreme Gradient Boosting Tree Algorithm produced more stable and superior performance than other models under more complex input scenarios. We also found that the performance of different AI&DM models are closely relevant to the reservoir elevations, sizes, and functionalities. Discussions were made about the sensitivity of AI&DM models' parameterizations and the key advantages of AI&DM models over the rule-based reservoir models. We further identify that the main advantage of AI&DM models is the flexibility in designing input structures, whereas the rule-based simulation model is rather limited. Future studies were suggested regarding the best way reservoir operators and researchers could use, select, and apply different AI&DM models in simulating reservoir releases under different natural and modeling environments. This comparison study also serves as a reference and a piece of groundwork for further promoting the practical uses of AI&DM models in assisting reservoir operation.

1. Introduction

Reservoirs and dams are fundamental, human-built, multi-functional water infrastructures that collect, store, and deliver fresh surface water for a multitude of uses, including flood and fire control, recreation, wildlife habitat, residential, industrial, and agricultural water supply, hydro-electric power generation, supply source during droughts, and more. The reservoir release decisions directly influence various aspects of social-economic functioning and our nation's security (Yang et al., 2016, Yang et al., 2020b). In recent years, more frequent and severe abrupt weather extremes, climate change, natural hazards, aging infrastructure, and increases in water demands due to population growth, have placed another great barrier to prevent effective, sustainable, and flexible operation for our nations' reservoir systems. For example, in May 2020, due to extended extreme precipitation, water behind the two consecutive reservoirs in Michigan reached the reservoir storage capacity and caused catastrophic dam breaks, flooding the Tittabawassee River, and completely drained the reservoirs (CNN, 2020). In 2017, the Addicks and Barkers reservoirs near the Houston Area were intentionally operated to release additional water downstream. This

* Corresponding author.
E-mail address: tiantian.yang@ou.edu (T. Yang).

https://doi.org/10.1016/j.jhydrol.2021.126723
Received 17 March 2021; Received in revised form 26 June 2021; Accepted 20 July 2021
Available online 26 July 2021

operation happened during the same period when hurricane Harvey hit the Houston area, causing an additional 8000 houses near these two reservoirs were flooded. The federal judge ruled the U.S. Army Corps of Engineers (USACE) is liable for flooding these homes (CNN, 2019, HoustonChronicle, 2019, HoustonPublicMedia, 2019). In the same year, a sudden water flux reaching the designed spillway capacity of the Oroville dam, California, caused irreversible damage to the dam and triggered a large-scale emergency evacuation of more than 180,000 people living downstream (MercuryNews, 2017, NewYorkTimes, 2017). These recent dam failure cases indicate the need for continuous developments of effective and flexible reservoir operation tools and modeling schemes.

The approaches for reservoir operation and decision support can be categorized into optimization models and simulation models (Labadie, 2004, Reddy and Kumar, 2006, Yeh, 1985). Differs from reservoir optimization models, reservoir simulation models are useful in assisting operators in estimating reservoir yields and quantifying system behaviors based on pre-defined operating rules (Louks and Sigvaldason, 1981). In early studies, Sigvaldson (1976) developed an innovative approach for simulating reservoir responses using a priority ranking concept. Chaturvedi and Srivastava (1981) developed a screen-simulation model based on linear programming methods for a large complex water resources system. The reservoir simulation models have rapidly evolved from excel sheets-based models in early times to coupled hydrological and hydraulic models to support various types of operations, such as predicting how the system behaves under the current hydrological situation (inflow, storage, extractions, releases) on different temporal scales. Many reservoir simulation models have become operational and widely used in the U.S., e.g., the HEC-5 model developed by USACE (Bonner, 1989), DWRSIM developed by CDWR (Barnes and Chung, 1986, Chung et al., 1989), the WEAP21 model (Yates et al., 2005), the Calsim model (Draper et al., 2004), the RiverWare models (Zagona et al., 2001), and the CRAM Water Resources Modeling Tool (LynkerTech, 2018), etc. Lund and Guzman (1999) concluded that simulation models were more likely to be trusted as a standard by operators as compared to reservoir optimization models.

In practice, these reservoir simulation models rely on so-called reservoir operating rule curves, which define an empirically desired reservoir storage-release relationship (Louks and Sigvaldason, 1981). These rule curves are subject to approval by governmental authorities and are defined beforehand. Usually, the formulation of the release rule is based on historical data or design scenarios. The set of rules must be defined in such a way that for as many as situations and the operational goals (e.g., power production, water supply, minimum flow) are conservatively met under given constraints (e. g. dam safety requirements, flood control requirements, environmental obligations). An advantage of rule-based operations is that the set of rules is usually transparent, clear, and can easily be integrated into simulation models for the water system.

However, Oliveira and Loucks (1997) pointed out in many situations, the operators will still operate the system in a way that deviates from these pre-defined rule curves to adapt to specific conditions, objectives, or constraints that may change over time. Draper et al. (2004) also criticized that many simulation models were severely restricted by the explicit implementation of operating rules as hard model constraints, which jeopardized the flexibility of using such tools to adjust to different environmental settings. In other words, a drawback of rule-based operations is that the control actions are not necessarily the optimal strategy for the current situation and could not cover various conditions from changing environment. To give an example: a target water level for a reservoir on a specific day in the year would account for both dry and wet situations in order to be able to cope with both water scarcity and flood issues. In a dry situation, it might be possible to operate the reservoir with a higher target water level. This would be beneficial for hydropower production and water supply, but is of course, only suitable if one can afford to operate with a lower flood control room. If, on the contrary, a substantial increase in inflow is expected, the target water level should be even lower than the rules might say. In such a situation, it will make sense to pre-release water in order to generate a sufficient flood control room.

With respect to these advantages and disadvantage of rule-based reservoir operation and simulation models, in recent years, the Artificial Intelligence and Data Mining (AI&DM) techniques become popular tools in assisting reservoir operation and decision making (Cancelliere et al., 2002, Chaves and Chang, 2008, Cheng et al., 2020, Coulibaly et al., 2001, Coulibaly et al., 2000, Esmaeilzadeh et al., 2017, Jain et al., 1999, Kişi, 2007, Maier et al., 2010, Wu et al., 2009, Yang et al., 2017b, Zhang et al., 2019). The AI&DM models are powerful tools in data classification and regression, but they purely rely on the statistical relationship between the target variables and the input data (e.g., the input features). By setting different combinations of model training data and target variables, the uses of AI&DM models also appear to be versatile. For example, these models can be used to simulate reservoir release, to extract the existing operation rules, and to predict reservoir inflows flows and uncertainties, and to manage reservoir storage and water levels, etc. (Ashaary et al., 2015, Bessler et al., 2003, Castelletti et al., 2012, Castelletti et al., 2010, Chang et al., 2016, Cheng et al., 2008, Rahnamay Naeini et al., 2020, Wei and Hsu, 2008).

However, one big challenge of applying these AI&DM tools in assisting reservoir operation is the lack of scalability and transferability. Different AI&DM models employ distinct data classification and regression philosophies, in which none of them are subject to the mass-balance equations with physical constraints. This hinders the practitioner from trusting the modeling results, especially when operators are used to the traditional rule-based simulation models. In addition, the AI&DM model's performance may substantially vary based on user-selected model structural parameters and the way different AI&DM models are set up. In short, it is an extremely tedious work for reservoir operators to vet a set of AI&DM models, identify the most suitable approach and associated parameters for one particular application, and carry out verification experiments for another reservoir case or another problem setting to make sure the AI&DM models are transferable and scalable under different simulation environment. Differs from the rule-based reservoir simulation models, the performance of AI&DM models are likely to change when the training data changes. Furthermore, the field of artificial intelligence is still rapidly evolving. Newer and stronger AI&DM models are becoming readily available for applications before an older and simpler model being thoroughly evaluated is assisting of reservoir operation and decision-making. This makes reservoir operators even more hesitant to trust and practice an alternative AI&DM model over their existing and functional rule-based reservoir simulation models.

In the research community, there have been numerous studies to apply a variety of AI&DM models to assist reservoir operation, hydrology, and water resources management (Adnan et al., 2019, Shabani et al., 2020, Shamshirband et al., 2020, Yuan et al., 2018). Some popular AI&DM models include linear regression model, support vector machines (SMV), k-Nearest Neighbors regression (kNN), Decision Tree (DT) model, Multi-Layer Reception (MPL) model (i.e., Artificial Neural Network Model), and Deep Learning Algorithm (i.e., the Convolutional Neural Network and Recurrent Neural Network family), etc. However, each AI&DM model is subject to specific pros and cons, and there is no commonly accepted agreement on which modeling scheme is consistently effective than others across different study cases.

The linear regression model is the most straightforward statistical technique and early approach used to quantify the linear relationship in hydrological time series. Though it is relatively simple and old-fashioned, it has been widely used to investigate hydrologic variables (Adnan et al., 2020, Caldwell et al., 2015, Li et al., 2016, Ombadi et al., 2020, Ren et al., 2020, Sahoo and Jha, 2013, Schmidt et al., 2020, Yuan et al., 2018), derive reservoir operation rule/policy (Ghimire et al., 2020, Liu et al., 2019, Zhou et al., 2016), predict reservoir inflow and

streamflow (Adnan et al., 2019, Lima and Lall, 2010b, Masselot et al., 2016), estimate hydraulic behaviors (Adnan et al., 2021), and assist water quality management (Zhao et al., 2018) and drought prediction (Li et al., 2020). The linear regression model's advantage is its simplicity and efficiency when the decision variables and target variables have an underlying linear correlation. However, the disadvantage is also apparent: If only a nonlinear correlation exists between the decision variables and target variables, such always fails to capture such a complex relationship and can only simplify the regression with a linear estimation.

Unlike the linear models, the Support Vector Machine (SVM) model acknowledges the existence of a possible nonlinear relationship between the features and the target variable. The SVM models could be further categorized into Support Vector Classification (SVC) and Support Vector Regression (SVR) models based on the nature of problems. In the SVM model, a hyperplane will be created and used to separate the feature samples in the feature space. This hyperplane could be either linear or nonlinear based on the user-selected kernel functions. Because the hyperplane in SVM models could adaptively partition the data samples, the SVM models could address the disadvantage of linear models for nonlinear regression and guarantee a unique and globally optimal solution when searching for the hyperplane (Lin et al., 2006). Theoretically, the SVM models could minimize the errors in the learning process and effectively reduce data overfitting if a proper kernel function is applied (Lal and Datta, 2018, Yu et al., 2006b). The SVM model has been applied to solve various water resources problems, such as reservoir operation (Aboutalebi et al., 2016, Bozorg-Haddad et al., 2018, Ji et al., 2014, Liu et al., 2017a, Xie et al., 2012), and reservoir inflow and streamflow forecasting (Babaei et al., 2019, Feng et al., 2020, Liu et al., et al., 2017b, Malik et al., 2020, Samadianfard et al., 2019, Tao et al., 2018). The SVM model's advantage is that it can easily overcome the high dimensionality problem (Hand, 2007). However, a major drawback associated with the SVM model is its low training efficiency (Wei, 2015). When applying the SVM models to datasets with large samples, the training time tends to increase exponentially with the total number of data samples, which prohibits some real-world applications, especially when quick model training and decision making are needed.

The k-Nearest Neighbors (kNN) model is an instance-based learning, or lazy-learning method, which was originally developed by Fix (1951), and further improved by many others (Altman, 1992, Coomans and Massart, 1982, Cover and Hart, 1967). The kNN model can be applied to both classification and regression problems. The classification procedure uses a neighbor search algorithm to recursively find the closeness of a total number of k training examples in the feature space. After classification, a regression could be carried out by averaging the target values from the k nearest neighbor samples (Atkeson et al., 1997). The applications of kNN for reservoir operation and water resources management are also numerous. For example, Nikoo et al. (2014) applied kNN for water and wastewater allocation in the Dez reservoir–river system in Iran and obtained good model statistical performance. Ahmadi et al. (2010) combined kNN and a Genetic algorithm in a reservoir simulation–optimization model and successfully incorporated forecast uncertainties of inflow into optimal reservoir operation. Yang et al. (2020a) tested a kNN forecasting model to generate medium- to long-term inflow forecasts for the Danjiangkou Reservoir in China, and the results proved the validity and reliability of the proposed kNN prediction method. The advantages of kNN lie in its capability in achieving computational tractability (Toussaint, 2005), and high effectiveness in approximating the target variables using a limited number of decision variables in a small subset of data samples. Two drawbacks of kNN include (1) the training procedure sometimes overlooks the similarity and statistical relationship of the entire training samples, thus, did not work well on datasets that contain a high level of noises, and (2) all calculations are deferred until classification (Bremner et al., 2005). Similar critiques on kNN models also exist. For example, Akbari et al. (2011) pointed out that successful applications of the kNN model rely on the similarity of the output values within the defined neighbors, where the feature vectors are to be relatively close to each other. However, the number of neighbors in each group can be different, and some neighbors may belong to none of the clusters. In such cases, the prediction made by kNN algorithms may be risky and unreliable. In other words, the application of kNN models would require a strong and local correlation between decision and target variables, though such a correlation relationship may not exist globally. Nevertheless, the use of kNN models in reservoir operation is still popular in the literature.

Another popular AI&DM model set in support of reservoir management is the Decision Tree (DT) model (Breiman, 2001, Breiman et al., 1984, Chen and Guestrin, 2016, Loh, 2014, Quinlan, 1986). The DT model is a set of "white-box" machine learning models, which rely on building a sequence of simple Boolean "If-Then", and "True-False" logic to explain how complex data samples could be partitioned into smaller subsets or classes based on the feature values. A comprehensive review of DT models can be found in Mosavi et al. (2018). In other words, each data sample is regarded as a sequence of logical decision outcomes from the feature variables, and similar decision outcomes could be traced back to the feature values thresholds used in the training process or the tree-growing process. Therefore, the advantage of DT models is the transparency to users, and it shares a very similar procedure of how reservoir release decision is made. For example, dam operators typically use current storage level and rule curves (i.e., the relationship between discharge and storage) to decide whether to release a certain amount of water from reservoirs (Raso et al., 2014, Schwanenberg et al., 2012, Schwanenberg et al., 2014, Uysal et al., 2020, Zagona et al., 2001, Zhang et al., 2020). Some recent applications of DT models to reservoir release simulation have proven their usefulness in assisting different reservoir operation, such as release scheduling (Ji et al., 2016, Rahnamay Naeini et al., 2020, Wei, 2012, Yang et al., 2015, Yang et al., 2020b, Zhang et al., 2019), and reservoir inflow forecast (Erdal and Karakurt, 2013, Tongal and Booij, 2018, Yang et al., 2017b). Specifically, Yang et al. (2016) tested out a few different DT models to simulate the reservoir releases from 9 major reservoirs in California and concluded that the advantage of DT models is their effectiveness in capturing how reservoir releases are following the storage and inflow conditions, which is a similar process used in the traditional Rule-based simulation model. In another study, Yang et al. (2020b) found that different DT models in various input conditions could generate varying simulations with high variations and concluded that DT models are very sensitive to training data and may have a low transferability. In general, the implementations of DT models for reservoir simulation are still limited in practice. One reason is the lack of accuracy of ensemble forecasts and significant hydrological uncertainty, and the other reason is that over- and underfitting can occur in DT models for datasets with a small number of samples, or the training samples value becomes too sparse. In short, low robustness and accuracy are identified when scaling up DT models to different reservoir case studies (Brodley and Utgoff, 1995, Dietterich, 1995, Mingers, 1989).

The Multiple-Layers Perceptron (MLP), or Artificial Neural Network (ANN) model, is also one of the popular AI&DM model sets in the broader field of hydrology community. MLP or ANN model has been widely applied to solve many types of problems, including the rainfall-runoff simulation, reservoir operation, and streamflow forecast problems (Coulibaly et al., 2000, Dawson and Wilby, 1998, Kişi, 2004, Xu and Li, 2002). The main advantage of the ANN model is its ability to detect complex nonlinear relationships between input features and outputs through the flexible learning process, and the ANN model can incorporate all necessary relationships through training procedures without requiring a priori knowledge of the underlying process (Daliakopoulos et al., 2005, French et al., 1992). For example, Sattari et al. (2012) applied the ANN model with an early stopped training approach to enhance the model's predictive performance. They demonstrated that the ANN model approach has substantially better accuracy than other baseline models in forecasting the daily reservoir inflow time series.

Zhang et al. (2018) compared three AI models, including an ANN model, an SVR model, and a deep learning model, to assist reservoir operation at different time scales, concluding that the ANN model's applicability on limited amounts of data is better than the other two AI models. Niu et al. (2019) also compared three AI models (ANN, SVR, and extreme learning machine) and a multi-variable linear regression for deriving the hydropower reservoir operation rule of Hongjiadu reservoir in southwest China, concluding that the three ANN models have some unique merits and have better performance than the other employed regression models. In many other studies, the ANN or the MLP models have shown superior performance over other traditional statistical time series models (i.e., Auto-Regressive Integrated Moving Average; ARIMA) and/or other AI&DM models in different hydrological and reservoir operation studies (Adamowski et al., 2012, Babaei et al., 2019, Jain et al., 1996, Lin et al., 2009, Lohani et al., 2012, Raman and Sunilkumar, 1995, Valipour et al., 2013).

However, the MLP or ANN models still suffer from some weakness caused by 1) the lack of physical interpretation of its structure, 2) difficulties in finding the global optimum of connecting weights, and 3) uncertainties from the model parameters and settings (Govindaraju, 2000, Kim et al., 2020). Specifically, the inputs, number of layers and nodes, activation function, and training algorithm are considered as the major sources of uncertainty (Kasiviswanathan and Sudheer, 2017). For example, the most commonly used activation function is the logistic/sigmoid function (Hsu et al., 1995). However, the use of the hyperbolic tangent activation function could be better than the logistic/sigmoid function in other applications (Zadeh et al., 2010). In addition, the generalization of the ANN model can be highly challenging as the model performance can vary from one training dataset to another, even with the same model structure and parameters. In other words, when applying ANN or MLP models to different reservoir case studies, practitioners need a lot of efforts to tune the model parameters, and the parameters selection process could be challenging and time-consuming. These drawbacks of MLP or ANN models substantially prevent a broader use of the models in assisting various types of reservoir operation practices.

The Long Short Term Memory (LSTM) model is a class of deep machine learning algorithms and is rapidly gaining popularity in more recent years. The LSTM model is a special kind of Recurrent Neural Network (RNN), which enables the output in the previous step to be used as input in the current step; thus, it has merits for simulating and predicting sequential time series data. Several recent studies have reported the superior performance of applying the LSTM model to predict hydrological time series (Apaydin et al., 2020, Fan et al., 2020, Kao et al., 2020, Sahoo et al., 2019, Yuan et al., 2018), and most of the studies are relatively new as compared to other AI&DM models. For example, Kratzert et al. (2018) tested the LSTM model in rainfall-runoff forecasting and compared the results to the well-known process-based hydrologic model (i.e., Sacramento Soil Moisture Accounting Model, or SAC-SMA model). They concluded that the LSTM model is competitive in predicting runoff from meteorological observations compared to the process-based SAC-SMA model. Zhang et al. (2018) tested out the LSTM model in predicting the long-term inflow time series of the DanjiangKou Reservoir, which is the head-water source for China's South-To-North Water Diversion Project middle route, and verified the outstanding advantage of the LSTM model in effectively learning the sequential dependencies of reservoir inflow variabilities. Zhu et al. (2020) developed an improved LSTM model, applied it to predict the daily streamflow at four stations in the upper Yangtze River, and compared the performance against three benchmark models (ANN model, generalized linear model, and heteroscedastic Gaussian Process model). The results showed that the improved LSTM model has satisfying performance compared to the benchmark models for high flow forecasting. Zolfaghari and Golabi (2021) compared a few DT models, the LSTM model, and six other benchmark models to predict the hydropower generation of The Mahabad Dam in Iran, and concluded that the LSTM model is one of the emerging AI&DM models and has great potential in assisting reservoir operation and hydropower scheduling.

The major advantage of the LSTM model is its capability of taking sequential data as inputs instead of independent training samples. This feature benefits to model's capability of dealing with more extended historic hydrologic observations with temporal dependence (Wu et al., 2020), which is a common feature associated with many types of hydrological time series. However, the LSTM model also has some limitations inherited from the "black-boxed" model, which is the lack of explicit internal representation of the water balance (Kratzert et al., 2018). Moreover, the training of LSTM could also be computationally expensive in some cases, since the model trains on data sequence with user-defined time steps. The longer the time steps or the data sequence, the more expensive the computation will be. Nonetheless, many recent studies demonstrated a strong potential of the LSTM model in hydrological time series forecasting and reservoir operation (Bai et al., 2021, Sahoo et al., 2019, Xiang et al., 2020).

