Data-driven artificial intelligence-based streamflow forecasting, a review of methods, applications, and tools

Received: 19 March 2024 | Accepted: 8 August 2024

DOI: 10.1111/1752-1688.13229

AMERICAN WATER RESOURCES ASSOCIATION
JAWRA
Journal of the American Water Resources Association

LITERATURE REVIEW

Data-driven artificial intelligence-based streamflow forecasting, a review of methods, applications, and tools

Heerbod Jahanbani | Khandakar Ahmed | Bruce Gu

Institute for Sustainable Industries and Liveable Cities (ISILC), Victoria University, Melbourne, Victoria, Australia

Correspondence
Heerbod Jahanbani, Institute for Sustainable Industries and Liveable Cities (ISILC), Victoria University, Melbourne, VIC, Australia.
Email: heerbod.jahanbani@live.vu.edu.au

Abstract

Data-driven artificial intelligence (DDAI) prediction has gained much attention, especially in recent years, because of its power and flexibility compared to traditional approaches. In hydrology, streamflow forecasting is one of the areas that took advantage of utilizing DDAI-based forecasting, given the weakness of the old approaches (e.g., physical-based approaches). Since many different techniques and tools have been used for streamflow forecasting, there is a new way to explore them. This manuscript reviews the recent (2011–2023) applications of DDAI in streamflow prediction. It provides a background of DDAI-based techniques, including machine learning algorithms and methods for pre-processing the data and optimizing or enhancing the machine learning approaches. We also explore the applications of DDAI techniques in streamflow forecasting. Finally, the most common tools for utilizing DDAI techniques in streamflow forecasting are presented.

KEYWORDS
streamflow forecasting, data-driven, artificial intelligence (AI), machine learning, stochastic data

1 | INTRODUCTION

With climate change and population growth, the essential phenomenon on earth (water) has been significantly under threat in the last few decades. According to the most recent Intergovernmental Panel on Climate Change (IPCC) studies on climate change (Masson-Delmotte et al., 2021), the weather extremes we are seeing now (heatwaves, heavy rainfall, and droughts) are getting more intense and more frequent, directly linked to our emissions of greenhouse gases. This interpretation emphasizes that there will be more climate change/variability in the

Abbreviations: AI, artificial intelligence; ANFIS, adaptive neuro-fuzzy inference system; ANN, artificial neural network; BA, bat algorithm; CPN, counterpropagation network; DDAI, data-driven artificial intelligence; DL, deep learning; DWT, discrete wavelet transform; EC, evolutionary computation; ENSO, El Niño Southern Oscillation; FFA, firefly optimization algorithm; FFNN, feedforward neural network; FFS, forward feature selection; FL, fuzzy logic; FS, fuzzy set; GA, genetic algorithm; GB, gradient boosting; GP, genetic programming; GT, gamma test; HHT, Hilbert–Huang transform; IPCC, Intergovernmental Panel on Climate Change; K-NN, K-nearest neighbors; LSSVM, least square support vector Machine; LSTM, long short-term memory; ML, machine learning; MLR, multiple linear regression; PCA, principal component analysis; PSO, particle swarm optimization; RBF, radial basis function; RFE, recursive feature elimination; RNN, recurrent neural network; SARSA, State Action Reward State Action; SPEI, Standardized Precipitation Evapotranspiration Index; SPI, Standardized Precipitation Index; SRL, systematic literature review; SSI, Standardized Streamflow Index; SVM, support vector machine; WT, wavelet transform.

Paper No. JAWR-24-0037-L of the *Journal of the American Water Resources Association* (JAWR).

Discussions are open until six months from publication.

Research Impact Statement

This review will serve as a guide for researchers and practitioners to select the most appropriate data-driven AI/ML methods for their specific streamflow forecasting needs.

future, making it more challenging to predict streamflow accurately. The intense change/variability can negatively impact water resource decisions by making them more complicated and less accurate in the future.

Hydrological models are designed to simplify the complexity of nature and simulate water behavior such as runoff. Physical-based hydrological models use mathematical and numerical models that try to achieve the goal. However, they come with limitations (Tan et al., 2022), such as (a) they contain a set of rules that cannot fully capture the complexity of nature, (b) are difficult to measure parameters accurately and do not always perform well, and (c) they are not designed for future prediction. Therefore, another alternative is statistical models. Statistical models achieve their performance by relying on statistical analysis from the available data and gaining information and patterns rather than pre-knowledge based on the system's characteristics (Sit et al., 2020). This capability has been a great motivation for applying data-driven artificial intelligence (DDAI) techniques in hydrological modeling.

It is vital to note that artificial intelligence (AI) and machine learning (ML) tools can sometimes uncover correlations that are either physically inaccurate or defy clear explanations. Therefore, it is imperative to repeatedly be cautious against blind faith in any method without a thorough grasp of the underlying processes.

AI is a relatively new science in which programs/machines experience, learn, act, and achieve goals using the gained intelligence similar to humans. ML is one of the significant parts of AI that brings the algorithms of automatic learning, improvements, and increased chance of success (Hrnjica & Mehr, 2020). ML performs a skillful job by training in which the learning (intelligence) is based on statistical approaches (Mohammed et al., 2016). In other words, ML consists of techniques that utilize adaptive pathways to organize data and predict. The adaptive pathway has a training process that learns from experience to reach the target (Bonaccorso, 2018). Deep learning (DL) is a subcategory of artificial neural networks (ANNs) with many layers and training processes (Hrnjica & Mehr, 2020). In other words, DL is a sophisticated connected pattern matching. DL can solve more complex problems based on the knowledge built on the previous solutions (Ian et al., 2016) (Figure 1).

Hydrological modeling and forecasting use four aspects of AI—(1) regression and classification by utilizing ML algorithms, (2) fuzzy set (FS) and fuzzy logic (FL) approach, (3) evolutionary computation (EC), and (4) pre-processing methods. The first aspect (regression and classification) is the fundamental part of data-driven streamflow forecasting. The other aspects play the role of analyzing the input variables, improving the regression or classification, and consequently improving the prediction. In recent years, the combination of aspects one with two to four has been called hybrid DDAI in hydrology (Ibrahim et al., 2022). FL/FS approach (Zadeh, 1965) works with uncertainties using the FS membership. EC helps with the optimization of the parameters of the regression or the membership function of the FL. Pre-processing approaches such as wavelet conjunction models are used for analyzing and improving the input time series.

Aside from streamflow forecasting, AI has been used in different areas of hydrology such as drought (Belayneh et al., 2014, 2016; Dehghani et al., 2014; Ganguli & Reddy, 2014; Khan et al., 2020; Mohamadi et al., 2020), flood (Bahram Saghafian & Dehghani, 2017; Chang et al., 2014; Gizaw & Gan, 2016; Jimeno-Sáez et al., 2017), water level forecasting (Phan & Nguyen, 2020; Sapitang et al., 2020; Zhu et al., 2020), precipitation prediction or downscaling (Ahmed et al., 2020; Sachindra et al., 2018; Vandal et al., 2018), and evapotranspiration (Jahanbani et al., 2011; Jahanbani & El-Shafie, 2011).