With respect to the above literature review, though AI&DM models are becoming more and more popular in the field of reservoir operation, the performance of different models still vary, and each model has its own strengths and weakness. Large-scale comparison is critically needed to comprehensively evaluate the performance of various AI&DM models, as well as to prevent AI&DM models from becoming an "alchemy" (Hutson, 2018) in the hydrology and water resources management research community. It is essential to establish a standard evaluation testbed that includes a large number of case studies (at least 10+) to identify whether the AI&DM models and the commonly employed parameters are transferable from one case to another, and to verify the model's predictive performance is stable and reliable across different reservoir cases. Such a comparison study can further verify the usefulness of AI&DM models in assisting reservoir operation and help decision-makers to rebuild the confidence for future uses of AI&DM models in practice. However, as far as the authors' knowledge, such a study that compares multiple methods over a large number of study cases for reservoir operation is rare in the literature. Most of the existing studies only compare two or three AI& DM models in a limited number of studies without further examining the transferability of their proposed models.

Therefore, in this study, we apply a total of 12 popular AI&DM models, which consist of two linear models, two SVM models, two kNN models, three DT models, two ANN models, and one Deep Learning model, to simulate the daily reservoir releases over 33 study cases over the upper Colorado region in the U.S. In order to comprehensively examine the models' performance, we compared the simulated release time series with observation via six commonly accepted statistical metrics under three different input scenarios. Through this study, we want to answer the questions that (1) whether these AI&DM models could reasonably mimic the human's release decisions using the limited information of historical reservoir inflow and storage time series? (2) which model performs the best or worse on what conditions and evaluation criterion? (3) can the employed AI&DM models generate reliable and stable predictions across different study cases? And (4) whether a model is transferable to other untested cases with high statistical confidence? The experiment setting and findings of this study will also provide a technical reference of model evaluation and operation guidance on the topic of applying AI&DM methods to reservoir operation for interested researchers and operators.

The rest of the paper is organized as follows: Section 2 summarizes the applied AI&DM models and employed statistical measures; The data and experiment settings are described in Section 3; In Section 4, we demonstrate the experiment results; The discussion and conclusion are provided in Sections 5 and 6, respectively. The Supplementary Material includes all the calculated statistics of each reservoir under different scenarios and the detailed model setting and parameters. The main body of the paper provides further analysis and a summary of the obtained simulation results.

2. Methodology

In this study, we applied a total of twelve AI&DM models with different model complexity and parameterizations. The twelve models could be categorized into six groups based on the data classification and regression mechanism used: (1) The first modeling group includes a basic multi-variable linear regression model and a second linear model termed the ridge regression model (Hoerl and Kennard, 1970, Marquardt and Snee, 1975). Both of them belong to the most basic linear model and the simplest regression model set; (2) The second model set consists of two types of Supportive Vector Regression (SVR) models (Boser et al., 1992, Cortes and Vapnik, 1995, Vapnik, 2013, Vapnik, 1999): one SVR model is implemented with the radial basis function kernel (SVR_rbf), and another uses the polynomial function kernel (SVR_poly); (3) The third model group consists of two k-Nearest Neighbors (kNN) Regression models (Altman, 1992, Coomans and Massart, 1982, Cover and Hart, 1967), with the number of neighbors being 3 and 10 (kNN_3 and kNN_10), respectively; (4) The fourth model group belongs to the Decision Tree (DT)- based models (Breiman et al., 1984, Mingers, 1989, Quinlan, 1986). There are three DT models being tested in this study, namely the Classification And Regression Tree (CART) (Breiman et al., 1984), the Random Forest (RF) algorithm (Breiman, 2001), and the Extreme Gradient Boosting Tree (XGBoost) model (Chen and Guestrin, 2016); (5) The fifth group includes two Multiple Layer Perceptron (MLP) Models (Jain et al., 1996), implemented with the hyperbolic tangent activation function (MLP_Tanh), and the polynomial activation function (MLP_poly); Last, (6) the sixth model group contains one popular deep learning model, i.e., the Long-Short-Term-Memory (LSTM) model (Hochreiter and Schmidhuber, 1997). In the following sections, we briefly introduce these employed AI&DM models. For conciseness, we only summarize the key concept used in each model. The detailed mathematical definition is summarized in Supplementary Material Section 1 for interested readers.

2.1. Linear model

Linear regression is one of the most widely used mathematical techniques to predict a target variable or vector (y_i), based on the values of a set of dependent variables, or called feature vector (x_i) (Kutner et al., 2005, Makridakis et al., 2008, Montgomery et al., 2012). A linear relationship is assumed to be existing between the dependent variable and the target (Makridakis et al., 2008, Montgomery et al., 2012). The coefficient vector β_n could be estimated by fitting a linear line between the historical observations of $x_{n,i}$ and y_i. In the context of machine learning, the similar process is called model training. Once the values of the coefficient vector β_n being identified from the training process, the same set of β_n could be used to predict the values of the target variables using a new data point $x_{n,i}$, which is from either in a future phase or a testing dataset that the training process never uses (Marill, 2004, Seber and Lee, 2012). Specifically, for linear regression models, we have the following Eq. (1):

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i} + \varepsilon \quad (1)$$

where y_i is the ith observation of the dependent variable, $x_{1,i}, x_{2,i}, \cdots, x_{n,i}$ are respectively theith observation of the independent variables $x_1, x_2, \cdots, x_n$ (n is the number of variables), $\beta_0, \beta_1, \cdots, \beta_n$ are the model parameters, which are also called regression coefficients. ε is the error term assumed to be normally distributed with zero mean and variance σ^2.

The Linear Ridge regression is a modified model of the basic linear regression model with an L2-norm penalty regularization term (Hoerl and Kennard, 1970, Marquardt and Snee, 1975). The penalty term intentionally shrinks and controls the regression coefficient of the linear model to avoid the poorly determined and high variance coefficient problems. The process of introducing the penalty term to a regression model is called regularization. The goal of regularization is to improve the conditioning of the prediction problem (i.e., overfitting or underfitting), and essentially to reduce the prediction variance when the input variables have some levels of dependency or the training dataset is biased. Adding an L2-norm penalty term onto the linear regression model (Eq.1) could effectively constrain the values of the coefficient vector, and control the bias-variance trade-off (Hastie et al., 2009, Myers and Myers, 1990). Hence, the Linear Ridge regression algorithm may build a model with a fewer number of parameters than the simple linear multivariable regression model, and it is found to be less sensitive and less overfitting than the regular linear model (Lima and Lall, 2010a, Yu and Liong, 2007). The procedure of building in the L-2 panelty term into the standard multi-variable linear regression model is presented in Supplementary Material Section 1.1.

2.2. Support Vector Machine (SVM) and support vector regression (SVR)

The Support Vector Machine (SVM) is a supervised machine learning algorithm, which was introduced as a statistical learning algorithm for complicated data classification and regression (Boser et al., 1992, Cortes and Vapnik, 1995, Vapnik, 2013, Vapnik, 1999). Based on the purpose of use, the SVM can be further categorized into the Support Vector Classification (SVC) and Support Vector Regression (SVR). The key concept of SVM is to identify and search for a hyperplane, i.e., an conceptual and high-dimensional fitting line, which will optimally partition the training data by its feature values. The identified hyperplane will become the decision boundary, and be further used to predict the continuous output in the regression model.

Comparing to the basic linear models, the SVM model acknowledges the presence of non-linearity in the data, i.e., the input features are nonlinearly correlated, making it hard to identify a linear line or plane to effectively separate the input data. To address this issue, the SVM models will first create an N + 1 higher dimensional feature space with N equals to the dimension of original input data. Because in the higher dimensional feature space, all input features will be linearly correlated, the formulation of a linear separation line (i.e., the hyperplane) will be feasible (Gunn, 1998, Noble, 2006, Smola and Schölkopf, 2004). The identified hyperplane will be transferred back to the original N-dimensional space to partition the input data and further used as the regression fitting line. To find the optimal hyperplane in the N + 1 dimentional space, the SVM algorithm implements a learning algorithm that provides a globally optimal solution by minimizing the upper bound of the generalization error between the support points (Deka, 2014, Haykin and Network, 2004, Hipni et al., 2013). In summary, the SVM algorithm consists of two essential steps: 1) to project the input data into a higher dimensional feature space, and 2) to find a global optimal hyperplane to split the data by evaluating the offsets of each data point to this hyperplane (Deka, 2014, Vapnik, 2013).

In the SVM model framework, one of the key parameters is the selection of the kernel function. A kernel function is used to transform the inputs into a required dot product format, as well as to describe how input features are linearly separated in the N + 1 dimensional feature space. In this study, we tested two common kernel functions, namely the radial basis function kernel and the polynomial kernel function, following the prior suggestions from the study in applying SVR on hydrological time series forecasting and reservoir simulation using the SVM models (Adnan et al., 2020, Cheng et al., 2020, Yu et al., 2006a, Zhang et al., 2018):

1. Polynomial kernel function: $K(x_i, x_j) = [\gamma(x_i \cdot x_j) + c]^d$
2. Radial basis kernel function: $K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2)$

where γ is the structural parameter in the polynomial function,c is the residuals, and d is the degree of the polynomial term, x_i and x_j represents the data in the original data space and the transformed N + 1 dimensional space. For conciseness, the detailed mathematical

derivation and representation of the SVM model are summarized in the Supplementary Material Section 1.2 for interested readers.

2.3. k-nearest neighbors (kNN) regression model

The kNN model is an unsupervised machine learning approach, which was originally introduced by Fix (1951). The kNN algorithm was further developed by many others (Altman, 1992, Coomans and Massart, 1982, Cover and Hart, 1967). Over the years, the kNN model has become a simple but effective method for solving both classification and regression problems in a variety of research fields. The key procedure in the kNN model is to group the training samples into multiple classes with the same user-predefined size (k) or the number of data samples in each class. The data separation process follows the rule that the k data points in each class will be the closest with respect to any other new data point in the training sample. In other words, the kNN algorithm partitions the data based on the closeness or the shortest distance among the proximity of a total of k samples (Bhatia, 2010, Cunningham and Delany, 2020). The distance can, in general, be any metric measure, while the standard Euclidean distance is the most common choice (Imandoust and Bolandraftar, 2013). The number of samples (k) can be a user-defined constant or vary based on the local density of points (radius-based neighbor learning). In this study, we set the number of samples in the nearest neighbors as 3 and 10, as kNN_3 and kNN_10, respectively (also see Supplementary Material Section 1.3). Some initial test was carried out to identify the values of this parameter that the number of nearest neighbors beyond 10 did not improve and sometimes deteriorate the model performance. More discussion will be provided in later sections about the choices of AI&DM model parameterization and sensitivity analysis.

2.4. Decision Tree (DT)-based models: CART, RF, and XGBoost Tree model

The Decision Tree (DT) model is a non-parametric, "white-box" statistical learning approach (Breiman et al., 1984, Mingers, 1989, Quinlan, 1986) that uses a tree-structured classifier to recursively partition the training dataset into smaller subsets (i.e., nodes), following identified splitting rules. When splitting the data into smaller subsets, each data partition will primarily follow a simple "true or false" Boolean logic, i.e., whether the value of a feature or data greater or smaller than a threshold (Freund and Mason, 1999, Rokach and Maimon, 2005). By repeatedly partitioning the data into smaller subsets, the final classes will only contain the samples that are distinguishable from other subsets of data based on a sequence of characteristics in the feature space (Buntine and Niblett, 1992). The major advantage of the DT is that it follows a highly intuitive and efficient splitting rule, where decision-makers can easily see the logic of interpreting the data (James et al., 2013, Krishnan et al., 1999, Yang et al., 2020b). Furthermore, the establishment of splitting rules to partition the data, or the so-called tree-growing process, requires less effort for data preparation and pre-processing. In other words, there is no need to make strict assumptions about the distribution of data or the raw value scales of the training data. The tree-based models are also found suitable for dealing with unbalanced data classification and regression (Ganganwar, 2012, Ke et al., 2017, Pradhan, 2013).

However, there are still some disadvantages of the DT model. For example, over- and under-fitting can occur on datasets with a small number of samples, or the training samples value becomes too sparse, the DT model could result in a lack of robustness and low accuracy for unseen data (Brodley and Utgoff, 1995, Dietterich, 1995, Mingers, 1989). To deal with those problems in DT models, primarily two enhancement strategies were developed, i.e., the bagging and boosting techniques. In this study, we employed three DTs models, namely the classical Classification And Regression Tree (CART) (Breiman et al., 1984), the Random Forest (RF) algorithm (Breiman, 2001), and the Extreme Gradient Boosting (XGBoost) Tree algorithm (Chen and Guestrin, 2016). The CART model is one of the early and classical DT models, while the RF and XGBoost algorithms are two newer DT models, which incorporated with the bagging and boosting enhancement techniques, respectively.

The Classification And Regression Tree (CART) was originally introduced by Breiman et al. (1984). It divides the data items into homogenous subsets using binary recursive partitions aiming to classify datasets and has been proven as a powerful tool for both classification and prediction problems (De'ath and Fabricius, 2000, Loh, 2014, Steinberg and Colla, 2009). The CART algorithm provides a base regression model for the later developments of both RF and XGBoost algorithms. The mathematical description of the CART algorithm, its tree-growing process, and splitting criteria are briefly introduced in the Supplementary Material Section 1.4, which follows the summary from Hastie et al. (2009) and the original development from Breiman et al. (1984).

The Random Forest (RF) is an ensemble learning algorithm introduced by Breiman (2001). The key concept used in the RF algorithm is to build and combine multiple candidates of the standard CART models based on a bagging strategy to avoid overfitting and/or underfitting (Liaw and Wiener, 2002). Specifically, the RF algorithm starts with generating several bootstrapped samples of input features from a given training dataset. Then, different CART models are trained on each of the bootstrap samples. During this process, both strong and weak learners are presented. Finally, the model output can be achieved by aggregating the outputs from all the candidate CART models, which are individually built from the bootstrapped samples (Efron, 1987, 1992, Johnson, 2001). The ensemble strategy used in the RF algorithm ensures the prediction model contains both weak learners and strong learners from the original training dataset. By assembling multiple candidates CART models, the final ensemble model will be more robust and less overfitting than the single CART model (Brokamp et al., 2017). Mathematically, let's assume that there are E number of trees and their model output is y_i ($i = 1,2,\cdots,E$). Then, the final output of the RF regression can be obtained by averaging the outputs of all the trees as presented in Eq. (2):

$$\hat{y} = \frac{1}{E}\sum_{i=1}^{E} y_i \tag{2}$$

where y_i is the prediction results from each ith CART model trained on different bootstrapped data and input features; E is a user-defined parameter and represents the total number of ensemble candidate CART models to be used as ensemble candidates; And $\hat{y}$ is the final ensemble model of the RF algorithm. The detailed model parameters used in the RF algorithm are summarized in Supplementary Material Section 1.5.

In contrast to the RF and CART algorithm, the XGBoost Tree model is another recently developed DT algorithm, which was introduced by Chen and Guestrin (2016). The XGBoost Tree algorithm was based on the greedy gradient boosting framework for parallel tree boosting (Friedman, 2001, 2002), but was enhanced by a novel tree learning algorithm is for handling sparse data and a theoretically justified weighted quantile sketch procedure to enable the handling of instance weights (Chen and Guestrin, 2016). Like the RF model, the XGBoost model uses an ensemble of CART. However, unlike the RF algorithm that develops ensemble candidates on bootstrapped training data and features, the XGBoost model is further designed to "boost" the performance of weak learners by performing additive training strategies (Boutaba et al., 2018, Ke et al., 2017). The Gradient Boosting approach allows new models to be trained to predict the residuals (i.e., errors) of prior models. In other words, instead of training multiple models in isolation of one another, the gradient boosting strategy trains models in succession. The new model will be trained to correct the errors made by the previously trained models. The "boosted" models are then added together

sequentially until no further improvements could be made. Eventually, by developing an ensemble of the "boosted" model candidates, the model's predictive performance will be improved as the performance of all the candidate ensemble models have been "boosted" by correcting the residuals. The mathematical derivation of gradient boosting tree from a based CART algorithm is briefly introduced in our Supplementary Material Section 1.6, and interested readers shall also refer to the original development research from Chen and Guestrin (2016).

2.5. Multiple Layer Perceptron (MLP) models or Artificial Neural Network (ANN)

The Multiple Layer Perceptron (MLP) model or Artificial Neural Networks (ANN) model (Hoskins and Himmelblau, 1988, McCulloch and Pitts, 1943) is one of the most widely used machine learning algorithms in many fields. The development of the MLP or ANN model was inspired by the biological neural network of the human brain (Jain et al., 1996). The key concept of ANN is to build a network-like model structure with the goal of finding the intrinsic patterns or relationships in a given dataset through a learning (or training) process called back-propagation (ASCE, 2000, Goh, 1995, Hecht-Nielsen, 1992). In general, an ANN model consists of three types of basic layers (e.g., the input, hidden, and output layers), and the nodes between these layers are interconnected by tunable connection weight parameters. The input layer takes either the raw or normalized training data, and the value of one particular node in the input layer is weighted and passed to a successor node in the hidden layer. The corresponding node in the hidden layer will process the weighted sum value and information obtained from the input layer. Within each node in the hidden layer, the weighted sum values from prior nodes will be further processed by a pre-defined transformation function, or called activation function, and then becomes the output from the hidden node. The same weighting procedure is performed to combine the outputs from hidden layer nodes into the information feed to the nodes in the output layer node. If an MPL model has multiple hidden layers, the procedures of both weighting and transformation are repeated until the output layer is reached. The values of the connection weights are identified by calculating the accumulated errors from all output layers and hidden layers and by minimizing a loss function.

In the training process, the role of the activation function is important, as it enables nonlinear statistical modeling with complex data (Hsu et al., 1995, Zealand et al., 1999). In this study, we used two types of activation functions: the Hyperbolic Tangent activation function (Tanh) and the logistic activation function (Log). The Tangent and Logistic functions for any variable t are defined by Eqs. (3) and (4), respectively.

$$\mathrm{Tanh}(t) = \frac{2}{1+e^{-2t}} - 1 \tag{3}$$

$$\mathrm{logistic}(t) = \frac{K}{1+Ce^{-rx}} \tag{4}$$

where, C is the constant from integration, r is the proportionality constant, and K is the threshold limit. Assuming the K, C, and r all equal to 1, the logistic activation will become the standard sigmoid activation function. The detailed mathematical definition of an MLP or ANN model, and the applied model parameters (number of layers, number of hidden nodes, learning rate, regularization term, training optimizer, maximum iteration numbers, etc.) are introduced in Supplementary Material Section 1.7.

2.6. Long short term Memory (LSTM) model

The LSTM model is a deep learning algorithm introduced by Hochreiter and Schmidhuber (1997). It is a particular type of recurrent neural network (RNN) that is specialized in dealing with time-sequential data prediction (Gers et al., 2002, Graves and Schmidhuber, 2005, Rumelhart et al., 1986). Differs from the standard feedforward ANN or MLP models, the LSTM model assumes the training samples are temporally correlated. Instead of importing continuous data as independent training samples, the LSTM model trains the time series of both feature and target variables with a user-defined time step. The key benefit of the LSTM model is that it is operated by using memory cells with input, output, and forget gates, which capable of not only learning long-term range decencies but also overcoming the gradient vanishing and exploding problems of RNN (Hochreiter, 1998, Hochreiter et al., 2001). The basic structure of LSTM consists of input, hidden, and output layers similar to the MLP model, and it has an unfolded structure due to the recurrent connections of hidden states in the hidden layer. The mathematical description of the internal operations of a memory cell in LSTM can be briefly described in Gers et al. (1999) and Yu et al. (2019) and summarized in Supplementary Material Section 1.8.

2.7. Statistical measures

In this study, we employed six different statistical measurements based on the suggestions from Moriasi et al. (2007). The employed statistical measures include the Correlation Coefficient (CORR), the Root Mean Square Error (RMSE), the Nash-Sutcliffe Model Efficiency Coefficient (NSE) (Nash and Sutcliffe, 1970), the Kling-Gupta efficiency (KGE) (Gupta et al. 2009), the RMSE-observation standard deviation ratio (RSR), and the Percentage of Biases (PBIAS). The equations, value range, and optimal values are listed in Table 1. In Table 1, the $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated reservoir daily discharges at the time step t, respectively. The $\overline{Q}_{obs,i}$ and $\overline{Q}_{sim,i}$ represent the mean of the observed and simulated values, respectively. The variable n is the total number of time steps in the compared time series. In the KGE calculation, μ_s and σ_s represent the mean and standard deviation of the simulated discharges, and μ_o and σ_o are the mean and standard deviation of the observed hydropower releases, respectively.