2 | SCOPE OF THIS PAPER

Several studies (Ibrahim et al., 2022; Sit et al., 2020; Yaseen et al., 2015) reviewed the application of AI in hydrology and water resources. However, there is no comprehensive and most recent review of the application of DDAI-based approaches in streamflow forecasting. This manuscript intends to review the recent papers from 2011 to 2023, focusing mainly on works published between 2015 and 2021. This paper reviews the application of DDAI-based approaches in daily or longer-than-daily streamflow prediction and does not review cases like real-time forecasting or flood forecasting. Although we briefly describe the most common types of DDAI-based approaches (e.g., ML algorithms), the paper mainly focuses on DDAI used in streamflow forecasting since the review of DDAI and their mathematical equations have been explained in several books. The main differences between this paper with similar works are summarized below:

a. most recent comprehensive review of the application of DDAI-based approaches;
b. introduction of stochastic data in streamflow prediction;

AMERICAN WATER RESOURCES ASSOCIATION JAWRA JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

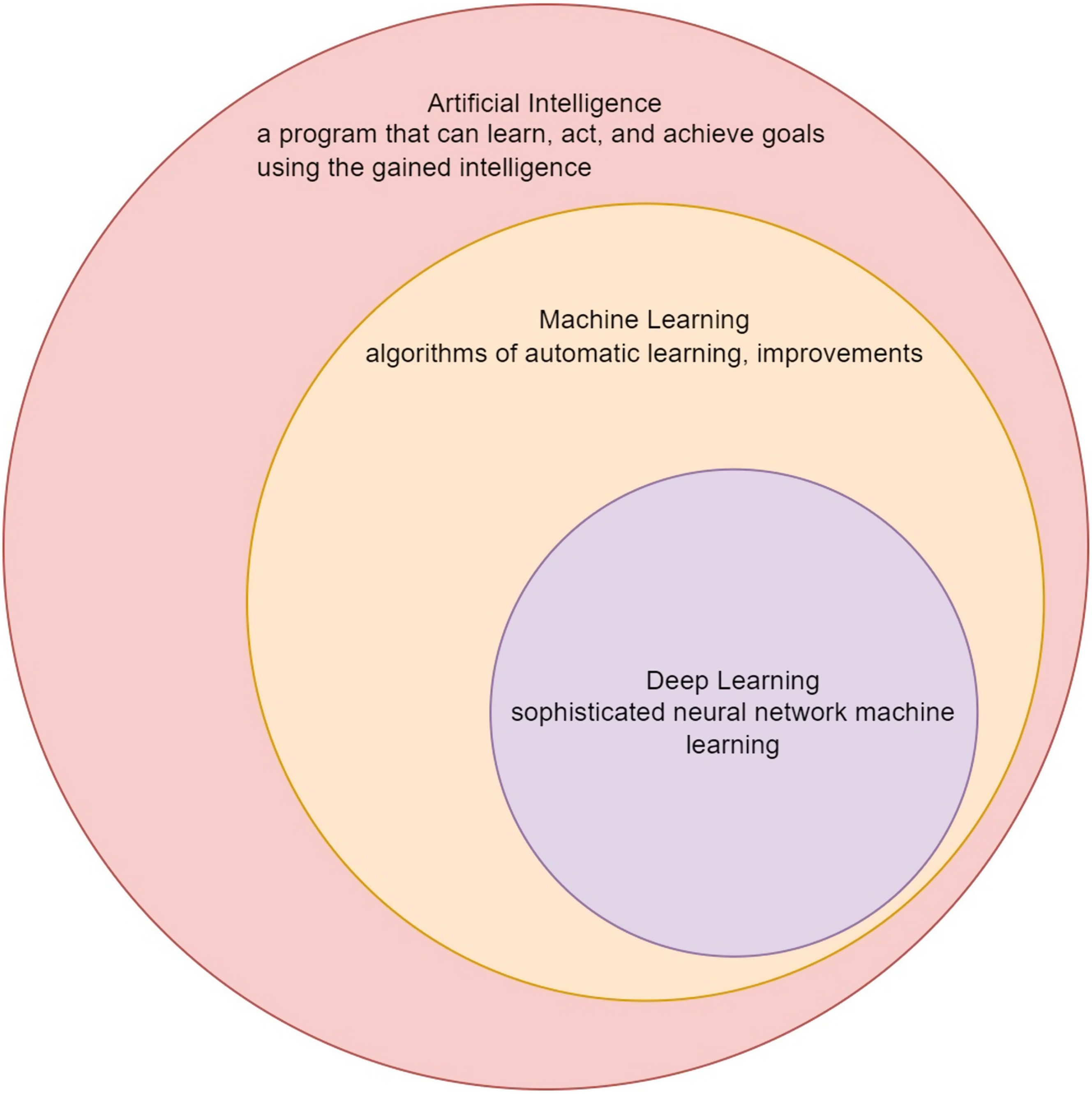

FIGURE 1 Artificial intelligence (AI) versus machine learning versus deep learning.

c. summary of tools for DDAI-based modeling; and
d. practical example of how to use the tool with a simple example of streamflow forecasting (Appendix B).

Figure 2 visualizes a summary of the paper structure.

3 | SYSTEMATIC LITERATURE REVIEW

The systematic literature review (SLR) for this review paper includes three stages: planning, conducting, and reporting (Kitchenham, 2004).

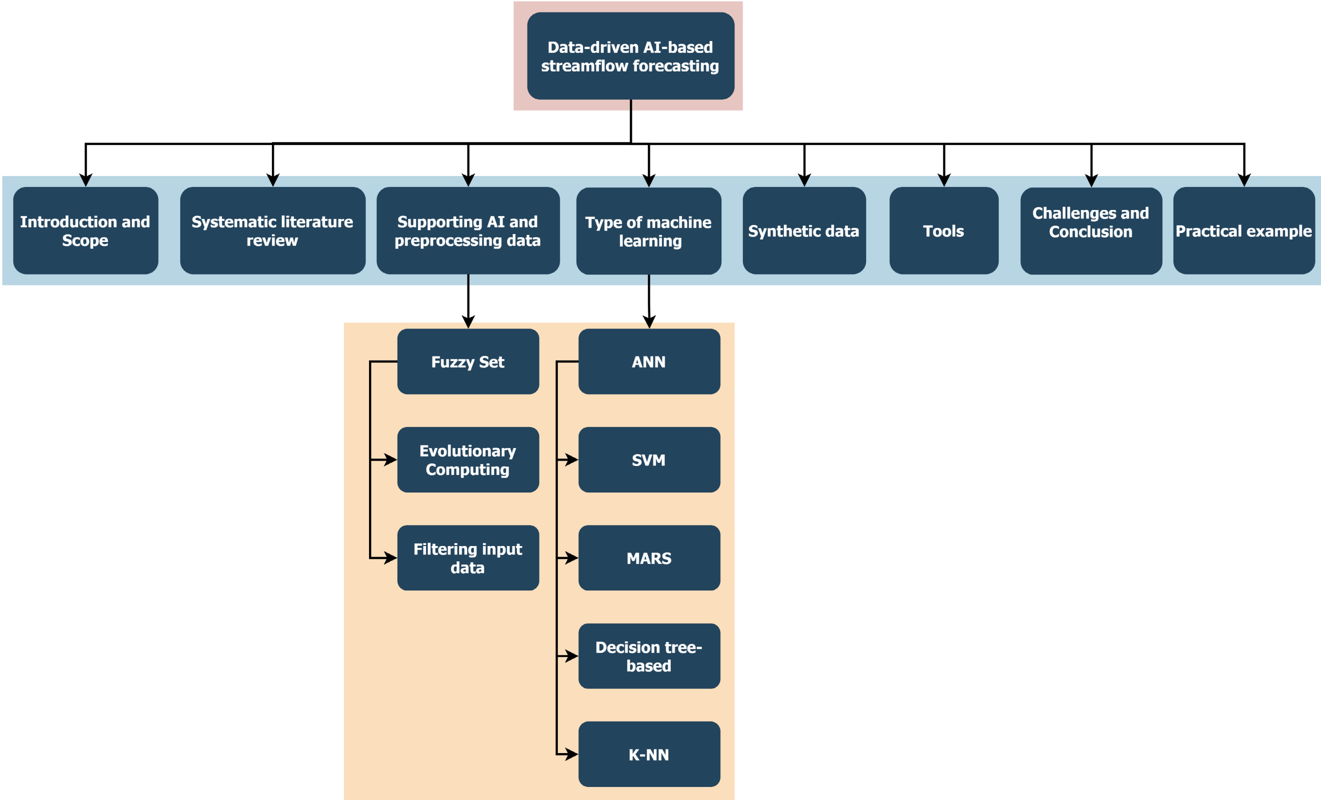

FIGURE 2 Flowchart of the structure of the paper. ANN, artificial neural network; K-NN, K-nearest neighbors; MARS, multivariate adaptive regression splines; SVM, support vector machine.

3.1 | Planning for SLR

Planning for the SLR in this paper requires following three steps: (1) formulating the research questions, (2) identification of research threads and inclusion, and (3) data source selection.

3.1.1 | Formulating the research questions

As mentioned in the Introduction, in hydrological forecasting, AI techniques have been utilized for many different areas, such as precipitation forecasting, evaporation forecasting, flood forecasting, and streamflow forecasting, which is the topic of this review paper. Hence, the main aim of this SLR is to investigate the recent (last 11 years) state of research in DDAI-based streamflow forecasting. Therefore, the high-level question to be answered in formulating the research question is what are AI-based methods and techniques for streamflow forecasting. This question can be broken down into the following subquestions:

1. What are the ML techniques for streamflow forecasting?
2. Can combining ML algorithms with AI-based techniques (hybrid model) improve the modeling results?
3. How to utilize synthetic data for streamflow forecasting?
4. What tools are available for DDAI-based streamflow forecasting?
5. What are the challenges and gaps in the literature?

3.1.2 | Identification of research strings, inclusions, and exclusions

By combining terms from the expertise of the topic and some of the most frequently cited articles on streamflow forecasting, we came up with the search strings. Initially, this review paper was limited to machine algorithms studies only. However, by reviewing further studies, we

AMERICAN WATER RESOURCES ASSOCIATION JAWRA JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

realized that other aspects of AI, such as FSs and evolutionary computing (EC), have shown significant enhancement to streamflow forecasting. Also, some industry practice knowledge has been less published, such as stochastic data generation and its application in streamflow forecasting. However, this paper excludes real-time flood forecasting focusing on the longer-term streamflow forecasting that can be used for water resources planning.

3.1.3 | Data source selection

The references for this review paper are sourced from: (a) high-rank journal papers available in online digital publishers like Springer Link, Scopus, Elsevier Science Direct, and Google Scholar; (b) proceedings of the repudiated conferences such as Hydrology and Water Resources Symposium (HWRS), International Congress on Environmental Modelling and Software Modelling and Simulation Society of Australia and New Zealand Inc. (MSSANZ); (c) scientific manual of DDAI-based tools; and (d) discussion with industrial experts. We found that the search engines of the most well-known scientific libraries behaved differently when the search string was specified during the definition of the search strings. Depending on the library, multiple methods are used to conduct the exact search (i.e., using different syntax). There were a variety of ways to search the content in each library. For instance, they allowed you to look for keywords in an article's title. In some cases, alternative options allow the exact keywords to be searched in abstracts, full text, or a mix of the three (e.g., title, abstract, and keywords).

3.2 | Conducting the SLR

For this review paper, all the recent (2011–2023) valuable studies and knowledge around DDAI-based streamflow forecasting limited to the scope of this paper were reviewed. Section 5 provides the outcome of the SLR for this review paper. The keywords were a combination of a few of the following words: “streamflow,” “forecasting,” “prediction,” “machine learning,” “artificial intelligence,” “model,” “statistical,” “ensemble,” “data-driven,” “re-view,” and “survey.” Different combinations were tested in different publishers' sites and it was noted that there was no significant change in the outcome. For better management of the references and to avoid problems like duplication, they were stored in the Mendeley Reference Manager.

3.3 | Reporting and analyzing

Figure 3 demonstrates the distribution of the journal article publications, and Table A1 (Appendix A) summarizes the most recent categorizing methods of DDAI-based streamflow forecasting. Also, the last section of this paper concludes all the reviewed papers and opens the identified challenges. In addition, suggestions were made for future research.

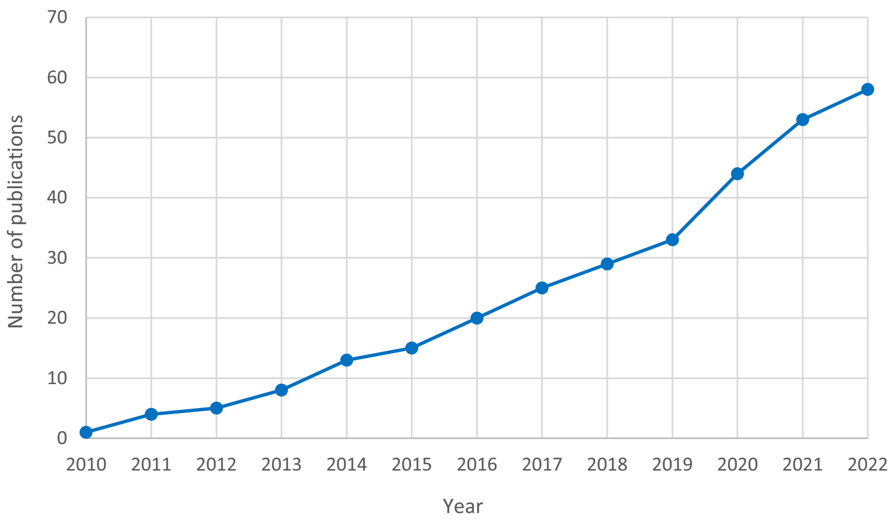

FIGURE 3 Cumulative number of publications related to the scope of this paper since 2010.

4 | SUPPORTING AI AND PREPROCESSING TECHNIQUES

As mentioned in the Introduction, for streamflow forecasting, ML plays the prominent role of regression (or classification) and other AI techniques (briefly explained in the subsections below) act as the supporting tool to improve the regression. In recent years, combining supportive tools and ML has been called hybrid models. Figure 4 summarizes supporting AI and prepossessing techniques for streamflow forecasting.

4.1 | Fuzzy logic

The theory of FL was introduced by Lotfi A. Zadeh in 1965 (Zadeh, 1965). It is a rule-based approach (if and then) and the main advantage of using FL in any kind of modeling is reducing ambiguity, and complexity and considering the uncertainty. This can be accomplished by either preferring the use of input data or considering the input data as interval rather than crisp point data. In streamflow forecasting, the FL approach theory has shown significant improvements when combined with ML algorithms relative to ML solely. In other words, FL plays an important supporting tool to the ML to make the prediction more accurate and provide better solutions to the complexity of the forecast. In the section on ML algorithms, after introducing each ML algorithm, the combination of those particular algorithms with FL will be explored and the advantages of such a combination will be provided.

4.2 | Evolutionary computing

EC is a general name for the approaches and techniques that are based on natural evolution (Sivakumar & Berndtsson, 2010). The EC algorithms are stimulated by the Darwinian theory of evolution, in which instead of solving the problem directly, it proceeds by evolving solutions to the problems (Yaseen et al., 2015). Some examples of EC (out of many) are genetic programming (GP) (Koza, 1992), genetic algorithms (GA) (Holland, 1975), evolutionary programming techniques, firefly optimization algorithm (FFA), bat algorithm (BA), particle swarm optimization (PSO), evolutionary strategies, swarm intelligence, gradient-based optimization (GBO), growing neural gas, and self-organizing maps. EC has