The selection of these statistical measures is based on the commonly accepted standards of streamflow simulation and model evaluation in the field of hydrology. Specifically, the measurements of CORR, RMSE, and NSE are widely used statistical measures to quantify how the simulated streamflow matches the observed streamflow. CORR measures how a simulated time series will vary in corresponding to observation. RMSE quantifies the accumulated biases between the simulation and observation, and it is a similar measure of the Mean Absolute Error (MAE) with relatively higher sensitivity due to the mathematical operation of taking the square of the error term. NSE is a combined statistic of both RMSE and CORR, in which both the bias and temporal variation will affect the NSE value. In addition, according to Gupta et al. (2009), the Kling-Gupta efficiency (KGE) was developed amending some shortcomings of the NSE measurements. The KGE measurement was able to decompose the NSE values into linear correlation, bias, and variability components between simulated and observed time series, and it is thusly able to analyze the relative importance of each of the terms that contribute to the NSE index. According to Moriasi et al. (2007) and Singh et al. (2005), the RMSE-observation standard deviation ratio (RSR) standardizes RMSE using the observations standard deviation, and it combines both an error-index and the additional information (Legates and McCabe, 1999). According to Moriasi et al. (2007) and Gupta et al. (2009), the RMSE and RSR with a value equal to zero, and/or the NSE or KGE value of 1, is the indication of the best accuracy between a model simulated time series and the observation. The larger the RMSE and RSR value, or the smaller the NSE and KGE value, the poorer performance a model is. Some reference ranges for the NSE and KGE values are summarized below (Gupta et al., 2009, Moriasi et al., 2007): Unsatisfactory (NSE $\leq$ 0.4, Acceptable (0.40 $<$ NSE/KGE $\leq$ 0.50), Satisfactory (0.50 $<$ NSE/KGE $\leq$ 0.65), Good (0.65 $<$ NSE/KGE $\leq$ 0.75) and Very Good (0.75 $<$ NSE/KGE $\leq$ 1.00).

The variable PBIAS quantifies the percentage of biases between a

Table 1
Information about employed statistical measurement.

Measure	Equation	Range	Ideal Value
Correlation Coefficient (CORR)	$\frac{\sum_{i=1}^{n}\left(\left(Q_{sim,i}-\overline{Q}_{sim,i}\right)\left(Q_{obs,i}-\overline{Q}_{obs,i}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_{sim,i}-\overline{Q}_{sim,i}\right)^{2}\sum_{i=0}^{n}\left(Q_{obs,i}-\overline{Q}_{obs,i}\right)^{2}}}$	−1 to 1	1
Root Mean Square Error (RMSE)	$\sqrt{\sum_{i=0}^{n}\left(Q_{sim,i}-Q_{obs,i}\right)^{2}/n}$	0 to ∞	0
Nash-Sutcliffe Model Efficiency Coefficient (NSE)	$1-\frac{\sum_{i=1}^{n}\left(Q_{obs,i}-Q_{sim,i}\right)^{2}}{\sum_{i=1}^{n}\left(Q_{obs,i}-\overline{Q}_{obs,i}\right)^{2}}$	-∞ to 1	1
Kling-Gupta efficiency (KGE)	$1-\sqrt{\left(1-CORR\right)^{2}+\left(1-\sigma s/\sigma o\right)^{2}+\left(1-\mu s/\mu o\right)^{2}}$	-∞ to 1	1
RMSE-observation standard deviation ratio (RSR)	$\frac{\sqrt{\Sigma_{i=1}^{n}\left(Q_{obs,i}-Q_{sim,i}\right)^{2}}}{\sqrt{\Sigma_{i=1}^{n}\left(Q_{obs,i}-\overline{Q}_{obs,i}\right)^{2}}}$	0 to ∞	0
Percent bias (PBIAS)	$\frac{\Sigma_{i=1}^{n}\left(Q_{obs,i}-Q_{sim,i}\right)*100}{\Sigma_{i=1}^{T}\left(Q_{obs,i}\right)}$	-∞ to ∞	0.0

simulated time series and the reference. Positive and negative values indicate underestimation and overestimation relative to the measured data, respectively. While values of 0 are desired, satisfactory values of PBIAS vary for different constituents and must consider the level of measurement uncertainty. The higher the PBIAS value, either towards positive or negative, the worser the model accuracy.

Besides the tabular and numerical statistical measurements, in this study, we also employed a graphical model evaluation tool, termed the Taylor Diagram (Taylor, 2001). The Taylor diagram provides a way of showing how three complementary model performance statistics, i.e., the correlation coefficient R, the standard deviation (sigma), and the (centered) root-mean-square error, simultaneously in one single 2-D graph. According to Taylor (2001), the plotting of multiple statistics in one graph is based on the geometric relationship (cosine-law) between the correlation coefficient (R), the centered Root Mean Square Error (RMSE), and the standard deviations between the simulation and observation. The Taylor diagram has been extensively used in climate model studies (Miao et al., 2014, Tao et al., 2018, Yang et al., 2018), as well as the topic of AI&DM model evaluation and comparison in the field of hydrology (Adnan et al., 2021, Kargar et al., 2020, Shabani et al., 2020).

3. Data and experiments setting

3.1. Study cases

The Upper Colorado River Basin is comprised of four states, including Colorado, New Mexico, Utah and Wyoming. Collectively, the Upper Colorado Basin contributes a majority of the fresh surface water supplies coming into the entire Colorado River Basin, primarily through winter snowpack and streamflow. With the impacts of climate change altering the amount of snowpack and timing of spring runoff, water supply in the Colorado River is increasingly strained. The reservoir systems in the Upper Colorado Region plays an inevitably crucial role in managing the surface water for multiple uses, such as flood control, hydropower, creation (ski/boating/fishing), as well as direct water supplies to residential, industrial, and irrigation over the states of Colorado, New Mexico, Utah, and Wyoming. The Colorado River provides water to nearly 40 million people and drives a $1.4 trillion economy. Climate change and increasing water demand due to an expanding population is and will continue to present significant challenges. If left unaddressed, the varying weather and climate will impact our regional and national economies, degrade the environment, challenge our agricultural heritage and food production, and limit recreational opportunities from fishing and boating to skiing. Therefore, the focus on the Upper Colorado basin reflects the importance of both regional and national social-economic benefits.

The upper Colorado region consists of complex terrains and is prone to changing climate of precipitation, snowpack, and temperature. The refills of the reservoir system over Upper Colorado primarily are from the spring snowpack melting and direct streamflow from rainfall-runoff hydrology. This setting is typical in many other regions where reservoir plays the role of changing the timing and amount of water flowing to downstream regions. As mentioned in the introduction, the development of advanced management and decision-making tools will significantly promote our capability in controlling the surface water resources, and better enable us to mitigate the potential impacts of climate. The Upper Colorado region, itself, serves as an ideal region that the impacts of climate and extreme rainfall/snows, could easily reflect in the reservoir inflows variability. Based on the above two reasons, the reservoir system within the Upper Colorado River basin is selected for this comparison study of different AI&DM models.

In this research, we choose 33 reservoirs in the Upper Colorado region under the jurisdiction of the U.S. Bureau of Reclamation (USBR). The following Table 2 lists the short name, full name, locations, and data lengths employed in this research. The locations of the selected reservoirs are presented in Fig. 1. We choose these reservoirs based on the criteria that (1) the reservoir shall have a complete set of data records for reservoir inflow, storage, and outflow at a daily time step, and (2) the data records are continuous without significant missing data over ten days. The reservoir inflow, storage, and outflow data are obtained from the USBR water operation archive (https://www.usbr.gov/rsvrWater/HistoricalApp.html). In this data repository, some reservoirs among the 33 selected ones have an earlier start date and longer data record. However, in our initial data screening, we found some data were either missing or unavailable at earlier records, and the numbers of missing data are significant. Therefore, in the experiments, we manually checked the data values for each reservoir and selected the start data when all daily inflow, storage, and outflow data are continuously available from that year. In other words, the starting date for each reservoir is different (Table 2), while the ending date for all simulations is set consistently as December 31st, 2020. Among the selected 33 reservoir cases, the shortest data record is about 10 + years, and the longest data record is about 50 + years. Table 2 also lists the elevation of each reservoir, which ranges from 1323 to 2847 m above sea-level.

3.2. Experiment setting

In this study, we use the daily inflow, storage, and seasonality (Months) as model inputs to estimate the daily reservoir outflow. According to several prior studies on the topic of reservoir time series analysis (Hejazi et al., 2008, Zhao et al., 2012), the one- and two-time

Table 2
Information of the Selected Reservoirs Over the Upper Colorado River Basin.

Initials	Names	Lat	Lon	Data Start Date	Data Length (Years)	Elevation (Meters)
BSR	Big Sandy Reservoir	42.24923	−109.429	1/1/1990	30	2060
CAU	Causey Reservoir	41.29019	−111.583	1/1/1999	21	1745
CRY	Crystal Reservoir	38.45359	−107.335	1/1/1978	42	2251
DCR	Deer Creek Reservoir	40.40667	−111.527	1/1/1987	33	1653
DIL	Dillon Reservoir	39.6074	−106.055	1/1/1985	35	2751
ECH	Echo Reservoir	40.96486	−111.432	1/1/1967	53	1691
ECR	East Canyon Reservoir	40.92053	−111.601	1/1/1992	28	1749
EBR	Elephant Butte Reservoir	33.15349	−107.191	1/1/2007	13	1323
FGR	Flaming Gorge Reservoir	40.91499	−109.422	1/1/1963	57	1828
FON	Fontenelle Reservoir	42.0283	−110.061	1/1/1990	30	1976
GMR	Green Mountain Reservoir	39.8783	−106.33	1/1/1977	43	2406
HNR	Huntington North Reservoir	39.34173	−110.947	1/1/1999	21	1774
HYR	Hyrum Reservoir	41.62663	−111.872	1/1/1999	21	1427
JOR	Jordanelle Reservoir	40.40729	−111.528	1/1/1997	23	1636
JVR	Joes Valley Reservoir	39.28848	−111.269	1/1/1996	24	2129
LCR	Lost Creek Reservoir	41.18417	−111.399	1/1/1998	22	1824
LEM	Lemon Reservoir	37.38171	−107.661	1/1/1965	55	2478
MCP	Mcphee Reservoir	37.57651	−108.572	1/1/1991	29	2073
MCR	Meeks Cabin Reservoir	41.02533	−110.58	1/1/1998	22	2647
MPR	Morrow Point Reservoir	38.4518	−107.538	1/1/1977	43	2184
NAV	Navajo Reservoir	36.80237	−107.613	1/1/1986	34	1801
PIN	Pineview Reservoir	41.25402	−111.843	1/1/1990	30	1495
RFR	Red Fleet Reservoir	40.58028	−109.442	1/1/1989	31	1721
RID	Ridgway Reservoir	38.19918	−107.742	1/1/1990	30	2101
ROC	Rockport Reservoir	40.79	−111.404	1/1/1969	51	1807
RUE	Ruedi Reservoir	39.3631	−106.818	1/1/1980	40	2349
SCO	Scofield Reservoir	39.7862	−111.119	1/1/1996	24	2338
SJR	Silver Jack Reservoir	38.23207	−107.543	1/1/1992	28	2725
STA	Starvation Reservoir	40.18876	−110.444	1/1/1981	39	1700
STE	Steinaker Reservoir	40.50567	−109.531	1/1/1976	44	1655
TPR	Taylor Park Reservoir	38.818	−106.607	1/1/1963	57	2847
USR	Upper Stillwater Reservoir	40.56	−110.699	1/1/1991	29	2445
VAL	Vallecito Reservoir	37.37775	−107.575	1/1/1986	34	2318

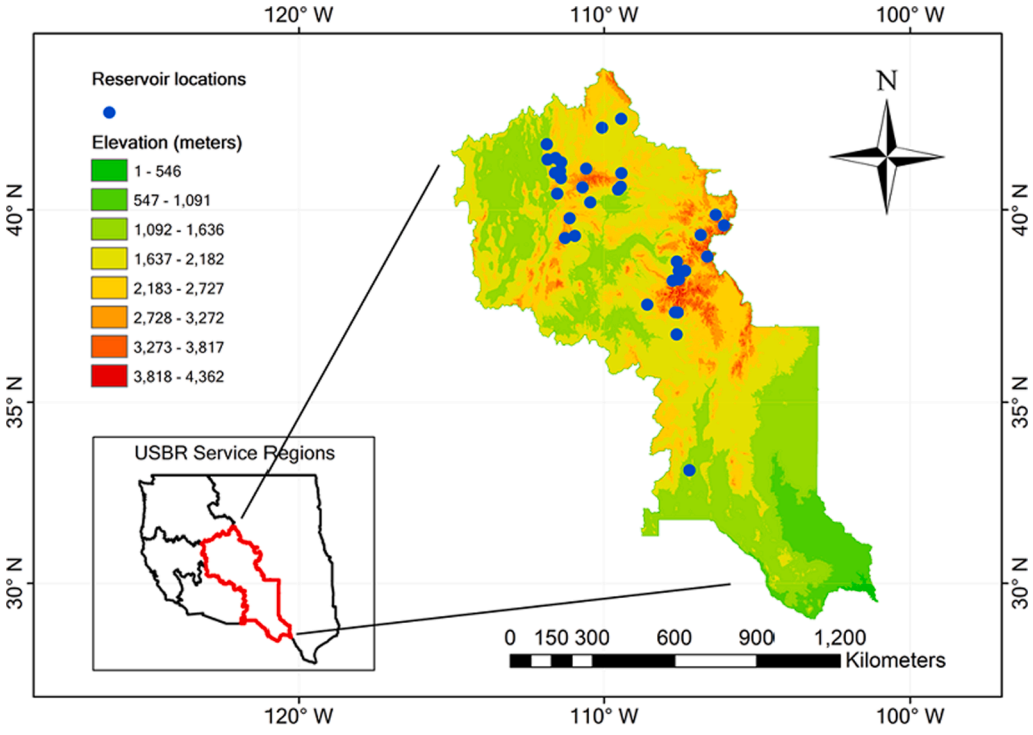

Fig. 1. Locations of the Selected Reservoirs.

step delayed reservoir inflow and storage information have strong correlations to the current time step reservoir outflow. Following their findings, we designed three different input scenarios to drive the AI&DM models, and our experiment design which is shown in the following Fig. 2. In the first designed simulation Scenario (S1), we use the current inflow, storage, and seasonality (Months) as the default inputs to AI&DM models. In the second simulation Scenario No. 2 (S2), the model input categories further extend to cover the 1-step (1 day) delayed

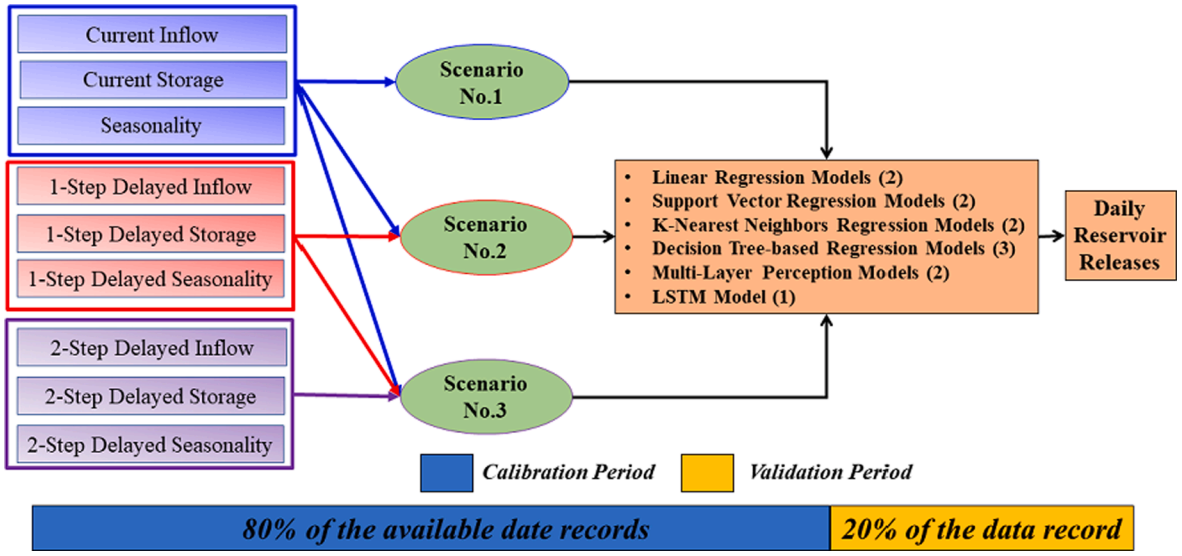

Fig. 2. Experiment Design.

information along with the default model inputs under S1. In the last and third Scenario (S3), we design the AI&DM model inputs to further include both the one- and two-step delayed inflow, storage, and seasonality, as well as the default input set defined under S1. In other words, the complexity and number of inputs to the AI&DM models are increasing from S1, S2, to S3. We expect the performance of AI&DM models will be sensitive to the input training time series being used and the differences will be analyzed in the later result section.

In all of the performed experiments, 80% of the data is used to train the AI&DM models, and reminder 20% of the data record is used as validation. The underline assumption is that the employed AI&DM models will be trained on a subset of the entire data, and be tested on a new subset of data that the model never sees during its training process. Note that this partition ratio of data (80/20) is consistent across each reservoir case. However, since each reservoir has different data lengths (Table 2), the data lengths used in training and validation are different for each reservoir. Nevertheless, for each reservoir, the training and validation data used are identical across different AI&DM models, which ensures the model comparison is fair.

Last but not least, when we train individual AI&DM models, all model inputs and target values are normalized into the range of [−1,1]. After the model training process and during the validation period, the model predicted values are transformed back to the normal range of daily outflows for each reservoir. The evaluation and assessment of statistical measurements are conducted using the transformed model predictions, which are in the normal range of daily reservoir outflow observations instead of the normalized range of [−1,1]. Most of the model hyperparameters are set prior to the experiments by manual trial-and-error to avoid overfitting, and this is also the reason we did not put a separate third sub-dataset as testing. The specific model hyperparameters and training settings for each employed AI&DM model are listed in Supplementary Material Section 1 for interested readers.

4. Results

4.1. Statistical results of the baseline scenario No.1 (S1)

In this section, we will present the obtained statistics of all reservoirs under the baseline Scenario No.1 (S1), and then compares the performance across different models and reservoir cases. Note that all calculated raw statistical measurements between simulated and observed reservoir outflows under all simulation scenarios are presented in the Supplementary Material Tables 1–33, and this section will present a further summary and analysis. Specifically, in the following Table 3, we summarize the maximum, minimum, and average of the statistical measures across all employed AI&DM models over the validation periods for each reservoir. These maximum, minimum, and average statistical values are drawn from the raw statistical performance of each individual reservoir (Supplementary Material Tables 1–33) across all employed AI&DM models. In order to obtain an overview of the statistical performance of all reservoirs, at the ending rows of Table 3, we further take a numerical maximum, minimum, and average of all corresponding values from the statistical summary of prior rows of Table 3 for all reservoirs). The results in Table 3 indicate how well the employed AI&DM methods could capture the variations of daily reservoir releases regardless individual model used. The ending bolded rows of Table 3 indicate how well the statistical performance between simulation and observation across all employed reservoir cases in general.

According to Table 3, in most of the reservoir cases, the employed AI&DM models could achieve satisfactory statistical performance over the validation periods, as evidenced by the minimum values row for each reservoir. The worst CORR value (e.g., “CORR” column and “Min” row for each reservoir) is 0.215 in the reservoir case of STA, followed by the second-worst CORR value being 0.295 in the case of STE. However, the average of all of the worst (“Min”) CORR values across all reservoir cases is 0.634 (e.g., “CORR” column and “Min” row at the “Average” section at the end of Table 3). The worst (“Min”) NSE value is observed as −0.476 for reservoir NAV, followed by the second-worst NSE value being −0.133 in the reservoir case of FGR. Nevertheless, in most of the cases, the worst NSE values across all reservoir cases are consistently above 0.5, which indicates still a good match between the simulated and observed daily reservoir outflow according to the NSE satisfactory category set by Moriasi et al. (2007). The average of all worst NSE values in all employed reservoirs is 0.351 (e.g., “NSE” column and “Min” row at “Average” section at the end Table 3). Table 3 also indicates the worst KGE value is −0.242 for reservoir USR as compared to other cases. The average of all of the worst KGE across all studied cases is 0.398, referred to the “average” row at the end of Table 3. Except for these worst scenario cases, in general, the statistical measures are rather satisfactory with high CORR, NSE, and KGE minimum values for each reservoir (i.e., see individual “Min” rows under each reservoir case).

For PBIAS, the highest absolute bias in the simulated time series appears to be −122% in the reservoir case of USR, indicating at least one of the employed AI&DM models significantly underestimated the reservoir outflow. However, in the same case of USR, the average PBIAS value is −25%, which means if averaging the performance of all

Table 3
The summary of statistical measures for all reservoirs over validation periods.