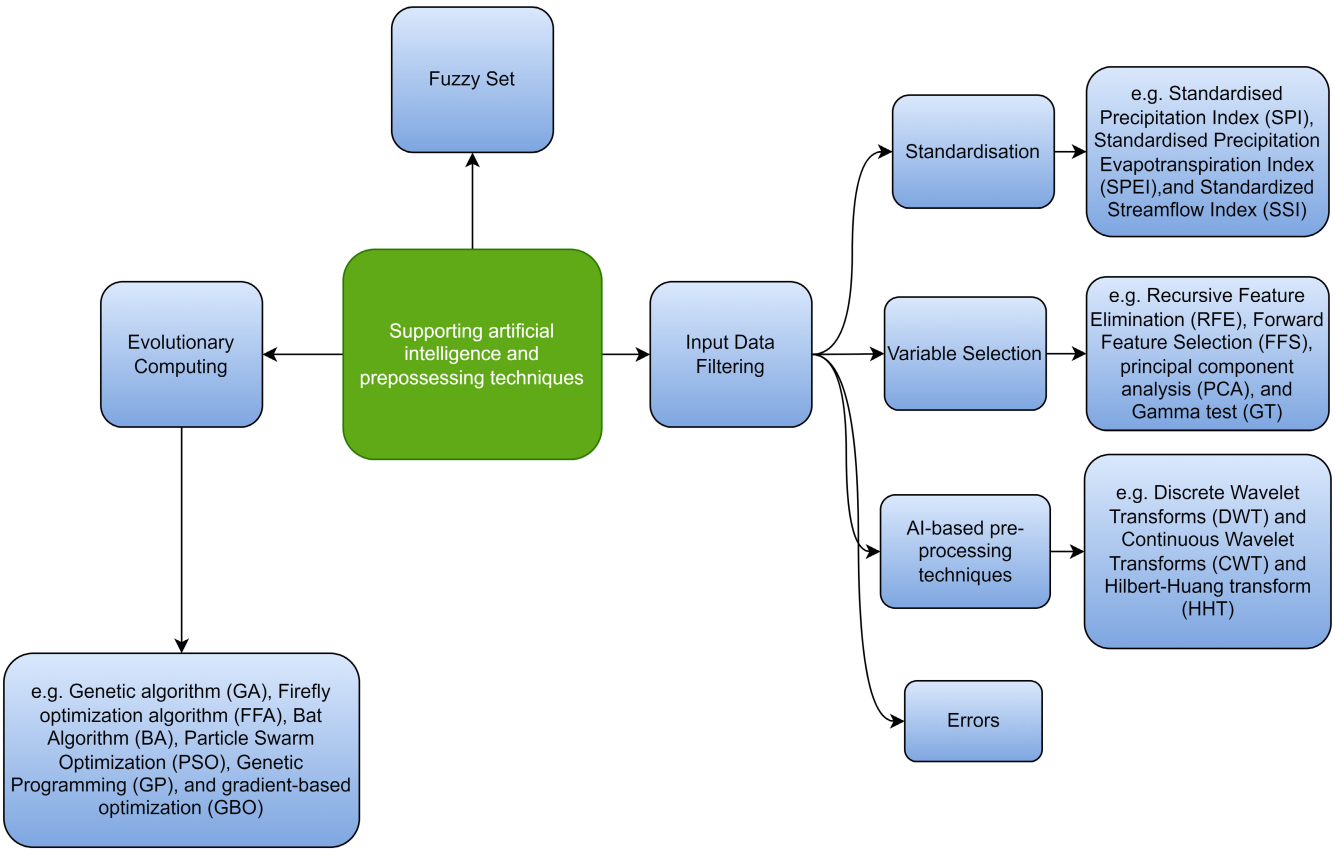

FIGURE 4 Summary of supporting AI and prepossessing techniques for streamflow forecasting.

been studied actively in the last couple of decades and the most common optimization techniques that have been used in hydrology and streamflow forecasting until now are GA, evolution strategies (Schwefel, 1981), GP, BA, PSO, GBO, and FFA. Similar to FL, EC works as a supporting tool to help MLs by optimizing the parameters and improving the streamflow forecast as well as direct applications in optimizing parameters of streamflow modeling and forecasting models. Details of how EC can combine with ML and FL are explained in the ML algorithms section.

4.3 | Filtering of input data

Hydrological modeling, especially data-driven streamflow modeling, normally requires a large dataset and different input variables. There could be errors in the measurement of the data, strong skewness, significant correlation between the variables, and nonstationary situations. Therefore, before doing any kind of modeling, including ML, the dataset should be screened to create the best input set that provides the most accurate outputs.

4.3.1 | Errors

The inputs for streamflow prediction usually are antecedent conditions (previous days or months of inflows), climate data (especially precipitation and temperature), and spatial data (landuse and land cover). The source of the data is the in situ measurement or remote sensing (or a mix of both), in which both are prone to errors including human errors, machine calibration, age of infrastructure, and damage (e.g., because of extreme weather: floods). One of the most common methods of finding errors is by looking at the time-series plot (visual identification). However, plots are good when the errors are outliers (too small or too big relative to other observations). Two easy ways of distinguishing error and outlier in the case of streamflow are by looking at rainfall nearby of the outlier event and also by checking streamflow and rainfall of nearby sites with similar hydrological characteristics. If errors are not outliers, they are not easily distinguishable. The systematic AI-based techniques for pre-processing of input data (for streamflow forecasting) have been explored in Section 5. However, there are techniques like prior data conflict (Egidi et al., 2022; Evans & Moshonov, 2006; Mutsvari et al., 2016) that provide systematic outlier exploration methods but have not been applied for streamflow forecasting purpose so far.

4.3.2 | Standardization and normalization

The following are a few reasons for the normalization or standardization of the input data, listed below:

- To tackle the non-error outliers.
- To avoid bias toward a few large/small events (Reis et al., 2021).
- Output has a different statistical distribution.
- Output is not homogeneous.
- Output needs to be on a particular scale.

There are several methods for standardization or normalization of data including min–max, decimal scaling, *z*-score, double sigmoid function, median and MAD, Box-Cox transformation, and tanh-estimators. Also, in hydrological prediction (including streamflow forecasting), there are three common applications of standardization: Standardized Precipitation Index (SPI) (Anshuka et al., 2019), Standardized Precipitation Evapotranspiration Index (SPEI), (Tirivarombo et al., 2018), and Standardized Streamflow Index (SSI) (Shamshirband et al., 2020).

4.3.3 | Variable selection

After all the filtering mentioned in previous sections, another option for improving the performance of the model is to find the best combination of the parameters and (if required) eliminate the less important ones. There are a few methods available for variable selection including recursive feature elimination (RFE), forward feature selection (FFS), principal component analysis (PCA), and gamma test (GT). One of the advantages of both RFE and FFS is they can work through ML runs. The RFE method selects all the variables then the less important ones are eliminated. The FFS works by adding one variable at a time to find the best combination (Reis et al., 2021). However, the need to reduce variables depends completely on the input data and method, for instance, the new iterative ensemble smoothers are scalable to millions of parameters (Chen & Oliver, 2012; Zhang et al., 2024). These methods have not been applied in streamflow forecasting.

4.3.4 | AI-based pre-processing techniques

In order to improve the streamflow forecasting models, input data should be analyzed and handled appropriately. One of the applications of input data pre-processing is recognizing and addressing the stationarity and nonstationarity conditions of the model. In a stationary model, parameters remain unchanged over time. Wavelet conjunction models (wavelet transform, WT) are AI-based time-series analyzers that have been commonly utilized in hydrology and streamflow forecasting. There are two types of WT: discrete wavelet transform (DWT) and continuous wavelet transform (CWT) wavelet signal process. Another (relatively new) method for data analysis is the Hilbert–Huang transform (HHT) (Huang & Wu, 2008). The base formula for CWT is as follows.

$$W_f(a,b)=\int_{-\infty}^{+\infty} f(t)\psi^*_{a,b}(t)dt, \tag{1}$$

where

$$\psi_{a,b}(t)=|a|^{-1/2}\psi(t-b/a)a, b\in R, a\neq 0, \tag{2}$$

where * represents the complex conjugate of the function, *a* and *b* are the parameters known as the scale and translation parameters, respectively, and *t* is the time. The coefficients of the CWT are calculated in each a (scale) and b (translation) and that leads to a full-time series on every scale which, hence, generates a large amount of data.

The DWT is defined by

$$W_f(j,k)=\int_{-\infty}^{+\infty} f(t)\psi^*_{jk}(t)dt, \tag{3}$$

$$\psi j,k(t)=a_0{}^{-j/2}\psi\left(a_0{}^{-j}t-b_0k\right). \tag{4}$$

5 | TYPES OF ML

There are a few ways of classifying ML approaches. One common method is based on the training process in which there are three types of ML: supervised learning, unsupervised learning, and reinforcement learning. Regression and clustering are two common learning types for supervised learning and clustering is one common type of learning for unsupervised learning. Regression learning works by finding a relationship between the dataset: input (independent) and output (dependent) variables. If the dataset has a timestamp, it is called time-series data. Prediction of time-series data is more complex and difficult in comparison to datasets without timestamps and time-series forecasting is one important area of ML (Hrnjica & Mehr, 2020) such as streamflow forecasting. Figure 5 demonstrates three types of ML, the learning process related to each type and common ML algorithms based on the learning process.