Reservoir		CORR	RMSE	NSE	KGE	PBIAS	RSR
	Max	0.930	3.037	0.847	0.776	20.731	0.669
BSR	Min	0.786	1.775	0.552	0.461	−10.337	0.391
	Ave	0.871	2.336	0.726	0.668	6.521	0.515
	Max	0.978	0.970	0.955	0.950	9.659	0.339
CAU	Min	0.947	0.606	0.885	0.754	−15.952	0.212
	Ave	0.962	0.806	0.918	0.877	1.037	0.282
	Max	0.995	10.213	0.990	0.969	7.372	0.291
CRY	Min	0.959	3.468	0.916	0.854	−9.771	0.099
	Ave	0.989	5.393	0.974	0.936	−1.921	0.153
	Max	0.904	5.904	0.810	0.899	2.511	0.712
DCR	Min	0.760	3.618	0.493	0.434	−10.209	0.436
	Ave	0.851	4.468	0.701	0.707	−4.787	0.539
	Max	0.821	6.531	0.664	0.787	13.177	0.739
DIL	Min	0.714	5.117	0.453	0.584	2.635	0.579
	Ave	0.785	5.612	0.594	0.727	7.018	0.635
	Max	0.912	6.713	0.822	0.794	12.582	0.759
ECH	Min	0.717	3.735	0.424	0.582	−24.383	0.422
	Ave	0.848	4.788	0.698	0.703	−9.329	0.541
	Max	0.830	23.222	0.689	0.800	20.408	0.997
EBR	Min	0.266	12.989	0.006	0.069	−6.816	0.558
	Ave	0.703	16.195	0.496	0.538	6.832	0.695
	Max	0.858	1.268	0.731	0.837	1.491	0.957
ECR	Min	0.359	0.687	0.084	0.219	−13.273	0.519
	Ave	0.739	0.891	0.527	0.664	−7.336	0.673
	Max	0.681	47.877	0.352	0.337	46.745	1.064
FGR	Min	0.391	36.210	−0.133	0.178	−13.729	0.805
	Ave	0.577	39.190	0.235	0.286	−2.090	0.871
	Max	0.897	32.645	0.782	0.751	16.703	0.694
FON	Min	0.785	21.983	0.519	0.431	−1.832	0.467
	Ave	0.844	26.293	0.684	0.700	11.353	0.559
	Max	0.887	7.122	0.782	0.803	6.842	0.707
GMR	Min	0.724	4.708	0.500	0.490	−2.229	0.467
	Ave	0.820	5.801	0.663	0.723	1.887	0.576
	Max	0.908	0.477	0.824	0.851	40.311	0.862
HNR	Min	0.694	0.232	0.256	0.216	−2.290	0.419
	Ave	0.840	0.348	0.690	0.607	26.263	0.629
	Max	0.932	1.775	0.866	0.922	8.204	0.484
HYR	Min	0.881	1.341	0.766	0.751	0.018	0.366
	Ave	0.916	1.482	0.836	0.880	3.192	0.404
	Max	0.884	5.899	0.767	0.812	7.195	0.721
JOR	Min	0.700	3.950	0.480	0.475	−5.718	0.483
	Ave	0.810	4.818	0.646	0.686	3.316	0.589
	Max	0.906	1.896	0.811	0.886	8.643	0.791
JVR	Min	0.642	1.044	0.375	0.593	−11.017	0.435
	Ave	0.808	1.421	0.634	0.771	−6.006	0.593
	Max	0.781	0.849	0.601	0.744	34.928	0.891
LCR	Min	0.508	0.602	0.206	0.113	−1.209	0.632
	Ave	0.719	0.685	0.476	0.531	8.868	0.719
	Max	0.892	2.117	0.788	0.806	12.639	0.697
LEM	Min	0.720	1.400	0.515	0.611	–22.832	0.461
	Ave	0.823	1.735	0.667	0.721	−10.363	0.571
	Max	0.738	8.781	0.533	0.703	45.681	0.850
MCP	Min	0.628	7.065	0.278	0.275	−49.236	0.684
	Ave	0.690	7.749	0.435	0.574	−5.466	0.750
	Max	0.958	3.333	0.907	0.891	15.077	0.464
MCR	Min	0.889	2.191	0.785	0.792	−8.087	0.305
	Ave	0.911	2.990	0.825	0.837	6.815	0.416
	Max	0.627	29.212	0.365	0.444	−4.199	0.897
MPR	Min	0.500	25.958	0.196	0.192	−20.276	0.797
	Ave	0.574	27.449	0.289	0.330	−10.556	0.843
	Max	0.641	26.968	0.378	0.422	19.271	1.215
NAV	Min	0.375	17.514	−0.476	0.082	−37.837	0.789
	Ave	0.480	22.295	−0.025	0.330	−24.477	1.004
	Max	0.832	7.350	0.682	0.745	1.425	0.925
PIN	Min	0.502	4.478	0.144	0.428	−14.819	0.564
	Ave	0.681	5.878	0.445	0.534	−7.075	0.740
	Max	0.873	0.740	0.739	0.806	−14.205	0.778
RFR	Min	0.689	0.485	0.394	0.641	−28.875	0.511
	Ave	0.817	0.584	0.614	0.735	−17.108	0.615
	Max	0.901	4.024	0.809	0.868	7.173	0.810
RID	Min	0.667	2.171	0.345	0.403	−31.254	0.437
	Ave	0.836	2.740	0.686	0.766	−3.223	0.551
	Max	0.863	3.775	0.714	0.716	37.266	0.742
ROC	Min	0.721	2.718	0.449	0.583	−26.091	0.534

(*continued on next page*)

Table 3 (*continued*)

Reservoir		CORR	RMSE	NSE	KGE	PBIAS	RSR
	Ave	0.804	3.139	0.616	0.671	−12.868	0.617
	Max	0.711	2.547	0.504	0.607	−1.870	0.913
RUE	Min	0.560	1.965	0.167	0.321	–33.155	0.704
	Ave	0.628	2.247	0.349	0.485	−9.155	0.805
	Max	0.777	1.817	0.535	0.673	−14.097	1.036
SCO	Min	0.361	1.196	−0.073	0.123	−49.197	0.682
	Ave	0.642	1.475	0.277	0.478	–32.361	0.841
	Max	0.963	1.764	0.917	0.942	14.631	0.422
SJR	Min	0.910	1.206	0.822	0.805	−13.289	0.289
	Ave	0.945	1.410	0.885	0.892	5.067	0.337
	Max	0.701	4.208	0.482	0.631	5.313	0.991
STA	Min	0.215	3.055	0.018	−0.024	−21.541	0.720
	Ave	0.547	3.558	0.286	0.415	−5.417	0.838
	Max	0.889	1.469	0.722	0.681	46.192	1.001
STE	Min	0.295	0.773	−0.003	−0.125	16.780	0.527
	Ave	0.746	0.996	0.513	0.442	28.887	0.679
	Max	0.824	3.962	0.652	0.620	36.337	0.855
TPR	Min	0.595	2.731	0.268	0.416	−10.537	0.590
	Ave	0.770	3.159	0.530	0.543	−0.778	0.682
	Max	0.927	5.197	0.830	0.739	20.940	0.822
USR	Min	0.639	2.609	0.324	−0.242	−122.843	0.413
	Ave	0.842	3.600	0.664	0.497	−25.450	0.570
	Max	0.926	6.783	0.839	0.841	21.699	0.598
VAL	Min	0.811	4.549	0.642	0.647	−11.141	0.401
	Ave	0.875	5.555	0.755	0.757	−3.051	0.490
Average	**Max**	**0.853**	**8.195**	**0.718**	**0.759**	**15.357**	**0.779**
	Min	**0.634**	**5.580**	**0.351**	**0.398**	**−18.495**	**0.506**
	Ave	**0.779**	**6.578**	**0.582**	**0.637**	**−2.478**	**0.616**

employed AI&DM models, the prediction bias could be reduced, and the high bias of underestimation was only associated with few modeling scenarios. When taking a further examination of raw statistics from the Supplementary Material Tables 32, 18, and 27, we found out that using SVR_rbf, SVR_poly, MPL_Tanh model in simulating the outflows from the reservoir of UBR, MCP, and SCO, individually, the PBIAS values are −122.843%, −49.236%, −49.197%, respectively. For the last RSR statistical measurement, the performance of all employed AI&DM models are similar, and the worst ("Max") value is observed to be 1.215 for the reservoir case of NAV, followed by the second worst value of 1.001 observed for the reservoir case of STA. In general, Table 3 shows that when simulating the daily reservoir outflow decisions using default reservoir inflow and storage time series, though the performance of each AI&DM model will vary, the averaged statistical performance over all 12 employed models and all 33 reservoir cases are still satisfactory with the following averaged statistical value ranges: $0.634 \leq \text{CORR} \leq 0.853$, $5.580 \leq \text{RMSE(Unit:m3/s)} \leq 8.195$, $0.351 \leq \text{NSE} \leq 0.718$, $0.398 \leq \text{KGE} \leq 0.759$, $-18.495 \leq \text{PBIAS(Unit:\%)} \leq 15.357$, and $0.506 \leq \text{RSR} \leq 0.779$, for all the reservoir cases.

4.2. Reservoir outflow simulation and model taylor diagram

In the following Figs. 3–5, we plot the AI&DM model simulated daily reservoir discharge against the observed time series for reservoirs BSR, CAU, CRY, DCR, DIL, EBR, ECH, ECR, FGR, FON, and GMR in Fig. 3; HNR, HYR, JOR, JVR, LCR, LEM, MCP, MCR, MPR, NAV, and PIN in Fig. 4; and RFR, RID, ROC, RUE, SCO, SJR, STA, STE, TPR, USR, and VAL in Fig. 5. The data records in Figs. 3–5 cover the validation periods for each reservoir listed in prior Table 2. In Figs. 3–5, the y-axis is daily discharge from the reservoir (m3/s) and the x-axis is the number of days in the validation periods. Different color lines indicate simulated reservoir outflows by different AI&DM models, and the observed time series are plotted with dotted black lines. Aside from each subplot in Figs. 3–5, we also draw the Taylor Diagram showing the model performance by color dots. In all the Taylor diagrams, the observation is located at the bottom axis. The closer of the color dots to the observation, the better performance in the context of the Pearson correlation coefficient, the root-mean-square error (RMSE) error, and the standard deviation by geometric cosine law (Taylor, 2001). According to Figs. 3–5, the variation of daily reservoir releases could be well captured with many AI&DM models, though the simulated reservoir outflows may deviate from observations from one simulation period to another. Among all studied reservoirs under S1, the performance of AI&DM models could also vary from one case to another.

In the prior statistical result (Table 3), we identify the reservoir cases FGR, STA, STE, NAV, and USR, which are contributing to the worst scenarios of different statistical measurements. In this section, we further analyze the results for these cases along with the time series plots and Taylor diagrams presented in Figs. 3–5**.** The simulated results for reservoir FGR, STA, STE, NAV, and USR are presented in Fig. 3(i), Fig. 5 (g), Fig. 5(h), Fig. 4(j), and Fig. 5 (j)**,** respectively.

According to the model simulated results of cases STA in Fig. 5(g), we observe that both Linear Model, the Ridge Model, and the LSTM model failed to identify the patterns of reservoir releases. These models showed significant overestimation over low flow conditions, whereas the reservoir release shall be close to zero, and demonstrated underestimation over high flow conditions. For example, the simulated reservoir outflow by the Linear, Ridge, and the LSTM models are fluctuating around 5 m^3/s, without capturing the outflow peaks that are above 10 m^3/s.

In contrast, the MLP_Log model could well capture the low flow conditions but exhibit some overestimation in a few peaks in these reservoir cases. By examining the detailed models' statistical measures in reservoir STA (Supplementary Material Table 29 Under Scenarios No.1), the CORR values obtained by the Linear Model, Ridge Model, and the MLP model with logistic activation function are 0.215, 0.215, and 0.255, respectively, while other models could reach a much higher CORR value, exceeding 0.540 in most of the employed AI&DM models. Similar behaviors of the linear model, ridge model, and the MLP_Log model are also observed in another reservoir case of SCO (Fig. 5e). For the reservoir case of SCO, the CORR values obtained by the Linear Model, Ridge Model, and MLP_Log model are 0.295, 0.295, and 0.361, respectively, under the default scenario No.1 (Supplementary Material Table 27). However, in other cases, such as reservoir EBR (Fig. 3f), ECR (Fig. 3h), and STE (Fig. 5h), the obtained CORR values by the MLP_Log model are consistently and significantly higher than the Linear Model

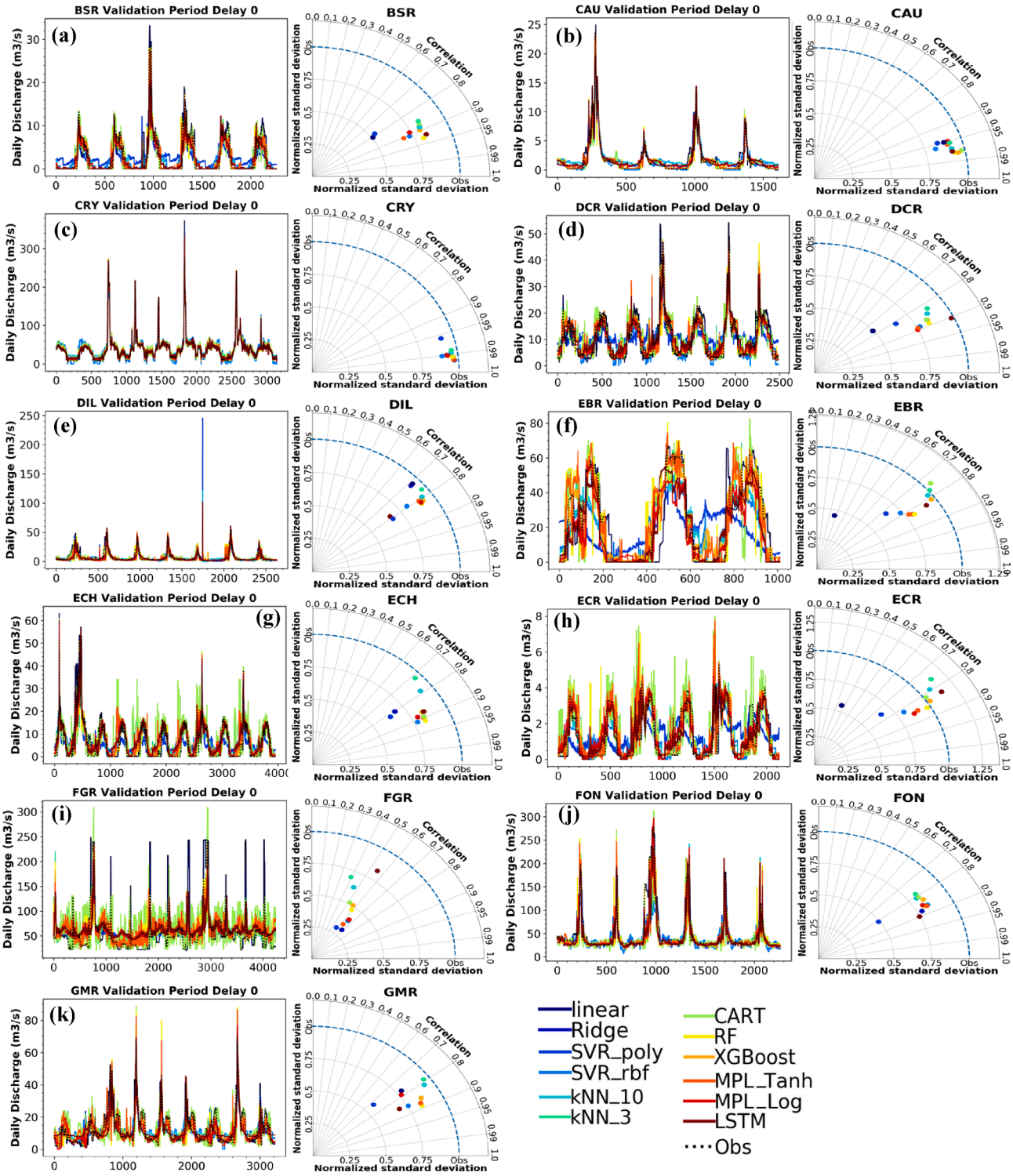

Fig. 3. Simulated and observed daily reservoir discharges and Taylor Diagrams for different AI&DM models - Part I.

and Ridge Model (Supplementary Material Tables 6, 8, and 30).

The non-competitive performances of the Linear Model and Ridge Models in simulating reservoir outflows are also observed in some other reservoir cases, for example, the results of BSR (Fig. 3a), HNR (Fig. 4a), and PIN (Fig. 4k), respectively; In all those mentioned cases, we observe that Linear and Ridge Models are performing poorly as compared to other more complex AI&DM models. This indicates the reservoir release decision is not a simple and linear process with respect to inflow amount and storage volumes. The linear assumption may not be valid when using reservoir inflow, storage to estimate reservoir outflow, and more complex AI&DM models are needed.

When we are analyzing obtained NSE values in the prior Table 3, we observe that the worst NSE value occurred in the reservoir case of NAV. With the time series plots and Taylor diagram of reservoir case NAV (Fig. 4j), we found that the worst NSE value is due to the poor performance associated with the CART model and kNN algorithms. The time series plots in Fig. 4 (j) indicate both the CART model (light green line) and kNN algorithms tend to overestimate the peaks and generate

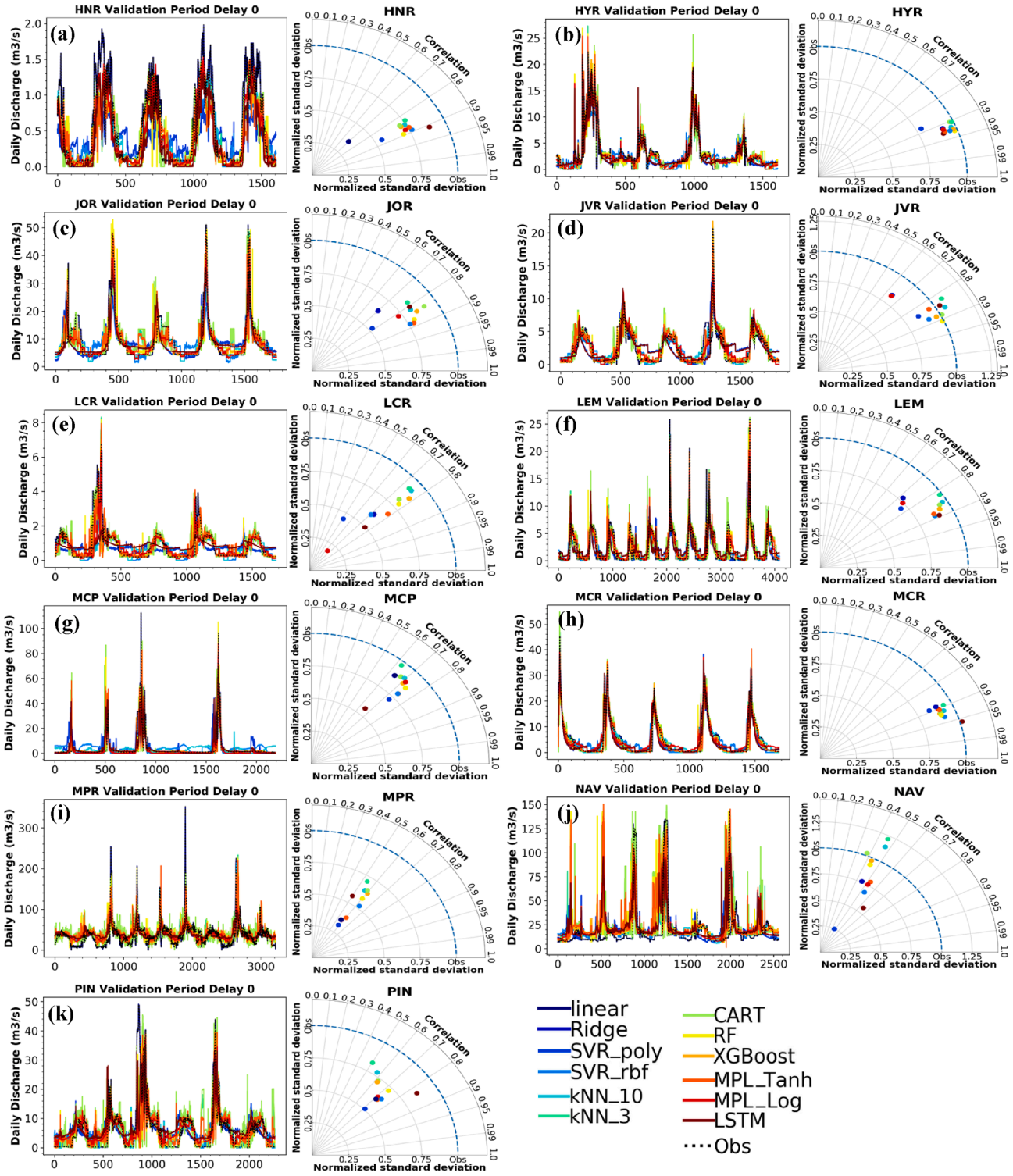

Fig. 4. Simulated and observed daily reservoir discharges and Taylor Diagrams for different AI&DM models - Part II.

significant overestimations across the validation periods. Based on the detailed statistics in Supplementary Material Table 21, the obtained NSE values for kNN_10, kNN_3, and CART models are −0.332, −0.476, and −0.365, respectively. These negative NSE values indicate unsatisfactory and poor model predictive performance of these models. Similar issues are also identified in other cases, including the reservoirs ECH, ECR, FGR in Fig. 3 (g), (h), (i), respectively; the reservoirs MPR and PIN inf Fig. 4 (c) and (k), respectively; and the reservoirs RFR, RID, SCO, STA, STE, and TPR in Fig. 5 (a), (b), (e), (g), (h), (i), respectively. In all the mentioned cases, the simulated daily reservoir releases from the kNN and CART models have relatively low NSE values compared to other AI&DM models.

Furthermore, in the case of FGR (Fig. 3i), we notice the LSTM model (dark red line) fails to predict the low flows of the reservoir releases, and the obtained NSE value for the LSTM model is the lowest (-0.133) as compared to other AI&DM models (Supplementary Material Table 9). Similar cases include the reservoir cases of LCR (Fig. 4e), PIN (Fig. 4k), RFR (Fig. 5a), SCO (Fig. 5e), and STA (Fig. 5g). In all these identified

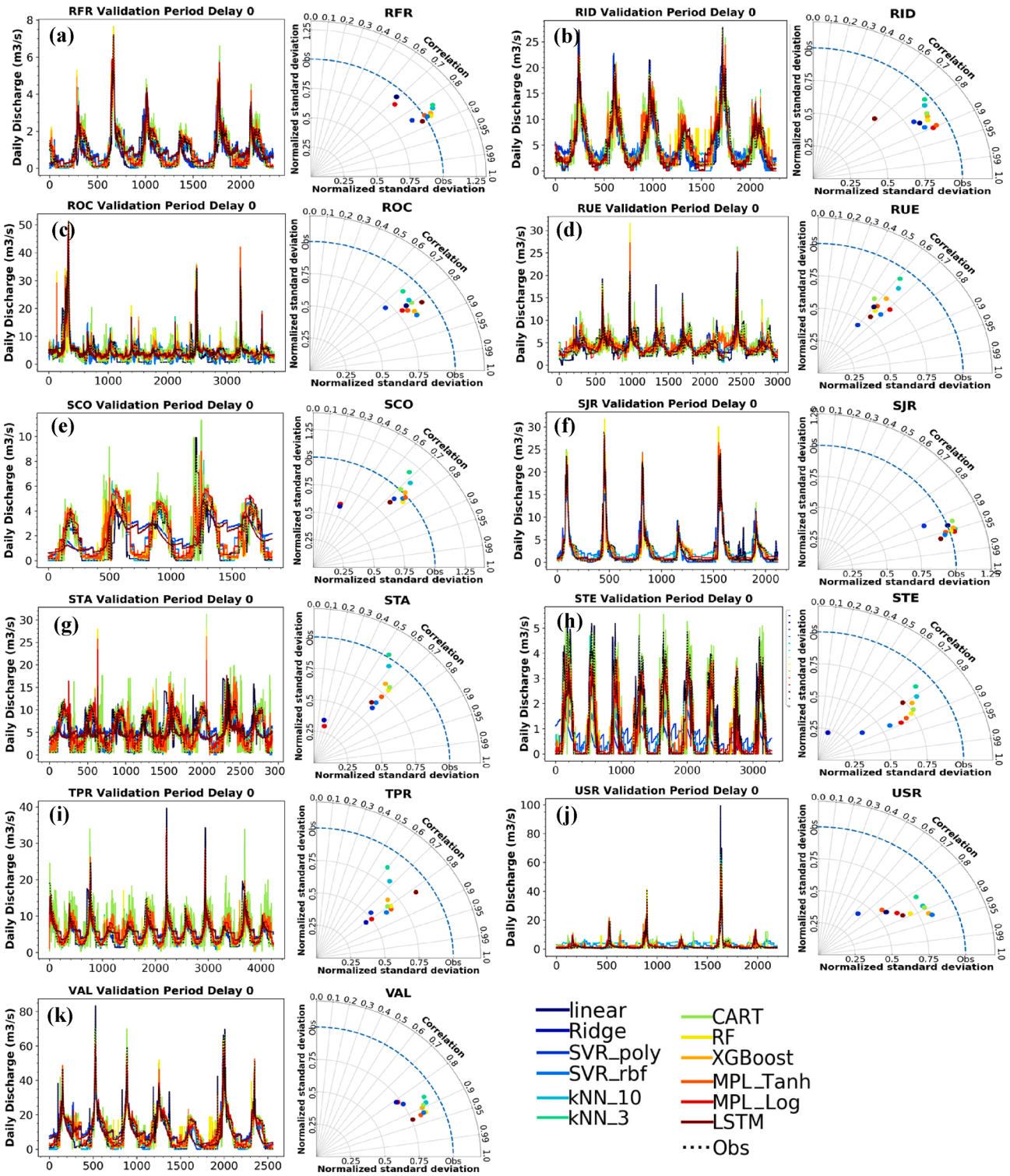

Fig. 5. Simulated and observed daily reservoir discharges and Taylor Diagrams for different AI&DM models - Part III.

cases, though the LSTM model could well capture peak flows, the low flows are consistently overestimated. The obtained NSE values by the LSTM model in these reservoirs are also lower than those obtained by other employed AI&DM models.

With respect to the KGE value, in earlier results, we identify among all employed models, the worst KGE value happens in reservoir USR with a negative KGE value of is −0.242 (Fig. 5j). According to Table 3, another two cases when KGE exhibits negative values are reservoir cases of STE and STA. We noticed that in reservoir STE and STA, both Linear Model and Ridge Model perform poorly. If excluding the linear models in reservoirs STE and STA, other models could consistently derive positive KGE values (Table 3). In reservoir USR (Fig. 5j), the two worst models are SVR models with radial basis and Polynomial kernels, which generate KGE values of −0.242 and 0.081, respectively (Supplementary Material Table 32). In the same case of USR, the simulation results derived from other models are consistently higher than 0.444, which belongs to a satisfactory KGE value. In addition, we also notice that in reservoir USR (Fig. 5j), the daily reservoir outflows are zeros in most of

the time steps, and there are only a few peaks within the validation periods. The only similar cases are DIL (Fig. 3e) and MCP (Fig. 4g), where most of the daily releases are very small, and a few peaks exist across the validation period. In the reservoir case of DIL (Fig. 3e), the two worst-performing models are LSTM and SVR with polynomial kernel function, which produces a KGE value of 0.584 and 0.608, respectively (Supplementary Material Table 5). However, the SVR model shows an evident overestimation in the later parts of the validation period (around data point 1700) (Fig. 3e). At this data point from DIL, many AI&DM models significantly overestimated the observation, including the LSTM model, the kNN_10 model, and the SVR_poly model. Among these problematic models, the SVR_poly model tends to perform the worst when compared to others. We further noticed that in the reservoir case of MCP (Fig. 4g) and according to the obtained KGE value over the validation period, the two worst performing models are LSTM and SVR_poly. The KGE values obtained from the simulated outflows from the LSTM and SVR_poly models over the MCP reservoir are 0.275 and 0.372 (Supplementary Material Table 18), respectively, under simulation scenarios No.1 (S1). In summary, we infer that if the reservoir presents consistent small daily releases (close to zeros) for most of the days, the LSTM and SVR_poly model seems to be not capable of obtaining satisfactory KGE values of the simulated reservoir outflow.

According to the Taylor diagrams presented along with each time series plot in Figs. 3–5, there are three major observations. First, it seems that the model simulation results from different types of AI&DM models are similar and consistently good for reservoir cases of CAU (Fig. 3b) and CRY (Fig. 3c). All the colored dots in the Taylor diagrams are nicely clustered and are consistently close to the observation point. Second, for reservoir cases of PIN (Fig. 4k) and MCR (Fig. 4h), the LSTM model performs significantly better than other AI&DM models, and the LSTM dot (dark red) is the closest to the observation point in the corresponding Taylor diagrams with apparent superiority. However, the LSTM model turns to be the worst performing model when switching to reservoir cases of MCP (Fig. 4g) and RID (Fig. 5b), and their corresponding locations in the Taylor diagram are further away from the observation as compared to that of other models. Third, in reservoir cases of RFR, SCO, and STA (Fig. 5a, e, and g, respectively), the MLP_Tanh model demonstrates relatively poorer performance than that in other reservoir cases. The level of predictive performance of the MLP_Tanh model is similar and low as it shows with some linear models. Note that these above findings are specifically based on the Taylor diagram results, which are mainly drawn from the closeness of model-generated color dots towards the observation. The inference of models' predictive performance may or may not correspond to the findings observed from the tabular statistical results. Nevertheless, the Taylor diagram provides an easy, integrated, and visually direct quantification of the raw model performance in the context of the standard deviation, CORR, and RMSE.

4.3. Model uncertainty and simulation stability

In the following Fig. 6, we present the box plot of the statistics obtained by each AI&DM model over the validation periods of all 33 reservoir cases under Scenario No.1. Note that this figure depicts a different aspect of statistical performance as compared to the statistical summary in the prior Table 3. Specifically, Table 3 indicates the AI&DM models' overall performance of maximum, minimum, and average for each employed reservoir, while Fig. 6 shows the performance of each employed AI&DM model across all 33 reservoir cases, and each box plot was made using 33 data points for each corresponding statistical measurement. According to Fig. 6, most of the AI&DM models could obtain the interquartile range (25–75 percentiles) of [0.6–0.9], [0.3–0.8], and [0.2–0.8], for CORR, NSE, and KGE (Fig. 6a–c), respectively, and [1.5–6.5], [−15 to 20], and [0.5–8.5] for the normalized RMSE, PBIAS and RSR (Fig. 6d-f), respectively. If comparing the statistical performance of all employed models, we notice that (1) the tree-based models (CART, RF, and XGBoost) could achieve a higher CORR and KGE values interquartile range, which indicates a better overall performance than other AI&DM models across all studied reservoirs, and (2) the linear models (Ridge and Linear) generally produce the poorest interquartile ranges as compared to others with respect to the statistics of CORR, NSE, KGE, and RSR. The LSTM model shows as a competitive model as compared to others in most of the reservoir cases. However, we also notice the ranges of interquartile obtained by the LSTM model are relatively larger than other AI&DM models for the statistic metrics of CORR, NSE, KGE, RSR, and PBIAS. This observation means that the LSTM model's performance are less stable than other AI&DM models, though in some cases, it can obtain a superior performance under the input Scenario No.1. Similar findings are also discussed in the prior section about the Taylor diagram. In general, the LSTM model performance indicated in this Fig. 6 and that in the prior Taylor diagrams are in agreement with each other. In addition, the LSTM model also generates the highest PBIAS interquartile range (Fig. 6e), which is not desired. In summary, the LSTM model has less transferability from one reservoir case to another under the baseline input scenario No.1, and it demonstrates low stability when applied in different simulation cases.

The results in Fig. 6 also indicate that (1) evaluating the performance

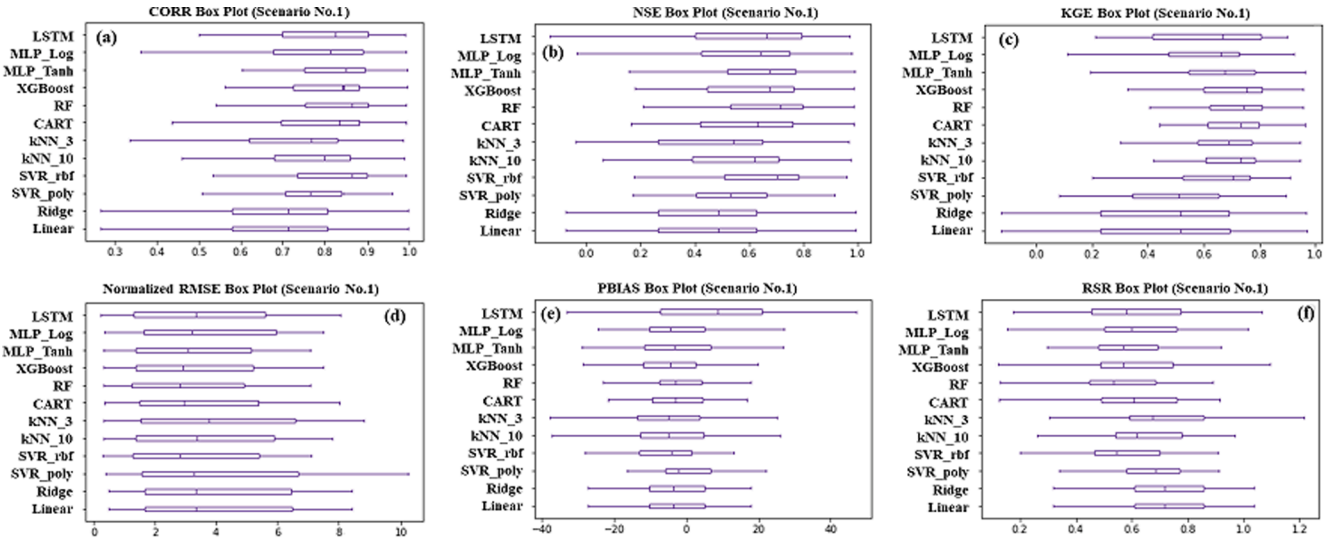

Fig. 6. Box Plots for Each AI&DM Model of All Employed Reservoir Cases Under Scenario No.1.

of AI&DM models needs an examination of a comprehensive set of statistical measurements as the superiority of one particular statistical performance cannot guarantee the same performance in other measurements, and (2) any conclusion about the performance of AI&DM models on single or few study cases would rather be biased, and when changing the study cases the model's performance are likely to vary. For example, we notice that kNN models (kNN_3 and kNN_10) could reach very satisfactory KGE interquartile ranges, and they are significantly better than the SVR_Poly model under Scenario No.1 (Fig. 6c). However, the simulated outflow from all 33 reservoirs by the SVR_Poly model has a lower bias (PBIAS value closer to zero) than all kNN models (Fig. 6e). Similarly, though the Multiple Layer Perceptron models (MLP_Log and MLP_Tanh) show superior performance over linear model sets in CORR, NSE, and KGE statistics (Fig. 6a-c), the overall performance of reducing RMES and PBIAS is similar to that of the linear model sets (Fig. 6d-e). More discussion will be provided in the later section regarding the best-performing models under different simulation scenarios and statistical measurements.

4.4. Statistical results for baseline of Scenario No.2&3 (S2&S3)

Instead of presenting the same statistical analysis similar to prior Table 4 and Figs. 3–5 under Scenario No.1 (S1), in this sub-section, we focus on comparing the results under Scenarios No.2 and 3 (S2 and S3) in terms of improvement or deterioration comparing to the baseline S1. All calculated raw statistical measurements under Scenarios No.2 and 3 are also available in Supplementary Material Tables 1-33, and this section is intended to highlight the performance changes of the employed AI&DM models when switching the input from baseline to more complex scenarios.

In the following Table 4, we summarize the averaged statistical improvement percentages for each reservoir under S2 and S3. Note that the improvement percentages in Table 4 are averaged across all employed AI&DM models for each reservoir. In another set of results (Table 5), we carried out a similar comparison and obtained the averaged statistical improvement percentages for each employed AI&DM model. Differs from the results in Table 4, the improvement percentages shown in Table 5 are averaged across all 33 reservoir cases, and the percentages are organized for each AI&DM model. In other words, Table 4 indicates how well the simulation of each reservoir could be improved by adding delayed inputs regardless of the employed AI&DM methods, while Table 5 reveals how sensitive each AI&DM model will respond to the additional inflow and storage information across different reservoir cases. In both Tables 4 and 5, a positive percentage implies an improvement with respect to the baseline Scenario No.1, and a negative percentage means the statistical performance deteriorates as compared to the baseline S1.

According to the results shown in Tables 4 and 5, there is a consistent observation that when delayed inflow and storage are added to the employed AI&DM models, the statistical performance will improve in general. Specifically, according to the S2 results in Table 4, the CORR, RMSE, NSE, KGE, PBIAS, and RSE values will improve 4.8%, 10.9%, 14.7%, 8.2%, 90.8%, and 10.9%, respectively, when averaging across all reservoir cases. If the AI&DM model inputs are further extended to include the two-step delayed inflow and storage values, the averaged model performance for all studied reservoir could further improve up to 7.6%, 16.5%, 28%, 12.8%, 96.1%, 16.5%, and 26.4%, for CORR, RMSE, NSE, KGE, PBIAS, and RSE (Table 4 last row), respectively. According to Tables 4 and 5, we observe that the additional inflow and storage inputs with delayed time series will consistently improve the outflow

Table 4
Percentage Improvements of All Methods for Each Studied Reservoir (Scenario No.1 as Baseline) Over the Validation Period.

Reservoirs/ Statistics	CORR		RMSE		NSE		KGE		PBIAS		RSE		Ave. for All Statistics	
	Scenario No.		Scenario No.		Scenario No.		Scenario No.		Scenario No.		Scenario No.		Scenario No.	
	2	3	2	3	2	3	2	3	2	3	2	3	2	3
BSR	0.6%	0.6%	9.4%	13.4%	2.1%	1.5%	3.9%	1.7%	36.7%	72.3%	9.4%	13.4%	**8.87%**	**14.72%**
CAU	0.8%	0.9%	9.6%	17.4%	1.6%	1.9%	1.3%	1.7%	89.3%	105.2%	9.6%	17.4%	**16.02%**	**20.63%**
CRY	0.3%	0.4%	35.2%	29.8%	0.8%	1.0%	1.7%	2.4%	57.3%	73.2%	35.2%	29.8%	**18.63%**	**19.50%**
DCR	1.1%	5.0%	10.6%	10.2%	2.3%	10.5%	1.4%	8.0%	62.6%	54.9%	10.6%	10.2%	**12.64%**	**14.12%**
DIL	−0.7%	−1.3%	7.2%	11.2%	−1.9%	−0.9%	−3.9%	−7.3%	85.4%	63.6%	7.2%	11.2%	**13.33%**	**10.92%**
ECH	3.5%	6.2%	17.0%	24.5%	7.5%	14.7%	5.3%	11.5%	62.8%	92.4%	17.0%	24.5%	**16.15%**	**24.85%**
EBR	4.9%	11.1%	4.7%	7.4%	196.5%	401.4%	13.1%	32.2%	89.7%	80.1%	4.7%	7.4%	**44.80%**	**77.09%**
ECR	1.3%	3.0%	4.4%	5.1%	0.3%	9.0%	0.2%	4.3%	131.6%	134.2%	4.4%	5.1%	**20.32%**	**22.96%**
FGR	8.1%	10.7%	11.4%	12.7%	16.9%	43.1%	42.9%	45.5%	44.8%	40.4%	11.4%	12.7%	**19.36%**	**23.60%**
FON	4.0%	7.8%	16.7%	27.7%	4.8%	17.9%	1.7%	11.8%	294.7%	50.3%	16.7%	27.7%	**48.37%**	**20.48%**
GMR	5.0%	8.6%	13.0%	25.4%	6.9%	18.3%	1.7%	9.5%	150.6%	145.3%	13.0%	25.4%	**27.19%**	**33.24%**
HNR	6.9%	9.0%	16.5%	23.2%	28.7%	39.3%	25.5%	34.6%	56.9%	27.2%	16.5%	23.2%	**21.57%**	**22.35%**
HYR	3.3%	4.0%	20.2%	31.4%	6.6%	7.9%	2.4%	2.6%	41.1%	46.1%	20.2%	31.4%	**13.38%**	**17.64%**
JOR	3.0%	1.5%	6.1%	6.7%	6.9%	3.7%	5.0%	4.5%	118.1%	149.2%	6.1%	6.7%	**20.73%**	**24.61%**
JVR	4.3%	2.4%	6.0%	4.4%	10.5%	6.4%	4.1%	1.6%	28.3%	25.9%	6.0%	4.4%	**8.47%**	**6.45%**
LCR	−0.2%	4.5%	9.3%	12.6%	7.5%	18.1%	22.2%	31.7%	95.6%	88.1%	9.3%	12.6%	**20.54%**	**23.94%**
LEM	11.4%	12.6%	30.9%	34.1%	26.7%	29.7%	14.4%	15.9%	46.5%	40.6%	30.9%	34.1%	**22.98%**	**23.84%**
MCP	3.2%	3.3%	4.3%	4.7%	8.3%	8.6%	6.1%	6.8%	832.9%	976.0%	4.3%	4.7%	**122.73%**	**143.44%**
MCR	3.2%	5.3%	17.1%	36.5%	6.5%	10.2%	2.2%	2.3%	44.4%	207.0%	17.1%	36.5%	**12.91%**	**42.54%**
MPR	1.8%	1.4%	1.5%	1.8%	6.0%	6.2%	1.4%	7.3%	31.7%	34.8%	1.5%	1.8%	**6.28%**	**7.60%**
NAV	11.0%	14.4%	6.9%	9.4%	31.0%	30.8%	24.4%	32.6%	29.7%	30.5%	6.9%	9.4%	**15.69%**	**18.17%**
PIN	10.5%	19.4%	16.0%	26.3%	25.4%	50.3%	15.3%	27.8%	79.2%	94.0%	16.0%	26.3%	**23.19%**	**34.87%**
RFR	0.7%	2.4%	3.0%	5.3%	1.5%	4.3%	−0.2%	−0.3%	17.7%	21.2%	3.0%	5.3%	**3.67%**	**5.47%**
RID	4.4%	5.2%	8.0%	17.5%	8.6%	13.5%	2.2%	7.4%	41.1%	53.5%	8.0%	17.5%	**10.33%**	**16.35%**
ROC	3.3%	4.4%	5.8%	10.5%	7.3%	9.5%	4.4%	3.9%	20.8%	21.3%	5.8%	10.5%	**6.76%**	**8.59%**
RUE	6.6%	14.6%	8.5%	13.4%	33.0%	47.3%	14.1%	23.5%	46.5%	61.7%	8.5%	13.4%	**16.77%**	**24.86%**
SCO	13.1%	19.9%	5.4%	13.1%	−141%	−195%	32.1%	48.0%	19.8%	20.2%	5.4%	13.1%	**−9.33%**	**−11.55%**
SJR	0.9%	1.5%	9.7%	13.9%	1.9%	3.1%	0.2%	0.6%	63.4%	32.6%	9.7%	13.9%	**12.25%**	**9.37%**
STA	28.6%	43.9%	6.3%	15.8%	110.7%	202.7%	7.6%	9.5%	86.9%	122.5%	6.3%	15.8%	**35.19%**	**58.62%**
STE	1.3%	0.7%	3.5%	2.7%	30.4%	48.3%	4.1%	−1.2%	7.7%	10.0%	3.5%	2.7%	**7.18%**	**9.02%**
TPR	6.4%	13.9%	9.3%	30.0%	17.3%	32.6%	12.6%	25.9%	108.4%	138.1%	9.3%	30.0%	**23.32%**	**38.65%**
USR	3.0%	4.8%	14.3%	17.2%	6.4%	9.9%	−4.6%	1.4%	44.1%	22.8%	14.3%	17.2%	**11.08%**	**10.49%**
VAL	3.1%	7.6%	13.3%	30.7%	6.9%	17.4%	5.7%	14.6%	30.3%	36.0%	13.3%	30.7%	**10.39%**	**19.56%**
Ave. of All Cases	**4.8%**	**7.6%**	**10.9%**	**16.5%**	**14.7%**	**28.0%**	**8.2%**	**12.8%**	**90.8%**	**96.1%**	**10.9%**	**16.5%**	**20.05%**	**25.36%**

Table 5
Percentage Improvements of All Methods for Each Studied Reservoir (Scenario No.1 as Baseline) Over the Validation Period.