5.1 | Supervised learning

In supervised learning, the algorithm is provided with labeled training information including input and expected output. The goal is to learn from mistakes (through iteration in most of the algorithm) until the modeled output is close to the provided (known) output (Bonaccorso, 2018). The learning process (also can be called calibration) occurs in the mathematical function by adjusting the parameters. One of the most common problems of supervised learning is overt-fitting, when the system experiences a high performance with the input data by which it has been trained, but may not necessarily show good results when new data are introduced. In other words, the calibration is good but not the verification. Common techniques for supervised learning are regression, language processing, and image recognition. Section 7 describes the common ML algorithms applicable in hydrological modeling and prediction.

5.2 | Unsupervised learning

In unsupervised learning, the learner predicts based on unlabeled training data. Therefore, quantitative measures of performance can be difficult (Mohri et al., 2012). Clustering, pattern recognition, automatic labelling, and object segmentation are common methods of unsupervised learning.

AMERICAN WATER RESOURCES ASSOCIATION JAWRA JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

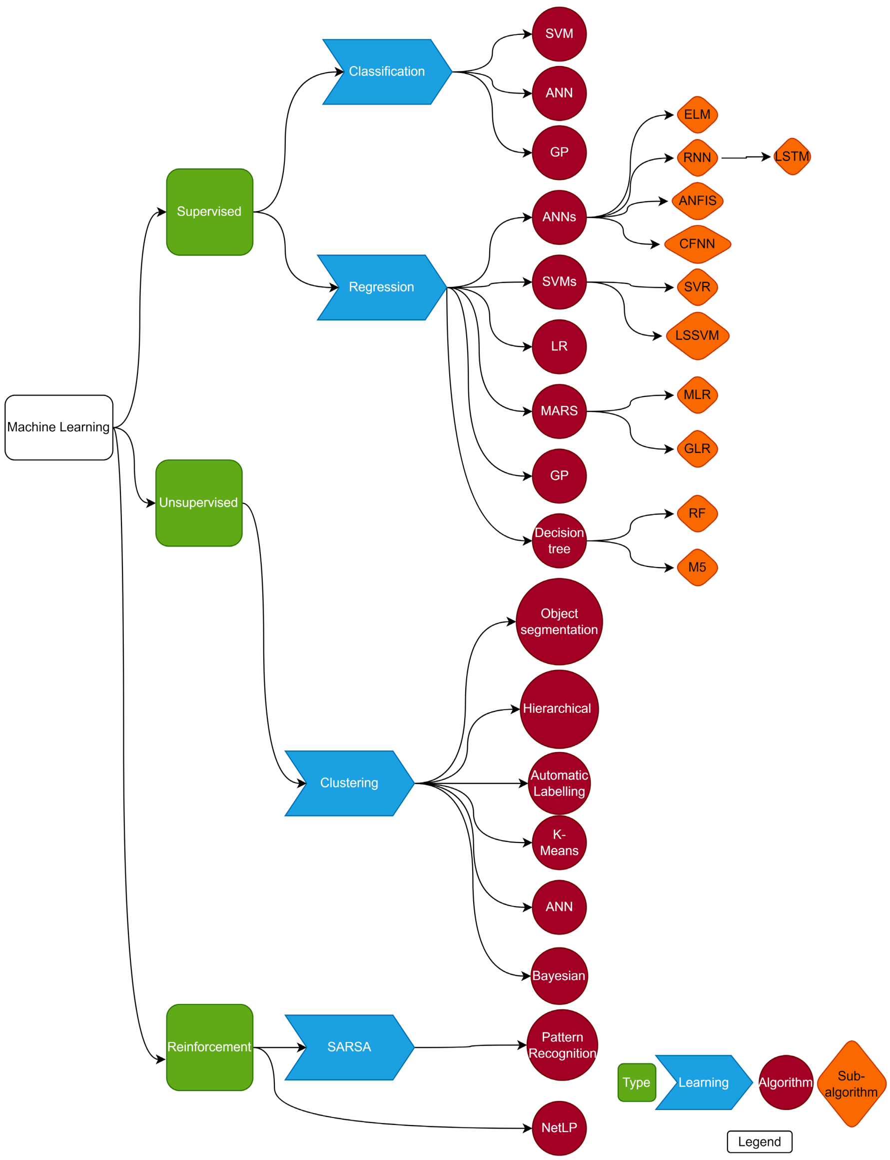

FIGURE 5 Machine learning types, learning approach and algorithms. ANFIS, adaptive neuro-fuzzy inference system; CFNN, cascade-forward neural networks; ELM, extreme learning machine; GLR, generalized likelihood ratio; GP, genetic programming; LR, linear regression; LSSVM, least square support vector machine; LSTM, long short-term memory; MLR, multiple linear regression; RF, random forest; RNN, recurrent neural network; SVR, support vector regression.

5.3 | Reinforcement learning

Reinforcement learning is based on the feedback provided by the environment. The feedback is in the form of reward and penalty and the goal is to collect as much reward as possible. However, the reward and penalty feedback may not be necessarily available and the learner must tradeoff between finding new information versus utilizing it (Mohri et al., 2012). One common example of reinforcement learning is deep neural network (Bonaccorso, 2018).

Reinforcement learning acts very similar to the way humans make decisions when there are multiple choices. For instance, one of the examples of water resources is the water transfer network. The water transfer network consists of nodes (e.g., interconnected reservoirs, supply/demand zones) and pipes. The transfer system should make decisions on sending water to which storage, transfer via which pipeline, when to spill, and so forth. The learner receives feedback from the environment in the form of award (incentive) and penalty (cost). For example, there is a high penalty for spilling and shortfall so the learner knows to avoid these as much as possible (until there is no other choice). There is an incentive on which reservoir to send water to and which pipeline to choose first or what is the demand priorities. These awards and penalties are sometimes set in a Network Linear Program (NetLP). Water resource modeling packages like REALM (REsource ALlocation Model, Perera, 2013) and eWater Source (Kelley et al., 2012) make use of NetLP for the water transfer/balance decisions.

6 | ML ALGORITHMS

Subsections below provide common ML algorithms in streamflow forecasting (Figure 5). Given the nature of streamflow forecasting, only supervised learning is applicable.

6.1 | Artificial neural network

An ANN is a mathematical model that tries to simulate the biological neural system and the human brain. The structure of ANN acts very similar to the human brain and nervous system and is able to create a nonlinear relation, which connects input and output data. ANN consists of simple processing elements (artificial neurons) and a high degree of interconnections (weights and bias) which work together to demonstrate complex behavior. Very similar to the human brain, it learns by examples and tries to improve the weight and bias to relate input and output, compared to the goal (target). There are three main applications of ANN: pattern recognition, data classification, and function approximation. In function approximation, by training the ANN, connections (weights) between the input and output are adjusted until the output is similar to the target. Because of the nonlinear correlation, ANN is a powerful implement for estimating complicated functions (Jahanbani et al., 2011; Jahanbani & El-Shafie, 2011) such as streamflow. Because of its relatively high speed and performance and ease of use, ANN-based approaches such as extreme learning machine (ELM) can be applied in complex nonlinear modeling such as streamflow forecasting using climate signals from climate change models (Zhu et al., 2019) or monthly streamflow forecasting (Akbarian et al., 2023). One of the most commonly used ANNs for hydrological purposes is feed-forward backpropagation (FFBP). It uses the Widrow–Hoff learning rule that the network weights move along a negative gradient of the performance function. However, depending on the applications and data type, in terms of performance, they sometimes sit second or third in the rank (Shortridge et al., 2016). One of the problems with ANN is that they are based on empirical risk minimization while other algorithms such as support vector machine (SVM) use structural risk minimization as well. As an example, Figure 6 shows the structure of ANN with five hidden layers and five neurons for forecasting Dec streamflow using 3-month antecedent flows.

6.1.1 | Recurrent neural network and long short-term memory

The traditional ANN with a feedforward network learning mechanism does not have any concept of dependency on the input data and the recurrent neural network (RNN) is designed to catch the sequential consequences when the previous information is required to be remembered for the future prediction. Therefore, RNN generally has two types of inputs: the current input which is the raw input fed to the model,

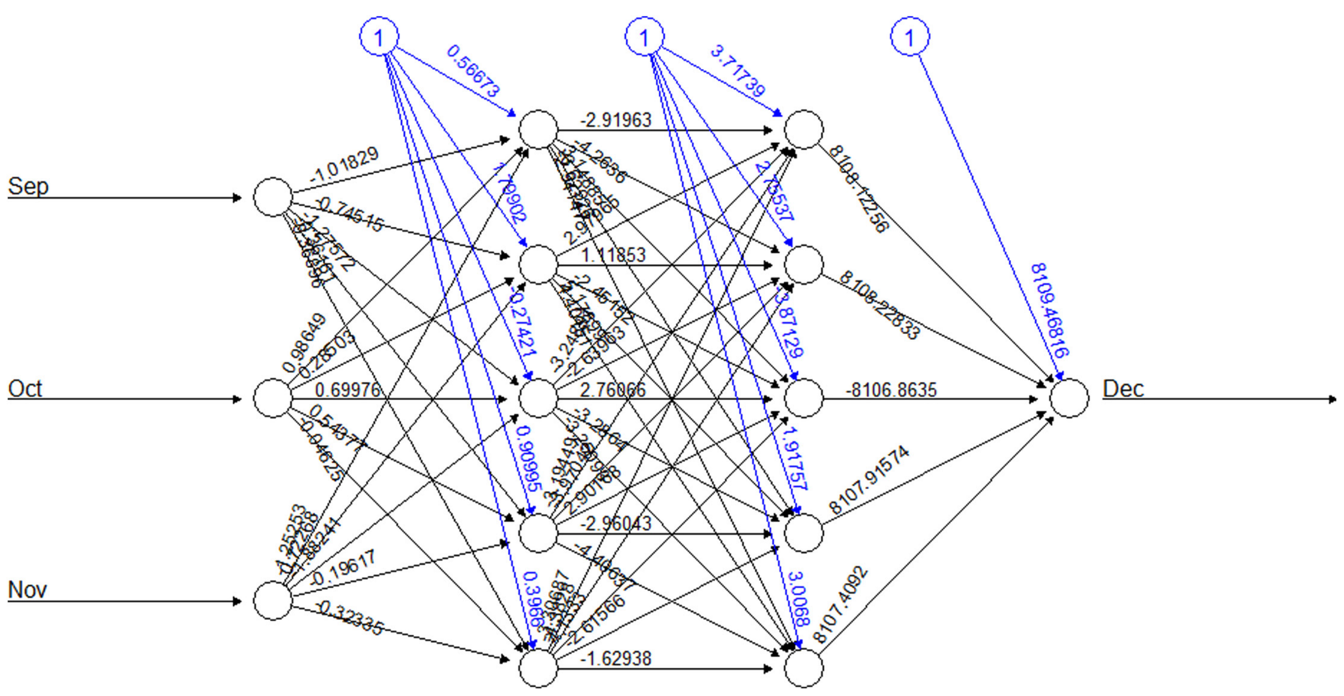

FIGURE 6 Structure of the ANN with five hidden layers and five neurons for forecasting Dec streamflow using Sep, Oct, and Nov flows.

and the input based on the data generated in the previous time step so-called backpropagation through time (BPTT). There are two types of issues with RNN: disappearance of the gradient and explosion of the gradient. The former is caused when the gradient is significantly small and weights are constant in each time step, which stops the network from further training. The latter is caused when the accumulation of updates of weight becomes very large and results in the explosion of the RNN (Hrnjica & Mehr, 2020). One solution to the abovementioned problems is the design of an RNN-based network called long short-term memory (LSTM). LSTM is a DL type of ML and generally, there are two main benefits of LSTM: the first one is allowing error flow through the network and the second one is introducing gates. Gates can control input flow, output flow, and storage of information. The forget gate makes the decision of which information to be stored and which one to be forgotten (Hrnjica & Mehr, 2020). The forget gate is very useful in time-series modeling in which it is required to capture time series and memorize long-term relationships. Therefore, LSTM performs very well in daily or monthly streamflow forecasting and can perform better than other ML approaches such as linear regression (LR), multi-layer perceptron (MLP), and SVM (Rahimzad et al., 2021).

6.1.2 | Adaptive neuro-fuzzy inference system

Adaptive neuro-fuzzy inference system (ANFIS) is a hybrid model that utilizes both fuzzy systems and neural networks. The FL provides a membership function for input data and the neural network acts as the learner. In such a nodes and rules structure, the rules can also relate the predictor to the predictand and nodes play as membership function (Figure 7). Because of such capability, ANFIS showed improved accuracy in streamflow forecasting in comparison to many other AI methods (Ashrafi et al., 2017; Sharma et al., 2015).

The basic formula and structure of ANFIS are briefly shown below:

$$\begin{aligned} &\mathrm{Rule\,1} = \mathrm{if}\,(x)\,\mathrm{is\,A1\,and}\,(y)\,\mathrm{is}\,(\mathrm{B1})\,\mathrm{then}\,(f1) = p1x + q1y = r1, \\ &\mathrm{Rule\,2} = \mathrm{if}\,(x)\,\mathrm{is\,A2\,and}\,(y)\,\mathrm{is}\,(\mathrm{B2})\,\mathrm{then}\,(f2) = p2x + q2y = r2, \end{aligned} \tag{5}$$

where $A1$, $B1$, $A2$, and $B2$ are membership functions, x and y are inputs, $f1$ and $f2$ are output functions, and $p1$, $q1$, $r1$, $p2$, $q2$, and $r2$ are linear parameter.

The following are some of the reasons why ANFIS improves streamflow forecasting:

1. Handling uncertainty and imprecision FL: ANFIS incorporates FL, which excels at handling uncertainty and imprecision inherent in streamflow data. It allows for the representation of linguistic terms like “low flow” or “high rainfall” using FSs, which better aligns with the natural ambiguity of hydrological processes. Membership Functions: ANFIS uses membership functions to model the degree to which input belongs to a FS, enabling a more nuanced understanding of the relationships between input variables and streamflow.

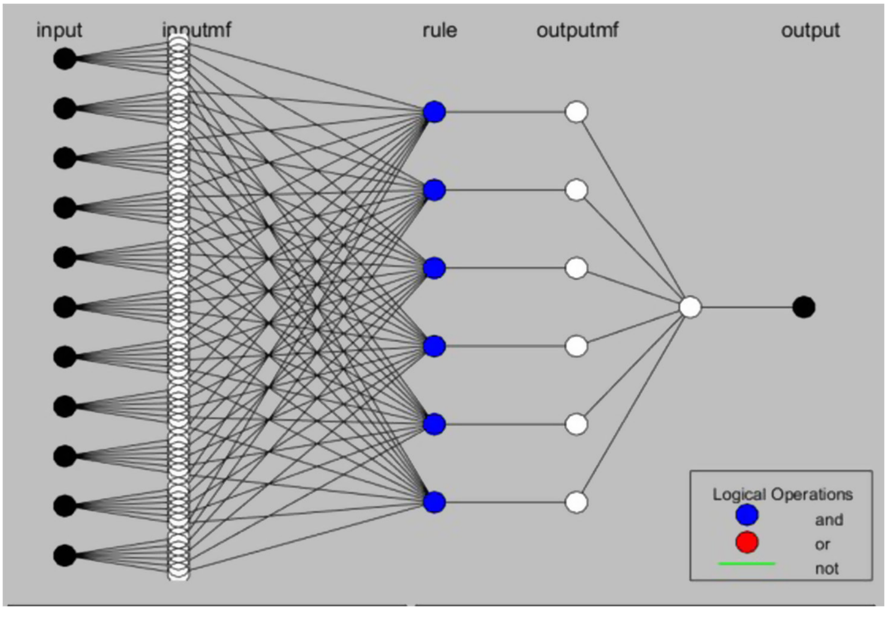

FIGURE 7 Structure of the adaptive neuro-fuzzy inference system model forecasting Dec flow based on Jan to Nov.