Models/Statistics	CORR		RMSE		NSE		KGE		PBIAS		RSE		Ave. for All Statistics	
	Scenario No.		Scenario No.		Scenario No.		Scenario No.		Scenario No.		Scenario No.		Scenario No.	
	2	3	2	3	2	3	2	3	2	3	2	3	2	3
linear	13.3%	19.3%	24.0%	30.4%	87.1%	148.5%	27.4%	36.6%	35.2%	48.8%	24.0%	30.4%	**30.15%**	**44.84%**
Ridge	9.8%	16.8%	15.3%	24.0%	63.2%	124.9%	13.4%	24.1%	24.2%	40.6%	15.3%	24.0%	**20.18%**	**36.33%**
SVR_poly	3.8%	5.0%	6.0%	7.8%	9.2%	12.1%	12.6%	17.3%	39.9%	54.0%	6.0%	7.8%	**11.08%**	**14.83%**
SVR_rbf	2.9%	5.7%	8.2%	14.8%	5.8%	11.3%	−0.3%	0.5%	79.0%	187.9%	8.2%	14.8%	**14.83%**	**33.57%**
kNN_10	2.3%	3.5%	5.5%	7.7%	7.1%	10.3%	3.3%	5.2%	247.2%	230.2%	5.5%	7.7%	**38.70%**	**37.82%**
kNN_3	3.1%	5.3%	6.6%	9.9%	18.9%	33.1%	4.5%	7.2%	25.6%	40.0%	6.6%	9.9%	**9.33%**	**15.06%**
CART	1.5%	1.9%	4.8%	6.8%	2.5%	3.5%	2.0%	3.1%	51.8%	55.3%	4.8%	6.8%	**9.61%**	**11.04%**
RF	1.2%	2.2%	3.8%	5.9%	1.9%	1.8%	2.0%	3.2%	29.7%	43.6%	3.8%	5.9%	**6.04%**	**8.95%**
XGBoost	5.8%	8.6%	13.7%	19.6%	8.7%	12.0%	6.7%	10.2%	94.7%	184.8%	13.7%	19.6%	**20.46%**	**36.38%**
MPL_Tanh	7.6%	12.8%	22.3%	41.1%	26.1%	43.2%	12.6%	26.1%	154.8%	99.6%	22.3%	41.1%	**35.12%**	**37.70%**
MPL_Log	11.0%	17.4%	18.7%	30.8%	−35%	−41%	21.3%	33.1%	131.8%	123.5%	18.7%	30.8%	**23.69%**	**27.79%**
LSTM	0.7%	−2.4%	15.8%	14.7%	−6.1%	−11.8%	−0.3%	−5.6%	235.6%	113.9%	15.8%	14.7%	**37.35%**	**17.65%**
Ave. for All Models	**5.8%**	**8.9%**	**10.5%**	**16.6%**	**12.7%**	**22.8%**	**8.8%**	**13.5%**	**91.1%**	**104.8%**	**10.5%**	**16.6%**	**19.91%**	**26.18%**

simulation accuracy, except for the reservoir SCO. According to the averaged improvement percentages shown in the last two columns of Table 4, we can conclude that in most of the reservoir cases, the statistical performance under S3 are better than that under both S1 and S2, though the magnitudes of improvement could slightly vary from one case to another.

According to the results shown in Table 5, when averaging the statistical improvement percentages across all studied reservoirs, most of the AI&DM models show similar improvement patterns as delayed reservoir inflow and storage information being added as additional model inputs. However, we notice that the predictive performance of the LSTM model may either deteriorate or improve when adding delayed information (Table 5), which is subject to the study case and employed statistical measurement. In other words, there is no consistent improvement or deterioration of the LSTM model, as we observed from other AI&DM models when switching from S1 to S2 and to S3. Specifically, when using the LSTM model, the RMSE and PBIAS values have been improved. However, the CORR, NSE, and KGE values show some levels of deterioration. In comparison, the Linear model, Ridge model, kNN models, and Tree-based models demonstrate consistent improvements when adding delayed information. We suspect this is because the LSTM model, by design, could take the current inflow and storage values, as well as its prior values as model inputs to estimate the outflow at the current time step. Therefore, manually setting delayed inflow and storage volumes as additional LSTM inputs did not essentially promote the model accuracy. Other employed AI&DM models do not carry the same feature as the LSTM model, and the delayed information is associated with complementary predictability than the original model inputs in the baseline S1. Furthermore, according to the results in Table 5, the CORR, RMSE, NSE, KGE, PBIAS, and RSE values are improved by 5.8%, 10.5%, 12.7%, 8.8%, 91.1%, and 10.5%, respectively, for all models under Scenario No.2. The same statistical measurement values are further improved by 8.9%, 16.6%, 22.8%, 13.5%, 104.8%, and 16.6%, respectively, when switching the model simulation to S3. According to the last two columns in Table 5, similar patterns are observed when comparing the percentage improvements for all statistical measures under S2 and S3.

Beyond the statistical comparison among different Scenarios, in the following Table 6, we further summarize the best and the second-best performing models for each statistic over the validation periods. The best-performing models are identified by comparing how many reservoir cases that a particular model can outperform other models on each individual statistical measurement. For example, according to Table 6, under the baseline scenario No.1, both RF and LSTM model could generate the best CORR values over the validation period for 10 out of the 33 employed reservoir cases. Therefore, RF and LSTM models are identified as the best performing models under S1 and for the CORR category. The second-best model under the same statistic measure (e.g., S1 and CORR) is the SVR model with radial basis kernel function (SVR_rbf), which outperforms 4 out of 33 studied cases. Using the same logic, we summarize the best and the second-best performing models, as well as the number of reservoirs that the corresponding model prevailed other employed AI&DM models in the following Table 6.

According to the results in Table 6, both RF and LSTM models are identified as the best- and second best-performing models under Scenario No.1(S1), respectively. They prevailed the highest and second-highest numbers of reservoirs than other employed models. This result indicates that both RF and LSTM models are more reliable and transferable than other models when simulating reservoir outflow using the current time step inflow and storage volume as model inputs. The performance of the RF and LSTM model are consistent across different statistical measurements, except in two cases when the SVR_rfb and kNN_3 model outperformed 4 and 6 reservoirs out of 33 studied cases and ranked as the second best-performing model under S1, respectively. However, when comparing the number of prevailing cases under S2 and S3, both the RF and LSTM model will no longer perform as the best and

Table 6
The Best Two Performing Models Under Each Scenario and The Number of Reservoir Experiments the Best Model Outperforms Others.

Scenario	BMP and Performance/Statistics	CORR	RMSE	NSE	KGE	PBIAS	RSR
No.1	The Best Performing Model	**RF/LSTM**	**RF**	**RF**	**RF**	**LSTM**	**RF**
	Number of Prevailing Cases	10 of 33	11 of 33	11 of 33	9 of 33	9 of 33	11 of 33
	The Second Best Performing Model	**SVR_rbf**	**LSTM**	**LSTM**	**LSTM**	**kNN_3**	**LSTM**
	Number of Prevailing Cases	4 of 33	9 of 33	9 of 33	5 of 33	6 of 33	10 of 33
No.2	The Best Performing Model	**MLP_Tanh**	**MLP _Tanh**	**MLP _Tanh**	**XGBoost**	**SVR_rbf**	**MLP _Tanh**
	Number of Prevailing Cases	10 of 33	11 of 33	11 of 33	13 of 33	7 of 33	11 of 33
	The Second Best Performing Model	**XGBoost**	**XGBoost**	**XGBoost**	**MPL_Tanh**	**LSTM**	**XGBoost**
	Number of Prevailing Cases	7 of 33	7 of 33	7 of 33	5 of 33	5 of 33	7 of 33
No.3	The Best Performing Model	**MLP _Tanh**	**MLP _Tanh**	**MLP_Tanh**	**XGBoost**	**MLP _Tanh/LSTM**	**MLP _Tanh**
	Number of Prevailing Cases	12 of 33	12 of 33	12 of 33	16 of 33	7 of 33	12 of 33
	The Second Best Performing Model	**XGBoost**	**XGBoost**	**XGBoost**	**MPL_Tanh**	**SVR_rbf**	**XGBoost**
	Number of Prevailing Cases	8 of 33	9 of 33	9 of 33	9 of 33	6 of 33	9 of 33

second-best models. Instead, the MLP model with Hyperbolic Tangent Activation function (e.g., MLP_Tanh) becomes the best performing model, and the XGBoost Tree model becomes the second-best performing model across different statistical measurements. The MLP_Tanh model shows slightly better performance over the XGBoost Tree model for statistics CORR, NSE, KGE, and RSR, while the XGBoost Tree algorithm outperforms the MLP model in KGE measurement under both S2 and S3. This result indicates that the performance of both MLP and XGBoost Tree models are stable and consistent across different statistical measurements when additional delayed information to simulate reservoir outflows.

In addition, the total number of reservoirs that the MLP_Tanh model prevails other AI&DM models slightly increased from 10, 11, 11, and 11 for statistics CORR, NSE, KGE, and RSR under S2, to 12, 12, 12, and 12 under the S3, respectively. Similarly, the total number of reservoirs that the XGBoost Tree model outperforms others also increases from 7, 7, 7, and 7 for the same statistics under S2 to the values of 8, 9, 9, and 9 under S3. This result implies that the reliability of the MLP_Tanh model and XGBoost Tree model is slightly improved with the increases of input complexity. Combined with the findings from earlier Tables 4 and 5, we further infer that these two models are capable of handling additional delayed information and can derive consistently good statistical performance with continuous improvements over the baseline input set under Scenario No.1.

The results in Table 6 also show that the SVR model can generate the best PBIAS values that are closer to zero than other AI&DM models for 7 out of 33 studied cases under S2, and prevails other models for 7 out of 33 reservoirs under S3, respectively. For the RSR measurement, both of the kNN and SVR models sometimes could outperform other models and succeed in the largest number of reservoirs under different input scenarios of S1-3. However, none of them could obtain better statistics than the MPL_Tanh, the LSTM, the RF, and the XGBoost Tree model with respect to all other commonly used statistical measurements (CORR, RMSE, NSE, and KGE).

5. Discussion

For most of the hydrological time series, it is likely that the water flow (either natural flow or controlled flow, e.g., the reservoir outflow) is continuous, and the time series inevitably have a certain level of autocorrelation with respect to previous time steps. One of the advantages of AI&DM models is the flexibility of handling additional inputs for data classification and prediction.

Echoing the presented simulation results in Section 4, we can reasonably expect that when adding delayed inflows and storage volumes in previous days, the AI&DM model could better capture the variability of reservoir outflows at the current time step. Similar correlation studies were available in the literature that the reservoir outflow decisions are also related to previous inflow and storage conditions up to two days in a retrospective manner (Hejazi et al., 2008, Zhao et al., 2012). Specifically, our designed Scenarios No. 2 and 3 experiments (Fig. 2) will answer the question of whether the employed AI&DM models with additional and delayed inflow and storage could achieve better or worse performance over the baseline Scenario No.1. The experiments with additional delayed AI&DM model inputs (S2 and S3) can help us further investigate the flexible uses of different AI&DM models in simulating the reservoir outflow decisions.

The results in Table 6 indicate that the LSTM and RF model could effectively use limited information from the inflow and storage at the current time step to simulate reservoir outflows by achieving the best statistical measurement values over a larger number of study cases than other AI&DM models. However, when manually adding delayed input information, the performance of MLP models and XGBoost tree-based models could be significantly improved and even outperform the LSTM and RF models in the baseline scenario. In other words, the MLP and XGBoost Tree models are less sensitive to the training information as compared to other AI&DM models as long as the proper information is used to train the model. The MLP_Tanh and XGBoost Tree models' performance stay competitive and stable when switching input scenarios from No.2 to No.3. This finding indicates that the selection of AI&DM models should rely on what information being used as predictors during the model training process and what statistical measurements are employed to evaluate the model performance after training. In general, there is no single model that could consistently outperform other AI&DM models with respect to all possible statistical measurements and across all studied reservoir cases.

The large-scale comparison studies in this paper are intended to explore the pros and cons of different AI&DM models in the field of hydrology and water resources. Based on our experiment results and findings, AI&DM model practitioners shall investigate as many similar cases as possible, and include multiple evaluation statistical measurements before putting them into real-world application. Diligent quality control and evaluation study of different AI&DM models will help operators comprehensively understand which model works well under what evaluation criteria. We believe that it is also highly subjective to simply claim the superiority of one particular AI&DM model based on a limited number of statistical measurements, as well as draw a conclusion from only a few study cases. Our current experiments also indicate that, though the nature of the problem could be similar, e.g., estimating reservoir outflows using inflow and storage, the performance of the AI&DM models will likely be significantly different when the training data changes, parameter changes, and the core regression techniques changes. It is also questionable to infer the performance of the AI&DM model based on one or a limited number of case studies to a large number of other untested cases without any evaluations. Most AI&DM models' predictive performance has uncertainty, and the obtained statistical measurements may also have large variations (Fig. 6) among different models. The prediction uncertainty and variability are inevitable because the AI&DM model purely relies on the training data, regardless of the physical dynamics and mass-continuity (Kim et al., 2021). There is no guarantee that the success of one or few case studies can become sufficient evidence that the same model functioning and behavior will occur in other study cases.

The authors would like to make the point clear that, though AI&DM models are popular and powerful tools in many current studies of hydrological simulation and reservoir operation, the pros and cons of each model still remain not fully understood. As evidenced in our simulation results (Figs. 3–5), not always one type of AI&DM model will consistently perform well in all study cases, and therefore, large-scale experiments, such as (1) the one presented in this study that intercompares various models on a large number of case studies, or (2) the investigation study to directly compare AI&DM models with physical rule-based or process-based modeling schemes, are critically needed before the practical uses of AI&DM models in any field. Due to the limitation of accessing the rule-curves of the 33 reservoirs, this study only serves as the former type of comparative study. A further comparison of the simulated AI&DM models against rule-curves, as well as the examination of how AI&DM simulated reservoir release decisions could reasonably meet all types of hard engineering constraints, soft operation limits, downstream supply, and environmental water and hydroelectric power demands, and regional ordinances are out of great importance to promote the practical applications of AI&DM models.

Furthermore, we suspect that the AI&DM model performance are also related to the elevations of the dam, maximum capacity of the reservoir forebay storage, as well as the primary functionalities of each reservoir (Table 2). For the influence of elevation, if taking a detailed analysis on some of the reservoirs with the highest elevations, for example, TPR (Fig. 5i), DIL (Fig. 3e), SJR (Fig. 5f) with elevations of 2847 m, 2751 m, and 2725 m, respectively, we notice most AI&DM models could produce consistent results with the normalized standard deviation above and higher than 0.5 (the circular lines in all Taylor diagrams). In contrast, when it comes to some of the reservoirs with the

lowest elevation, for instance, EBR (Fig. 3f), PIN (Fig. 4k) with elevations of 1323 m and 1427 m, respectively, we observe that results produced by different AI&DM models are having a larger discrepancy among each other. This is also evidenced by the normalized standard deviation shown in EBR (Fig. 3f) and PIN (Fig. 4k) that the color dots have a larger spread in the Taylor Diagram than other reservoirs. In other words, we infer that the reservoirs at lower elevations are more complex to manage due to the influences of water routing in upstream river basins and the possible discharges from reservoirs in higher elevations. Note that our finding is not contradictory with the conclusion about the impacts of elevation on reservoir discharge simulation from a prior study on reservoirs in California (Yang et al., 2016). Specifically, the elevations of reservoirs in Yang et al. (2016) are significantly lower than ours, and their studied reservoirs are more closer to water consumption areas, such as irrigation and residential districts. The reservoirs in our study region are also believed to be dominated by seasonal snowmelt, mountainous hydrology, and natural streamflow, instead of the direct reservoir refills due to the atmospheric river events that bring heavy precipitation over the Northern California region.

Regarding the influences of storage capacity, we infer that the larger the maximum reservoir storage, the more challenging the AI&DM models could capture the human's discharge decisions, and vice versa. This speculation is drawn based on the statistical performance of the following four reservoirs: the reservoir cases of FGR and EBR with maximum storage capacities of 4673.532 million m^3, and 2547.149 million m^3 (Table 2), respectively, which are the two largest reservoirs in this study. In contrast, the two smallest reservoirs are also analyzed, i.e., HNR and CAU, with storage capacities of 6.685 million m^3 and 9.707 million m^3 (Table 2), respectively. According to Table 3, the averaged NSE values derived by all AI&DM models for reservoir case FGR, EBR, HNR, and CAU are 0.235, 0.496, 0.690, and 0.918, respectively. The former two (i.e., the largest two reservoirs) are significantly lower than that from the latter two (i.e., the smallest two reservoirs). The same observation occurs in the KGE evaluation, in which the averaged KGE values across all AI&DM models for reservoir case FGR, EBR, HNR, and CAU are 0.286, 0.583, 0.607, and 0.877, respectively. In addition, we further noticed in the two largest reservoirs (FGR and EBR), at least one AI&DM model performed poorly, which is shown in the "Min" sub row under each reservoir row in Table 3. For NSE, the lowest values for FGR and EBR are −0.133 and 0.006, respectively. The corresponding worst performing model for the two smallest reservoirs (HNR and CAU) can produce an NSE value of 0.256 and 0.885, respectively. A similar pattern is also observed in other statistical measures summarized in Table 3. With this understanding, it is likely that larger reservoirs are associated with more complex constraints, natural environmental variabilities, and operating criteria, which conjunctively place a higher challenge for the applications of AI&DM models. On the contrary, smaller reservoirs tend to be easy to be managed and are more flexible in adjusting to variations of reservoir inflows and changing reservoir storage volumes. Therefore, different AI&DM models have stronger advantages in capturing the human's decision-making process, and effectively simulate the reservoir outflows than the applications on reservoirs with larger maximum capacities.

In addition, reservoir functionality and its primary purpose also play an important role in defining the complexity of the decision-making process. Both of the simulated and observed releases show a certain pattern in our studied cases. From the obtained results (Figs. 3–5), we observe that most of the flood control reservoirs (Table 2) have a distinct seasonal variation in the controlled reservoir outflows. Examples include the reservoir cases of FON (Fig. 3j), FGR (Fig. 3i), GMR (Fig. 3k), VAL (Fig. 5k), etc. Among these cases, there is a clear periodic cycle repeating itself from one year to another throughout the validation periods. Most of the employed AI&DM models are able to well capture these changes and identify that this is due to the operation to empty reservoir storage for the required flood control rooms during fall/winter time before the spring snowmelt. For reservoirs with the functionality of hydropower generation, the water diverted to the powerhouse is also included in the total outflow discharge simulation and observation. However, we notice in three of the reservoirs with hydropower functionalities, i.e., CRY (Fig. 3c), FGR (Fig. 3i), MPR (Fig. 4j), the observed and simulated reservoir outflows show many non-smooth variations at low flow regimes. This may be due to the corresponding hydropower generation during peak hours or intermittent hydroelectric supplies to meet the energy demands or the provision of spinning reserves in the power grid. All of these are the benefits of hydroelectric power and reservoir systems, which could be flexibly turned on and off quickly and effectively to stabilize the power grid variations (Ding et al., 2021).