2. Incorporating expert knowledge: Rule-based structure: ANFIS utilizes a rule-based structure that can explicitly integrate expert knowledge about streamflow dynamics. This allows for capturing insights from domain experts and refining model behavior. Transparent reasoning: The rules within ANFIS provide a more transparent understanding of how inputs are mapped to outputs, facilitating model interpretation and potential adjustments based on expert judgment.
3. Adaptive learning: hybrid approach: ANFIS combines the learning capabilities of neural networks with the interpretability of FL. It can adapt its parameters through a hybrid learning process that involves both neural network training and fuzzy inference system optimization.
4. Fine-tuning: This adaptive nature enables ANFIS to fine-tune its behavior based on the specific characteristics of the streamflow data, leading to better model performance.
5. Robustness to noise: Noise handling: ANFIS generally exhibits greater robustness to noise and outliers in the data compared to ANNs. The fuzzy membership functions can effectively smooth out noise and reduce its impact on the model's predictions.
6. Handling nonlinear relationships: Nonlinear modeling: Both ANNs and ANFIS excel at capturing nonlinear relationships present in streamflow data. However, ANFIS often demonstrates a superior ability to handle complex nonlinearities due to its FL component.
7. Interpretability: Explainable predictions: While ANNs can be viewed as "black boxes" due to their opaque internal structure, ANFIS offers a more explainable model. The fuzzy rules and membership functions provide insights into how the model arrives at its predictions, enhancing understanding and trust in the results.
8. Computational efficiency: Training speed: ANFIS often requires less training data and computational resources compared to ANNs, making it more efficient to train and implement.

There are two types of FISs: Sugeno–Takagi FIS and Mamdani FIS. Sugeno–Takagi FIS is the most commonly used in streamflow forecasting (Yaseen et al., 2017).

6.1.3 | Cascade-forward neural networks

Cascade-forward neural networks (CFNNs) are constructed based on forward only type of counterpropagation network (CPN). CPN has two types of full and forward-only and was first introduced by Hecht-Nielsen (1987). Because of combinations of FS approach and cascade-forward structure, CFNN demonstrates better performance than feedforward neural network (FFNN) in streamflow forecasting by creating a strong relationship between river flow and climate data (Hayder et al., 2020).

6.1.4 | ANFIS and EC

In order to improve ANFIS, the MFs parameters must be optimally selected, which requires optimization algorithms. EC techniques can help for optimum selection of the parameters of the membership functions of FL. It was shown that ANFIS optimized by FFA can perform much better than ANFIS (alone) in monthly streamflow forecasting (Yaseen et al., 2017). In another study, ANFIS with BA demonstrated decreases in the error more than ANFIS with PSO, and ANFIS with GA in monthly streamflow forecasting using predictors of lagged climate indices (Ehteram et al., 2019). In a more recent study, for predicting monthly streamflow using climate data and antecedent flow data, standalone ANFIS was compared with hybrid ones including ANFIS with gradient-based optimizer (ANFIS-GBO), ANFIS with Grey Wolf optimizer (ANFIS-GWO), ANFIS-PSO, ANFIS with ant colony optimization (ANFIS-ACO), ANFIS-GA, and ANFIS with differential equation (ANFIS-DE). It was demonstrated that all the hybrid ones improved the accuracy of the prediction and the best performance was with ANFIS-GBO (Adnan, Mostafa, et al., 2021).

6.2 | Support vector machine

SVM is a supervised ML that is used for classification and regression and works by finding the optimal hyperplane. In other words, by mapping the inputs to high- and infinite-dimensional feature space. SVM, which was first proposed by Cortes & Vapnik (1995) works in the opposite direction of neural network (NN). While NNs learn from extensive experiment to theory, SVMs work from theory to application. One of the powers of the SVMs is that they generate the number and value of the learning parameters based on the model capacity and somehow take advantage of nonparametric models (Kecman, 2005). Hence, SVM-based models like support vector regression (SVR) and least squares support vector machine (LSSVM) optimize globally (vs. ANN which is local) and they have less risk of over-fitting (Lin et al., 2006). Also, in contrast to physical-based models, SVMs require fewer data and can perform better in real-time forecasting (Zhang et al., 2018). Researchers showed SVM can outperform other machine leanings in streamflow forecasting (He et al., 2014; Hipni et al., 2013; Noori et al., 2011).

The SVM can be used for linear separation (classification) or nonlinear separation. Also, it can be applied in the input space or feature space.

One of the main differences between SVMs (and other similar algorithms like SVR) is the kernel function that is used to transform data and solve nonlinear problems. The kernel function plays an important role in the performance of the SVMs. The most common kernel functions used in streamflow forecasting are the polynomial, the radial basis function (RBF), and the linear function.

6.2.1 | Support vector regression

SVR is a type of SVM that has got regression functionality. SVR can provide fast and easy high performance in prediction such as in streamflow forecasting using General Circulation Model outputs (Zhu et al., 2019) and can perform better than NN-based algorithms such as FFNN in prediction such as monthly streamflow using historical (observed) flow (Adnan et al., 2017). In a simple form, SVR does the regression using conditions as expressed in the Equations (6–9).

Equation of hyperplane

$$Y = w + b, \tag{6}$$

Equations of the decision boundary

$$wx + b = +a, \tag{7}$$

$$wx + b = -a, \tag{8}$$

Hyperplane of SVR should satisfy

$$-\boldsymbol{a} < \boldsymbol{Y} - \boldsymbol{wx} + \boldsymbol{b} < +\boldsymbol{a}. \tag{9}$$

6.2.2 | Least square support vector machine

Since one of the problems with SVMs is the huge computational burden, LSSVM provides more consistent solutions while avoiding some of the complexity of the traditional SVM by replacing linear regression with quadratic programming. Because of this strength, LSSVM has been successful in predicting streamflow with a large amount of input data such as using large-scale climate signals (Kisi et al., 2019).

6.2.3 | SVM and input filtering

A large amount of input data is among the common issues in streamflow forecasting. Therefore, filtering the input data to find the best combination of input variables is essential as discussed in the "Filtering of input data" section. For instance, (Noori et al., 2011) compared PCA, GT, and forward selection (FS) for reducing the input variable and finding the optimum combination for the purpose of monthly streamflow forecasting with a maximum of 3 months lead-time. They found out that prepossessing the input data using PCA improves SVM (PCA performs better than GT and FS) while they applied PCA on ANN and demonstrated that PCA-SVM performs better than PCA-ANN.

6.2.4 | SVM with EC

SVM is a powerful ML algorithm for classification and regression tasks. However, the performance of SVM models can be sensitive to the values of the hyperparameters, such as the cost parameter (C) and the kernel parameter (gamma). PSO and GA are two popular evolutionary optimization algorithms that can be used to optimize the hyperparameters of SVM models.

However, one problem with SVM-PSO is that the standard implementation of the SVM with PSO cannot handle two or more objective functions simultaneously. This is because SVMs are designed to solve binary classification problems, where the objective is to find a hyperplane that separates the data points into two distinct classes. When dealing with multiple objectives, the optimization process becomes more complex and requires different approaches. There are several strategies to address this limitation and utilize SVM with PSO for multi-objective optimization:

- Scalarization: This method combines multiple objectives into a single scalar objective function. This can be achieved using weighted sums, Pareto optimality dominance, or other aggregation techniques. PSO then optimizes this single objective function to find a solution that balances the tradeoffs between all objectives.

- Vector evaluated PSO (VEPSO): This approach modifies the PSO algorithm to handle multiple objectives directly. Each particle in the swarm maintains a vector of fitness values corresponding to each objective. The algorithm then updates particles based on their dominance and crowding distance in the objective space, promoting solutions that are both good and diverse.
- Multi-objective evolutionary algorithms (MOEAs): These algorithms are specifically designed for multi-objective optimization problems. They often use population-based approaches and selection mechanisms that favor solutions based on both their objective values and diversity. Popular MOEAs include NSGA-II, SPEA2, and MOEA/D.
- Hybridization with other algorithms: SVMs can be combined with other optimization algorithms like GAs or ant colony optimization that are inherently capable of handling multiple objectives. This allows for leveraging the strengths of both approaches to achieve better solutions for multi-objective problems.

Choosing the best approach depends on the specific problem and its characteristics, such as the number of objectives, the nature of the objectives, and the desired level of diversity in the solution set. It is important to carefully consider these factors and select the most suitable method for your specific multi-objective optimization task involving SVMs and PSO.