One drawback of our current study is that the total number of reservoirs employed is still limited. Across the U.S. and worldwide, there is a variety of reservoirs with different functionalities, length of data records, different climate and weather conditions, physical settings of how water being controlled and releases, and the sources of reservoir inflows (e.g., either generated from snowmelt or direct runoff from upper watersheds). Based on specific features of the reservoirs, the experiment setting could be very different when applying the AI&DM models when designing model inputs, such as to include snowmelt, rainfall, lake evaporation, water losses, upstream water level, or reservoir downstream demand information, etc. The work presented here may only cover few subsets of reservoir features, and there is no possible way to across validate the model performance over an unlimited number of other reservoirs worldwide. Nevertheless, the way of how we apply the AI&DM models and test out the model sensitivity and variabilities are universally adaptable to other study cases. The employed experiment setting is based on the common knowledge that the reservoir outflow decision is governed by the rule curves, and is determined by the water conditions in reservoir forebay, afterbay, and storage, etc. Given that all existing rule-based simulation schemes are more or less built upon this key philosophy, we suggest that the uses of AI&DM models to assist reservoir operation shall not discard these underlying physics. With the help of AI&DM models, the investigation of whether additional decision factors, such as local hydrology, global climate interactions, anthropogenic influences, policy, and social-economic variables, are made possible. We highly believe that the AI&DM models are and will continue to be a promising tool in advancing the research frontier of the cross-disciplinary studies between computer and environmental sciences. The usefulness of AI&DM tools is not only limited to the topic of reservoir operation, but also beneficial to a broader aspect of hydrology, meteorology, climate, and integrated water resources management.

Another limitation in this current study is the lack of a full spectrum of sensitivity analysis on the model's structural parameters and the investigation of available prior/post-AI enhancement techniques. On one hand, different AI&DM models are associated with various model parameters, such as learning rate, other activation functions, number of hidden layers and hidden nodes for MLP models; maximum tree-depth, maximum for tree-based models; other kernel functions for SVM models; other numbers of neighbors in the kNN models, etc. The current study only selects a limited number of default settings when implementing these AI&DM models. Before conducting the outflow simulations for the 33 reservoir cases in this study, an initial test has been conducted to preselect the essential model parameters in order to control the level of overfitting for the employed AI&DM models. However, in the presented simulation results, we still observe some overfitting or underfitting phenomenon in some of the reservoir cases. A full spectrum of model parameter sensitivity tests is highly encouraged before applying a particular AI&DM model in practice, which is another grant challenge the operator faced when applying AI&DM models.

On the other hand, the employed AI&DM models in this study are rather the original and standard applications of the popular models in computer science, and we are not able to fully examine all possible AI&DM models with various customized further developments, training algorithms, and prior-/post-procedures. The number of different variations and successors of popular AI&DM algorithms is countless, and each

version may have its own strengths in a specific application. For example, fuzzy inference system (Adnan et al., 2019) could be jointly used with MLP models to adaptively improve the predictive performance; wavelet theory could also be embedded in AI&DM models to refine more representative predictors (Esmaeilzadeh et al., 2017); adding heuristic optimization algorithms could increase the model training accuracy but with a sacrifice of computational speed (Shamshirband et al., 2020, Yang et al., 2017a, Yuan et al., 2018); and hybrid modeling framework or model ensemble techniques could further reduce the prediction uncertainty of different AI&DM models (Tao et al., 2018, Yang et al., 2020b). We suggest future AI&DM model enhancements and applications in reservoir operation, hydrology, and water resources management shall at least include 3–5 standard and original machine learning algorithms with different core regression techniques for the purpose of comprehensive comparison.

Last but not least, the performed large-scale comparison experiments in this study indicate a good potential of different AI&DM models in accurately simulating the controlled reservoir outflows using flexible input designs, which are evidenced by the simulated 33 reservoirs outflows and their corresponding statistical performance (Figs. 3–5, Tables 3–5, and Supplementary Material Tables 1-33). The flexibilities of AI&DM models in taking various human-designed input features and identifying the intrinsic relationship between decision and target variables, remain as the main advantages of AI&DM models. Many implicit, hard-to-find, and internal causal relationships within the natural systems could be discovered and quantified through the proper uses of suitable AI&DM models. Comparison studies, especially on applying AI&DM models with hydrological and environmental data, shall include as many mature statistical measurements as possible to thoroughly examine different aspects of time series, i.e., biases, correlation, extremes, flow regimes, seasonality, trends, etc.

6. Conclusion

In this study, a total number of 12 AI&DM models with different parameterizations are employed to simulate the daily reservoir outflows of 33 reservoirs over the Upper Colorado Region in the United States. We designed three model input scenarios for model training and evaluation using the reservoir inflow, storage, and their corresponding delayed time series as inputs. A total of six commonly accepted statistical metrics were used to quantitatively measure the performance of different AI&DM models under three simulation scenarios. Overall, the simulation results are satisfactory, and the employed AI&DM models could achieve high statistical performance in most of the study cases over the validation periods. However, the performance of different AI&DM models may significantly vary based on the studied reservoir cases, the employed regression techniques, the statistical measurements evaluated, as well as the characteristics of the reservoir, i.e., elevation, size, and the primary functionality. A number of study cases and scenarios were further examined, whereas the simulation generated by different AI&DM models are with relatively low statistical performance. Based on our experiments, the following specific conclusions are drawn:

(1) Different AI&DM models may have individual strengths and weaknesses in simulating the reservoir outflows and assisting reservoir modeling. No single model could consistently outperform others in the 33 reservoir cases compared in this study. The model's performance is likely to vary by the modeling schemes, by the ways of training data structure, as well as by the statistical measurement used. It is suggested to include multiple statistical measurements and as many AI&DM models as possible to comprehensively understand the pros and cons of different AI&DM models when applying AI&DM models to reservoir simulation.
(2) The RF and LSTM models are found to be the best and second-best performing model when using the inflow, and storage at the current time step to simulate the reservoir outflows under the baseline Scenario No.1. However, the MLP model with hyperbolic tangent activation function and the XGBoost Tree algorithm appears to be the most reliable and stable models when additional delayed inflow and storage are further included in the model training process (Scenarios No.2 and 3). We infer that the MLP model and XGBoost Tree model are more capable of handling large and complex training data than other models, and the model's predictive performance are found to be more stable than other employed AI&DM models in the employed 33 reservoir cases.
(3) Reservoir elevation, maximum capacity, primary functionality, local hydrology could conjunctively affect the effectiveness of AI&DM models. We found that the reservoirs located at a lower elevation are more challenging to simulate, and the model daily outflow simulations tend to disagree with each other when different AI&DM models are used. This finding is possibly due to the high variabilities in water routing in upstream river basins, and the influences of discharge from an upper source reservoir/lake. In addition, we also found out that larger reservoirs are harder for the employed AI&DM models to capture the patterns of human controlled release decision, and vice versa. In our limited study, we observe that the statistical performance of AI&DM results over the two largest reservoirs are significantly poorer than that of the two smallest reservoirs. The functionality of reservoir and its primary purpose could also affect the daily and seasonal variations. Periodic patterns are observed in both observed and simulated reservoir outflow time series. The applications of AI&DM models should comprehensively consider these factors in practical uses.
(4) The delayed information reservoir operation time series could significantly increase the model performance in general. According to the comparison between S2/S3 versus the baseline S1, we identify that LSTM and Random Forest model are less sensitive to the manually delayed reservoir storage and inflow time series than many other models, but the LSTM and Random Forest model may not outperform other popular and standard AI&DM models under the scenarios that additional decision variables are added as model inputs. In contrast, properly configured MLP (i.e., ANN) and XGBoost models are more likely to be benefited from adding additional model inputs and produce better simulation results. Based on that, we suggest practitioners who use AI&DM models to simulate reservoir operation shall test out possible ancillary information and additional model inputs that are closely related to the reservoir decision-making process and select the most suitable AI&DM models in corresponding to the training data and study cases.
(5) From this study, we also conclude that the advantage of AI&DM models lies in their flexibility in incorporating different types of input data and identifying the implicit relationship between features and target variables. The traditional rule-based modeling scheme might be limited in this regard as the process-based governing equation can only take fixed inputs. It can be inferred that with different reservoir and problem settings, practitioners of AI&DM models can identify the most suitable combination of input data structure and model parameterizations to obtain the best possible model outcomes, but this process may require a large number of trial-and-error experiments. There are currently not commonly accepted "golden" rules on how to design the implementations of different AI&DM models in assist of reservoir operation. At this stage of research and scientific investigation, large-scale comparison and model evaluation studies are still needed in order to fully understand the pros and cons of different AI&DM models and to identify the best way of using AI&DM models to help manage the surface water resources, understand the local hydrology and water cycle, as well as to

support the decision-making of water infrastructures, such as dams and reservoirs.

(6) Future work may include (1) the direct comparison between AI&DM models with rule-based reservoir simulation and examination of the predicted AI&DM time series will meet various types of engineering hard and soft constraints, demands, and operation goals; (2) the development of hybrid modeling schemes that combine the physical rule-based model with AI&DM model to enhance simulation accuracy and our transcendent capability in the preparedness of possible weather and climate extremes; (3) the quantification study of reservoir inflow uncertainty and the corresponding experiments of multiple model ensemble techniques to reduce the uncertainty and discrepancy among the simulation results from different AI&DM models; (4) Incorporate ensemble hydrological forecasts to guide reservoir operation, given that one of the key strengths of AI&DM model is its structural flexibility that allows the inclusion of various types of forecast information to be added to support decision-making. Some recent studies over the Tres Marías dam (Mainardi et al., 2016), the Yuvacik dam (Uysal et al., 2018, Uysal et al., 2020), and Salto Grande Dam (Sinnige and Alvarado Montero, 2019) have demonstrated the combination of AI&DM models with model predictive control scheme could significantly improve the reservoir system performance in terms of flood reduction by more than 25% compared to deterministic forecast-based techniques. There are many innovative and new research domains that the AI&DM model could potentially contribute. Nevertheless, the fundamental model comparison and comprehensive evaluation are out of great importance to position the role of AI&DM models in assisting reservoir operation, and to better reveal the merits of these new technologies in solving any classical engineering and environmental science problems.

CRediT authorship contribution statement

Tiantian Yang: Conceptualization, Methodology, Software, Writing - original draft, Supervision, Funding acquisition. **Lujun Zhang:** Data curation, Visualization, Writing - review & editing, Formal analysis. **Taereem Kim:** Visualization, Software, Validation, Writing - review & editing. **Yang Hong:** Writing - review & editing. **Di Zhang:** Writing - review & editing. **Qidong Peng:** Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

This work is partially supported by the U.S. Department of Energy (DOE Prime Award # DE-IA0000018). This work is also financially supported by the National Key Research and Development Program of China (No.2018YFE0196000). The material is based upon work supported by the National Science Foundation under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the National Science Foundation under Grant No. NSF1802872.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2021.126723.

References

Aboutalebi, M., Bozorg-Haddad, O., Loáiciga, H.A., 2016. Simulation of methyl tertiary butyl ether concentrations in river-reservoir systems using support vector regression. J. Irrig. Drain. Eng. 142 (6), 04016015.

Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour. Res. 48 (1).

Adnan, R.M., Khosravinia, P., Karimi, B., Kisi, O., 2021. Prediction of hydraulics performance in drain envelopes using Kmeans based multivariate adaptive regression spline. Appl. Soft Comput. 100, 107008.

Adnan, R.M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O., Li, B., 2020. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. J. Hydrol. 586, 124371.

Adnan, R.M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., Kisi, O., 2019. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577, 123981.

Ahmadi, A., Karamouz, M., Moridi, A., 2010. Robust methods for identifying optimal reservoir operation strategies using deterministic and stochastic formulations. Water Resour. Manage. 24 (11), 2527–2552.

Akbari, M., Van Overloop, P.J., Afshar, A., 2011. Clustered K nearest neighbor algorithm for daily inflow forecasting. Water Resour. Manage. 25 (5), 1341–1357.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Statist. 46 (3), 175–185.

Apaydin, H., Feizi, H., Sattari, M.T., Colak, M.S., Shamshirband, S., Chau, K.-W., 2020. Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. Water 12 (5), 1500.

ASCE, 2000. Artificial neural networks in hydrology. I: Preliminary concepts. J. Hydrol. Eng. 5 (2), 115–123.

Ashaary, N.A., Wan Ishak, W.H. and Ku-Mahamud, K.R. (2015) Forecasting model for the change of reservoir water level stage based on temporal pattern of reservoir water level.

Atkeson, C.G., Moore, A.W., Schaal, S., 1997. Locally weighted learning. Lazy Learn. 11–73.

Babaei, M., Moeini, R., Ehsanzadeh, E., 2019. Artificial neural network and support vector machine models for inflow prediction of dam reservoir (case study: Zayandehroud Dam Reservoir). Water Resour. Manage. 33 (6), 2203–2218.

Bai, P., Liu, X., Xie, J., 2021. Simulating runoff under changing climatic conditions: a comparison of the long short-term memory network with two conceptual hydrologic models. J. Hydrol. 592, 125779.

Barnes Jr, G.W., Chung, F.I., 1986. Operational planning for California water system. J. Water Resour. Plann. Manage. 112 (1), 71–86.

Bessler, F.T., Savic, D.A., Walters, G.A., 2003. Water reservoir control with data mining. J. Water Resour. Plann. Manage. 129 (1), 26–34.

Bhatia, N., 2010. Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085.

Bonner, V., 1989. HEC-5: Simulation of Flood Control and Conservation Systems (for Microcomputers). Model-Simulation, Hydrologic Engineering Center, Davis, CA (USA).

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, pp. 144–152.

Boutaba, R., Salahuddin, M.A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., Caicedo, O.M., 2018. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. J. Internet Services Appl. 9 (1), 16.

Bozorg-Haddad, O., Aboutalebi, M., Ashofteh, P.-S., Loáiciga, H.A., 2018. Real-time reservoir operation using data mining techniques. Environ. Monit. Assess. 190 (10), 594.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Bremner, D., Demaine, E., Erickson, J., Iacono, J., Langerman, S., Morin, P., Toussaint, G., 2005. Output-sensitive algorithms for computing nearest-neighbour decision boundaries. Discrete Comput. Geometry 33 (4), 593–604.

Brodley, C.E., Utgoff, P.E., 1995. Multivariate decision trees. Machine Learning 19 (1), 45–77.

Brokamp, C., Jandarov, R., Rao, M., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos. Environ. 151, 1–11.

Buntine, W., Niblett, T., 1992. A further comparison of splitting rules for decision-tree induction. Machine Learning 8 (1), 75–85.

Caldwell, P., Segura, C., Gull Laird, S., Sun, G., McNulty, S.G., Sandercock, M., Boggs, J., Vose, J.M., 2015. Short-term stream water temperature observations permit rapid assessment of potential climate change impacts. Hydrol. Process. 29 (9), 2196–2211.

Cancelliere, A., Giuliano, G., Ancarani, A., Rossi, G., 2002. A neural networks approach for deriving irrigation reservoir operating rules. Water Resour. Manage. 16 (1), 71–88.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2012. Data-driven dynamic emulation modelling for the optimal management of environmental systems. Environ. Modell. Software 34, 30–43.

Castelletti, A., Galelli, S., Restelli, M., Soncini-Sessa, R., 2010. Tree-based reinforcement learning for optimal water reservoir operation. Water Resour. Res. 46 (9).

Chang, F.-J., Wang, Y.-C., Tsai, W.-P., 2016. Modelling intelligent water resources allocation for multi-users. Water Resour. Manage. 30 (4), 1395–1413.

Chaturvedi, M., Srivastava, D., 1981. Study of a complex water resources system with screening and simulation models. Water Resour. Res. 17 (4), 783–794.
Chaves, P., Chang, F.-J., 2008. Intelligent reservoir operation system based on evolving artificial neural networks. Adv. Water Resour. 31 (6), 926–936.
Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, pp. 785–794.
Cheng, C.-C., Hsu, N.-S., Wei, C.-C., 2008. Decision-tree analysis on optimal release of reservoir storage under typhoon warnings. Nat. Hazards 44 (1), 65–84.
Cheng, M., Fang, F., Kinouchi, T., Navon, I., Pain, C., 2020. Long lead-time daily and monthly streamflow forecasting using machine learning methods. J. Hydrol. 590, 125376.
Chung, F.I., Archer, M.C., DeVries, J.J., 1989. Network flow algorithm applied to California aqueduct simulation. J. Water Resour. Plann. Manage. 115 (2), 131–147.
CNN, 2019. Video shows flooding after Houston reservoir release. CNN news, 2019 Aug, 24th https://www.cnn.com/videos/us/2019/08/24/houston-addicks-and-barker-dams-reservoirs-harvey-release.cnn.
CNN, 2020. These before and after images show how much a Michigan dam failure drained a lake. CNN news, 2020 May, 20th https://www.cnn.com/2020/05/20/us/michigan-dam-failure-before-after-photos-trnd/index.html.
Coomans, D., Massart, D.L., 1982. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. Anal. Chim. Acta 136, 15–27.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.
Coulibaly, P., Anctil, F., Bobee, B., 2001. Multivariate reservoir inflow forecasting using temporal neural networks. J. Hydrol. Eng. 6 (5), 367–376.
Coulibaly, P., Anctil, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol. 230 (3–4), 244–257.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.
Cunningham, P., Delany, S.J., 2020. k-Nearest Neighbour Classifiers. arXiv preprint arXiv:2004.04523.
Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2005. Groundwater level forecasting using artificial neural networks. J. Hydrol. 309 (1–4), 229–240.
Dawson, C.W., Wilby, R., 1998. An artificial neural network approach to rainfall-runoff modelling. Hydrol. Sci. J. 43 (1), 47–66.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81 (11), 3178–3192.
Deka, P.C., 2014. Support vector machine applications in the field of hydrology: a review. Appl. Soft Comput. 19, 372–386.
Dietterich, T., 1995. Overfitting and undercomputing in machine learning. ACM Comput. Surveys (CSUR) 27 (3), 326–327.
Ding, Z., Wen, X., Tan, Q., Yang, T., Fang, G., Lei, X., Zhang, Y., Wang, H., 2021. A forecast-driven decision-making model for long-term operation of a hydro-wind-photovoltaic hybrid system. Appl. Energy 291, 116820.
Draper, A.J., Munevar, A., Arora, S.K., Reyes, E., Parker, N.L., Chung, F.I., Peterson, L.E., 2004. CalSim: Generalized model for reservoir system analysis. J. Water Resour. Plann. Manage. 130 (6), 480–489.
Efron, B., 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82 (397), 171–185.
Efron, B., 1992. Breakthroughs in Statistics. Springer, pp. 569–593.
Erdal, H.I., Karakurt, O., 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. J. Hydrol. 477, 119–128.
Esmaeilzadeh, B., Sattari, M.T., Samadianfard, S., 2017. Performance evaluation of ANNs and an M5 model tree in Sattarkhan Reservoir inflow prediction. ISH J. Hydraulic Eng. 23 (3), 283–292.
Fan, H., Jiang, M., Xu, L., Zhu, H., Cheng, J., Jiang, J., 2020. Comparison of long short term memory networks and the hydrological model in runoff simulation. Water 12 (1), 175.
Feng, Z.-K., Niu, W.-J., Tang, Z.-Y., Jiang, Z.-Q., Xu, Y., Liu, Y., Zhang, H.-R., 2020. Monthly runoff time series prediction by variational mode decomposition and support vector machine based on quantum-behaved particle swarm optimization. J. Hydrol. 583, 124627.
Fix, E., 1951. Discriminatory analysis: nonparametric discrimination, consistency properties, USAF school of Aviation Medicine.
French, M.N., Krajewski, W.F., Cuykendall, R.R., 1992. Rainfall forecasting in space and time using a neural network. J. Hydrol. 137 (1–4), 1–31.
Freund, Y., Mason, L., 1999. The alternating decision tree learning algorithm, pp. 124–133.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378.
Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. Int. J. Emerg. Technol. Adv. Eng. 2 (4), 42–47.
Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: Continual prediction with LSTM.
Gers, F.A., Schraudolph, N.N., Schmidhuber, J., 2002. Learning precise timing with LSTM recurrent networks. J. Machine Learn. Res. 3 (Aug), 115–143.
Ghimire, B.N., Shrestha, R.N., Bhatta, U.D., 2020. Advances in Water Resources Engineering and Management. Springer, pp. 27–42.
Goh, A.T., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. 9 (3), 143–151.
Govindaraju, R.S., 2000. Artificial neural networks in hydrology. II: hydrologic applications. J. Hydrol. Eng. 5 (2), 124–137.
Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks 18 (5–6), 602–610.
Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS Technical Report 14 (1), 5–16.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91.
Hand, D.J., 2007. Principles of data mining. Drug Safety 30 (7), 621–622.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
Haykin, S., Network, N., 2004. A comprehensive foundation. Neural Networks 2 (2004), 41.
Hecht-Nielsen, R., 1992. Neural Networks for Perception. Elsevier, pp. 65–93.
Hejazi, M.I., Cai, X., Ruddell, B.L., 2008. The role of hydrologic information in reservoir operation–learning from historical releases. Adv. Water Resour. 31 (12), 1636–1650.
Hipni, A., El-shafie, A., Najah, A., Karim, O.A., Hussain, A., Mukhlisin, M., 2013. Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). Water Resour. Manage. 27 (10), 3803–3823.
Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertainty Fuzziness Knowledge Based Syst. 6 (02), 107–116.
Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., 2001. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies. A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.
Hoskins, J.C., Himmelblau, D., 1988. Artificial neural network models of knowledge representation in chemical engineering. Comput. Chem. Eng. 12 (9–10), 881–890.
HoustonChronicle, 2019. How did Houston, counties escape blame for flooded homes in Addicks and Barker reservoirs? Houston Chronicle News December 2019 https://www.houstonchronicle.com/news/houston-texas/houston/article/How-did-Houston-counties-escape-blame-for-Harvey-14928293.php.
HoustonPublicMedia, 2019. Federal Judge Rules Army Corps Liable For Flooding Homes In Addicks And Barker Reservoirs During Harvey. Houston Public Media News Dec 17th 2019, https://www.houstonpublicmedia.org/articles/news/local/2019/12/17/354502/federal-judge-rules-army-corps-liable-for-flooding-homes-in-addicks-and-barker-reservoirs-during-harvey/.
Hsu, K.L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall-runoff process. Water Resour. Res. 31 (10), 2517–2530.
Hutson, M., 2018. Has Artificial Intelligence Become Alchemy? American Association for the Advancement of Science.
Imandoust, S.B., Bolandraftar, M., 2013. Application of k-nearest neighbor (knn) approach for predicting economic events: theoretical background. Int. J. Eng. Res. Appl. 3 (5), 605–610.
Jain, A.K., Mao, J., Mohiuddin, K.M., 1996. Artificial neural networks: a tutorial. Computer 29 (3), 31–44.
Jain, S., Das, A., Srivastava, D., 1999. Application of ANN for reservoir inflow prediction and operation. J. Water Resour. Plann. Manage. 125 (5), 263–271.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning, Springer.
Ji, C.-M., Zhou, T., Huang, H.-T., 2014. Operating rules derivation of Jinsha reservoirs system with parameter calibrated support vector regression. Water Resour. Manage. 28 (9), 2435–2451.
Ji, Y., Lei, X., Cai, S., Wang, X., 2016. Application of a classifier based on data mining techniques in water supply operation. Water 8 (12), 599.
Johnson, R.W., 2001. An introduction to the bootstrap. Teaching Statistics 23 (2), 49–54.
Kao, I.-F., Zhou, Y., Chang, L.-C., Chang, F.-J., 2020. Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. J. Hydrol. 583, 124631.
Kargar, K., Samadianfard, S., Parsa, J., Nabipour, N., Shamshirband, S., Mosavi, A., Chau, K.-W., 2020. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. Eng. Appl. Comput. Fluid Mech. 14 (1), 311–322.
Kasiviswanathan, K., Sudheer, K., 2017. Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. Stoch. Env. Res. Risk Assess. 31 (7), 1659–1670.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree, pp. 3146–3154.
Kim, T., Shin, J.Y., Kim, H., Heo, J.H., 2020. Ensemble-Based Neural Network Modeling for Hydrologic Forecasts: Addressing Uncertainty in the Model Structure and Input Variable Selection. Water Resour. Res. 56(6), e2019WR026262.
Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., Gourley, J.J., Hong, Y., 2021. Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: a case study of four watersheds with different hydro-climatic regions across the CONUS. J. Hydrol. 598, 126423.
Kişi, Ö., 2004. River flow modeling using artificial neural networks. J. Hydrol. Eng. 9 (1), 60–63.
Kişi, Ö., 2007. Streamflow forecasting using different artificial neural network algorithms. J. Hydrol. Eng. 12 (5), 532–539.
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22 (11), 6005–6022.

Krishnan, R., Sivakumar, G., Bhattacharya, P., 1999. Extracting decision trees from trained neural networks. Pattern Recogn. 32 (12).
Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., 2005. Applied linear statistical models, McGraw-Hill Irwin New York.
Labadie, J.W., 2004. Optimal operation of multireservoir systems: state-of-the-art review. J. Water Resour. Plann. Manage. 130 (2), 93–111.
Lal, A., Datta, B., 2018. Development and implementation of support vector machine regression surrogate models for predicting groundwater pumping-induced saltwater intrusion into coastal aquifers. Water Resour. Manage. 32 (7), 2405–2419.
Legates, D.R., McCabe Jr, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Li, X., Sha, J., Wang, Z.-L., 2016. A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. Hydrol. Res. 48 (5), 1214–1225.
Li, Z., Chen, T., Wu, Q., Xia, G., Chi, D., 2020. Application of penalized linear regression and ensemble methods for drought forecasting in Northeast China. Meteorol. Atmos. Phys. 132 (1), 113–130.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.
Lima, C.H., Lall, U., 2010a. Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge regression model. J. Hydrol. 380 (3–4), 438–449.
Lima, C.H.R., Lall, U., 2010b. Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge regression model. J. Hydrol. 380 (3), 438–449.
Lin, G.-F., Chen, G.-R., Huang, P.-Y., Chou, Y.-C., 2009. Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. J. Hydrol. 372 (1–4), 17–29.
Lin, J.-Y., Cheng, C.-T., Chau, K.-W., 2006. Using support vector machines for long-term discharge prediction. Hydrol. Sci. J. 51 (4), 599–612.
Liu, X., Chen, L., Zhu, Y., Singh, V.P., Qu, G., Guo, X., 2017a. Multi-objective reservoir operation during flood season considering spillway optimization. J. Hydrol. 552, 554–563.
Liu, Y., Qin, H., Zhang, Z., Yao, L., Wang, Y., Li, J., Liu, G., Zhou, J., 2019. Deriving reservoir operation rule based on Bayesian deep learning method considering multiple uncertainties. J. Hydrol. 579, 124207.
Liu, Y., Sang, Y.-F., Li, X., Hu, J., Liang, K., 2017b. Long-term streamflow forecasting based on relevance vector machine model. Water 9 (1), 9.
Loh, W.Y., 2014. Fifty years of classification and regression trees. Int Statistical Rev. 82 (3), 329–348.
Lohani, A., Kumar, R., Singh, R., 2012. Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. J. Hydrol. 442, 23–35.
Louks, D.P., Sigvaldason, O.T., 1981. Multiple Reservoir Operation in North America. ASCE, pp. 711–728.
Lund, J.R., Guzman, J., 1999. Derived operating rules for reservoirs in series or in parallel. J. Water Resour. Plann. Manage. 125 (3), 143–153.
LynkerTech, 2018. CRAM: Central Resrouces Allocation Model. https://lynkertech.shinyapps.io/cram/.
Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ. Modell. Software 25 (8), 891–909.
Mainardi, F., Schwanenberg, D., Alvarado, R., Assis dos Reis, A., Naumann, S., Collischonn, W., 2016. Performance assessment of deterministic and probabilistic weather predictions for the short-term optimization of a tropical hydropower reservoir, pp. EPSC2016-5156.
Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting Methods and Applications. John Wiley & Sons.
Malik, A., Tikhamarine, Y., Souag-Gamane, D., Kisi, O., Pham, Q.B., 2020. Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction. Stoch. Env. Res. Risk Assess. 34 (11), 1755–1773.
Marill, K.A., 2004. Advanced statistics: linear regression, part I: simple linear regression. Acad. Emerg. Med. 11 (1), 87–93.
Marquardt, D.W., Snee, R.D., 1975. Ridge regression in practice. Am. Statist. 29 (1), 3–20.
Masselot, P., Dabo-Niang, S., Chebana, F., Ouarda, T.B., 2016. Streamflow forecasting using functional regression. J. Hydrol. 538, 754–766.
McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bullet. Mathem. Biophys. 5 (4), 115–133.
MercuryNews, 2017. Oroville Dam: Feds and state officials ignored warnings 12 years ago. Mercury News Feb 12th 2017, https://www.mercurynews.com/2017/02/12/oroville-dam-feds-and-state-officials-ignored-warnings-12-years-ago/.
Miao, C., Duan, Q., Sun, Q., Huang, Y., Kong, D., Yang, T., Ye, A., Di, Z., Gong, W., 2014. Assessment of CMIP5 climate models and projected temperature changes over Northern Eurasia. Environ. Res. Lett. 9 (5), 055007.
Mingers, J., 1989. An empirical comparison of pruning methods for decision tree induction. Machine Learning 4 (2), 227–243.
Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to Linear Regression Analysis. John Wiley & Sons.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50 (3), 885–900.
Mosavi, A., Ozturk, P., Chau, K.-W., 2018. Flood prediction using machine learning models: literature review. Water 10 (11), 1536.
Myers, R.H., Myers, R.H., 1990. Classical and Modern Regression with Applications. Duxbury Press, Belmont, CA.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. 10 (3), 282–290.
NewYorkTimes (2017) What Happened at the Oroville Dam New York times News Feb 13th 2017, https://www.nytimes.com/interactive/2017/02/13/us/oroville-dam.html.
Nikoo, M.R., Kerachian, R., Karimi, A., Azadnia, A.A., Jafarzadegan, K., 2014. Optimal water and waste load allocation in reservoir–river systems: a case study. Environ. Earth Sci. 71 (9), 4127–4142.
Niu, W.-J., Feng, Z.-K., Feng, B.-F., Min, Y.-W., Cheng, C.-T., Zhou, J.-Z., 2019. Comparison of multiple linear regression, artificial neural network, extreme learning machine, and support vector machine in deriving operation rule of hydropower reservoir. Water 11 (1), 88.
Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24 (12), 1565–1567.
Oliveira, R., Loucks, D.P., 1997. Operating rules for multireservoir systems. Water Resour. Res. 33 (4), 839–852.
Ombadi, M., Nguyen, P., Sorooshian, S., Hsu, K.-l., 2020. Evaluation of methods for causal discovery in hydrometeorological systems. Water Resour. Res. 56(7), e2020WR027251.
Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.
Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.
Rahnamay Naeini, M., Yang, T., Tavakoly, A., Analui, B., AghaKouchak, A., Hsu, K.-L., Sorooshian, S., 2020. A model tree generator (MTG) framework for simulating hydrologic systems: application to reservoir routing. Water 12 (9), 2373.
Raman, H., Sunilkumar, N., 1995. Multivariate modelling of water resources time series using artificial neural networks. Hydrol. Sci. J. 40 (2), 145–163.
Raso, L., Schwanenberg, D., van de Giesen, N., van Overloop, P.J., 2014. Short-term optimal operation of water systems using ensemble forecasts. Adv. Water Resour. 71, 200–208.
Reddy, M.J., Kumar, D.N., 2006. Optimal reservoir operation using multi-objective evolutionary algorithm. Water Resour. Manage. 20 (6), 861–878.
Ren, K., Fang, W., Qu, J., Zhang, X., Shi, X., 2020. Comparison of eight filter-based feature selection methods for monthly streamflow forecasting – three case studies on CAMELS data sets. J. Hydrol. 586, 124897.
Rokach, L., Maimon, O., 2005. Data Mining and Knowledge Discovery Handbook. Springer, pp. 165–192.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.
Sahoo, B.B., Jha, R., Singh, A., Kumar, D., 2019. Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys. 67 (5), 1471–1481.
Sahoo, S., Jha, M.K., 2013. Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. Hydrogeol. J. 21 (8), 1865–1887.
Samadianfard, S., Jarhan, S., Salwana, E., Mosavi, A., Shamshirband, S., Akib, S., 2019. Support vector regression integrated with fruit fly optimization algorithm for river flow forecasting in Lake Urmia Basin. Water 11 (9).
Sattari, M.T., Yurekli, K., Pal, M., 2012. Performance evaluation of artificial neural network approaches in forecasting reservoir inflow. Appl. Math. Model. 36 (6), 2649–2657.
Schmidt, L., Heße, F., Attinger, S. and Kumar, R. (2020) Challenges in applying machine learning models for hydrological inference: a case study for flooding events across Germany. Water Resour. Res. 56(5), e2019WR025924.
Schwanenberg, D., Raso, L. ,and Student, H.P., 2012. Tree-Based Model Predictive Control for Optimizing Hydro Power under Uncertainty.
Schwanenberg, D., Xu, M., Ochterbeck, T., Allen, C., Karimanzira, D., 2014. Short-term management of hydropower assets of the Federal Columbia River power system. J. Appl. Water Eng. Res. 2 (1), 25–32.
Seber, G.A., Lee, A.J., 2012. Linear Regression Analysis. John Wiley & Sons.
Shabani, S., Samadianfard, S., Sattari, M.T., Mosavi, A., Shamshirband, S., Kmet, T., Várkonyi-Kóczy, A.R., 2020. Modeling pan evaporation using Gaussian process regression k-nearest neighbors random forest and support vector machines. Compar. Analysis. Atmosphere 11 (1), 66.
Shamshirband, S., Esmaeilbeiki, F., Zarehaghi, D., Neyshabouri, M., Samadianfard, S., Ghorbani, M.A., Mosavi, A., Nabipour, N., Chau, K.-W., 2020. Comparative analysis of hybrid models of firefly optimization algorithm with support vector machines and multilayer perceptron for predicting soil temperature at different depths. Eng. Appl. Comput. Fluid Mech. 14 (1), 939–953.
Sigvaldson, O., 1976. A simulation model for operating a multipurpose multireservoir system. Water Resour. Res. 12 (2), 263–278.
Singh, J., Knapp, H.V., Arnold, J., Demissie, M., 2005. Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. JAWRA J. Am. Water Resour. Assoc. 41 (2), 343–360.
Sinnige, R.P., Alvarado Montero, R.T., 2019. Application of the tree-based ensemble method for the Salto Grande case with RTC-Tools. Deltares RTC-Tool Report 2019-02(4).
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statist. Comput. 14 (3), 199–222.
Steinberg, D., Colla, P., 2009. CART: classification and regression trees. The top ten algorithms in data mining 9, 179.

Tao, Y., Yang, T., Faridzad, M., Jiang, L., He, X., Zhang, X., 2018. Non-stationary bias correction of monthly CMIP5 temperature projections over China using a residual-based bagging tree model. Int. J. Climatol. 38 (1), 467–482.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmospheres 106 (D7), 7183–7192.
Tongal, H., Booij, M.J., 2018. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. J. Hydrol. 564, 266–282.
Toussaint, G., 2005. Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. Int. J. Comput. Geometry Appl. 15 (02), 101–150.
Uysal, G., Alvarado-Montero, R., Schwanenberg, D., Şensoy, A., 2018. Real-time flood control by tree-based model predictive control including forecast uncertainty: a case study reservoir in Turkey. Water 10 (3), 340.
Uysal, G., Montero, R.-A., Schwanenberg, D., Sensoy, A., 2020. Real-Time Reservoir Operation by Tree-Based Model Predictive Control Including Forecast Uncertainty, p. 896.
Valipour, M., Banihabib, M.E., Behbahani, S.M.R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. J. Hydrol. 476, 433–441.
Vapnik, V., 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.
Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Networks 10 (5), 988–999.
Wei, C.-C., 2012. Discretized and continuous target fields for the reservoir release rules during floods. Water Resour. Manage. 26 (12), 3457–3477.
Wei, C.-C., 2015. Comparing lazy and eager learning models for water level forecasting in river-reservoir basins of inundation regions. Environ. Modell. Software 63, 137–155.
Wei, C.C., Hsu, N.S., 2008. Derived operating rules for a reservoir operation system: Comparison of decision trees, neural decision trees and fuzzy decision trees. Water Resour. Res. 44 (2).
Wu, C., Chau, K.W., Li, Y.S., 2009. Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour. Res. 45 (8).
Wu, Y., Ding, Y., Zhu, Y., Feng, J., Wang, S., 2020. Complexity to forecast flood: problem definition and spatiotemporal attention LSTM solution. Complexity.
Xiang, Z., Yan, J. and Demir, I. (2020) A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resources Research 56(1), e2019WR025326.
Xie, Z., Lou, I., Ung, W.K., Mok, K.M., 2012. Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs. Mathematical problems in engineering 2012.
Xu, Z., Li, J., 2002. Short-term inflow forecasting using an artificial neural network model. Hydrol. Process. 16 (12), 2423–2439.
Yang, M., Wang, H., Jiang, Y., Lu, X., Xu, Z., Sun, G., 2020a. GECA proposed ensemble–KNN method for improved monthly runoff forecasting. Water Resour. Manage. 34 (2), 849–863.
Yang, T., Asanjan, A.A., Faridzad, M., Hayatbini, N., Gao, X., Sorooshian, S., 2017a. An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. Inf. Sci. 418, 302–316.
Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S., Liu, X., 2017b. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resour. Res. 53 (4), 2786–2812.
Yang, T., Gao, X., Sellars, S.L., Sorooshian, S., 2015. Improving the multi-objective evolutionary optimization algorithm for hydropower reservoir operations in the California Oroville-Thermalito complex. Environ. Modell. Software 69, 262–279.
Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resour. Res. 52 (3), 1626–1651.
Yang, T., Liu, X., Wang, L., Bai, P., Li, J., 2020b. Simulating hydropower discharge using multiple decision tree methods and a dynamical model merging technique. J. Water Resour. Plann. Manage. 146 (2), 04019072.
Yang, T., Tao, Y., Li, J., Zhu, Q., Su, L., He, X., Zhang, X., 2018. Multi-criterion model ensemble of CMIP5 surface air temperature over China. Theor. Appl. Climatol. 132 (3), 1057–1072.
Yates, D., Sieber, J., Purkey, D., Huber-Lee, A., 2005. WEAP21—A demand-, priority-, and preference-driven water planning model: part 1: model characteristics. Water Int. 30 (4), 487–500.
Yeh, W.W.G., 1985. Reservoir management and operations models: a state-of-the-art review. Water Resour. Res. 21 (12), 1797–1818.
Yu, P.-S., Chen, S.-T., Chang, I.-F., 2006a. Support vector regression for real-time flood stage forecasting. J. Hydrol. 328 (3–4), 704–716.
Yu, P.-S., Chen, S.-T., Chang, I.F., 2006b. Support vector regression for real-time flood stage forecasting. J. Hydrol. 328 (3), 704–716.
Yu, X., Liong, S.-Y., 2007. Forecasting of hydrologic time series with ridge regression in feature space. J. Hydrol. 332 (3–4), 290–302.
Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 31 (7), 1235–1270.
Yuan, X., Chen, C., Lei, X., Yuan, Y., Adnan, R.M., 2018. Monthly runoff forecasting based on LSTM–ALO model. Stoch. Env. Res. Risk Assess. 32 (8), 2199–2212.
Zadeh, M.R., Amin, S., Khalili, D., Singh, V.P., 2010. Daily outflow prediction by multi layer perceptron with logistic sigmoid and tangent sigmoid activation functions. Water Resour. Manage. 24 (11), 2673–2688.
Zagona, E.A., Fulp, T.J., Shane, R., Magee, T., Goranflo, H.M., 2001. Riverware: a generalized tool for complex reservoir system modeling 1. JAWRA J. Am. Water Resour. Assoc. 37 (4), 913–929.
Zealand, C.M., Burn, D.H., Simonovic, S.P., 1999. Short term streamflow forecasting using artificial neural networks. J. Hydrol. 214 (1–4), 32–48.
Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X., Zhuang, J., 2018. Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. J. Hydrol. 565, 720–736.
Zhang, D., Peng, Q., Lin, J., Wang, D., Liu, X., Zhuang, J., 2019. Simulating reservoir operation using a recurrent neural network algorithm. Water 11 (4), 865.
Zhang, J., Cai, X., Lei, X., Liu, P., Wang, H., 2020. Real-time reservoir flood control operation enhanced by data assimilation. Hydrol. Earth Syst. Sci. Discuss. 1–37.
Zhao, J., Zhao, C., Zhang, F., Wu, G., Wang, H., 2018. Water quality prediction in the waste water treatment process based on ridge regression echo state network. IOP Conf. Series: Mater. Sci. Eng. 435, 012025.
Zhao, T., Yang, D., Cai, X., Zhao, J., Wang, H., 2012. Identifying effective forecast horizon for real-time reservoir operation under a limited inflow forecast. Water Resour. Res. 48 (1).
Zhou, Y., Guo, S., Liu, P., Xu, C.-Y., Zhao, X., 2016. Derivation of water and power operating rules for multi-reservoirs. Hydrol. Sci. J. 61 (2), 359–370.
Zhu, S., Luo, X., Yuan, X., Xu, Z., 2020. An improved long short-term memory network for streamflow forecasting in the upper Yangtze River. Stoch. Environ. Res. Risk Assess. 34 (9), 1313–1329.
Zolfaghari, M., Golabi, M.R., 2021. Modeling and predicting the electricity production in hydropower using conjunction of wavelet transform, long short-term memory and random forest models. Renewable Energy.