6.2.5 | SVM with WT and HHT

Another common problem in streamflow forecasting is the variability of the input climate or antecedent streamflow data to the SVM model. Techniques such as wavelet decomposition or HHT help divide the input time series into subcategories, which improves forecasting (Kambalimath & Deka, 2021) demonstrated that utilizing WT for pre-processing of input daily streamflow (dividing the daily historical streamflow time series to subcategories) will improve SVM in daily streamflow forecasting with lead-time 3–5 days relative to SVM with non-processed input data.

Therefore, SVM with the help of wavelet decomposition in this case, performs much better than SVM alone. Similarly, Kalteh (2013) coupled SVM and ANN with the DWT and shows SVM-DWT is more accurate than SVM alone and ANN-DWT. SVM can take advantage of HHT to decompose nonlinear, nonstationary input time series into subseries with meaningful time and frequency. Therefore, SVM with HHT can perform better than SVM with WT. Also, GA can be applied to optimize SVM parameters, and hence, SVM with HHT and GA perform much better than SVM with HHT under a strong nonstationary input time series (historical streamflow) to predict monthly streamflow up to 12-month lead-time (Meng et al., 2019).

6.3 | Multivariate adaptive regression splines

The multivariate adaptive regression splines (MARS) model creates a relationship between the input variable and output (e.g., streamflow) without any requirement of learning from the target since it uses antecedent condition (as the target) (Kisi et al., 2019). This method is particularly useful when there is not enough data available as the target or when there is a strong influence from the antecedent conditions on the prediction. Because of the above-mentioned advantage, MARS has shown achievement in rainfall-runoff modeling and streamflow prediction, in which antecedent conditions including climate data, landuse, vegetation cover, and streamflow strongly influence the streamflow prediction (Reis et al., 2021; Shortridge et al., 2016). Figure 8 shows MARS analysis of long-term monthly streamflows in predicting December flows using January to November antecedent flows using the R package earth. Multiple linear regression (MLR) is among other linear regression algorithms that have been used for streamflow forecasting. MLR alone does not have a great performance specifically in nonstationary conditions and can be improved when joined with pre-processing techniques such as WD or GA (Zhang et al., 2018).

6.4 | Decision tree-based models

Decision tree classification or predictive models have a structure similar to trees with branches. These models include a root node, internal node, and leaf node. Because of the nonparametric algorithm, they can deal with large and complicated datasets without bringing complicated parametric structures. Three common decision tree-based algorithms in streamflow forecasting are random forest (RF), M5 Model, and gradient boosting (GB).

6.4.1 | Random forest

RF is a supervised learning algorithm that is based on the decision tree. It combines tree predictors with ensemble learning in which more classifiers can solve complex problems (Breiman, 2001). RF is used for classification and regression and the binary recursive partitioning tree

AMERICAN WATER RESOURCES ASSOCIATION JAWRA JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

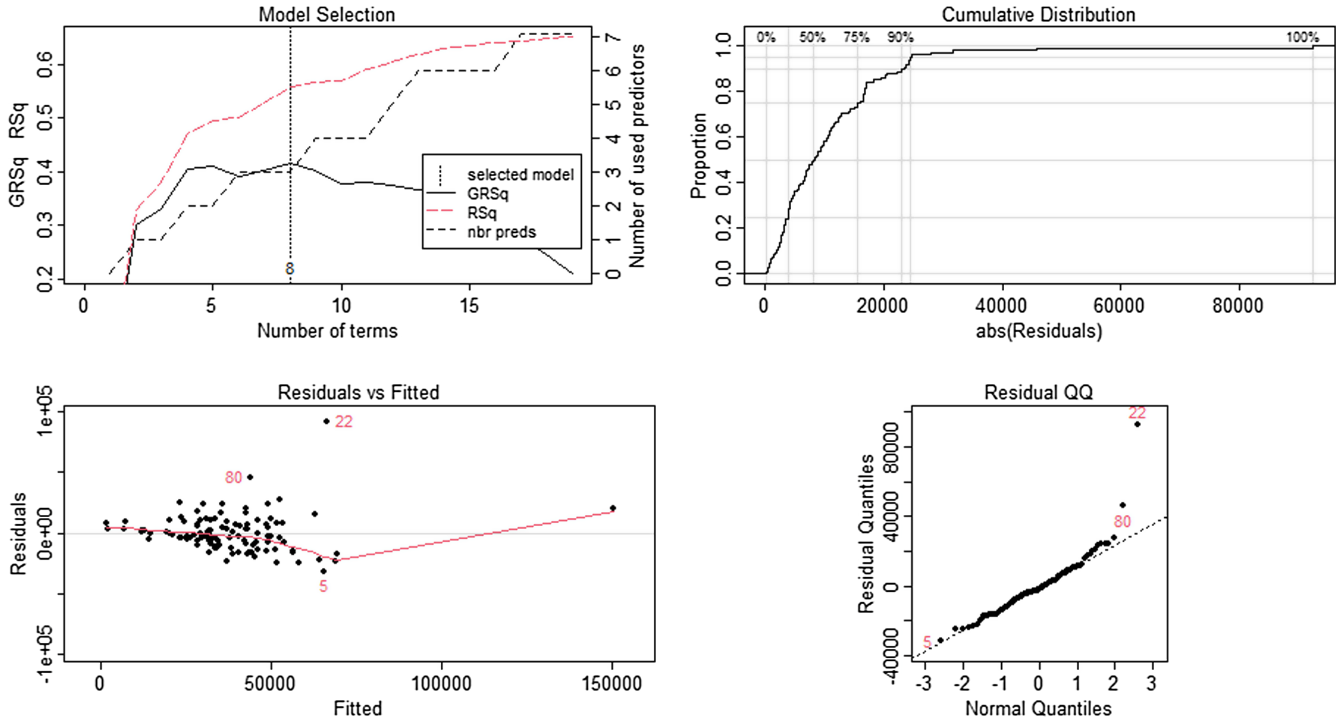

FIGURE 8 Multivariate adaptive regression splines analysis of monthly streamflow using R package "earth".

models, which are known as classification and regression tree (CART) models, make the foundation of the ensemble aspect of the methodology (Adnan, Zounemat-Kermani, et al., 2021).

Utilizing the RF algorithms (rule-based, nonparametric regression approach) showed high to low performances in different studies: high performance even in some cases better than ANN in predicting monthly streamflow using rainfall-runoff modeling approach (Shortridge et al., 2016), satisfactory performance for daily streamflow modeling (using different covariable) (Reis et al., 2021), mid-range performance relative to super learner regression and ANN and extremely randomized tree in daily streamflow forecasting (Tyralis et al., 2021), low range of performance compared to LSTM and ELM in monthly streamflow prediction using streamflow antecedent flow (Adnan, Zounemat-Kermani, et al., 2021; Akbarian et al., 2023).

6.4.2 | M5 model

M5 (or M5tree) model is another decision tree-based regression algorithm that tries to minimize error in each node. M5 has been used in streamflow forecasting and showed relatively high accuracy in monthly rainfall-runoff (Shortridge et al., 2016), lowest performance relative to LSSVM and MARS for monthly streamflow prediction using climate signals and antecedent flow (Kisi et al., 2019).

6.4.3 | Gradient boosting

GB is a powerful ensemble ML technique that sequentially builds a series of decision trees, each focused on correcting the errors of its predecessors. It is proven effective in streamflow forecasting due to its ability to model nonlinear relationships, handle complex interactions between variables, and adapt to diverse hydrological patterns. Advantages of GB for streamflow forecasting are it can handle nonlinear relationships, incorporates multiple variables, robust to noise and outliers, and adaptable to diverse patterns. GB can adapt to different hydrological regimes and seasonal variations, making it suitable for diverse streamflow forecasting applications (Ahmadianfar et al., 2023; Akbarian et al., 2023; Ni et al., 2020).

6.5 | K-nearest neighbors

K-NN is one of the most basic data mining methods and has been utilized for classification and regression based on the closeness of data to each other. In classification, K-NN returns the most common class label among neighbors. For example, text classification, image classification, etc. In regression, K-NN defines the average value among K's nearest neighbors. K-NN has been used for modeling and forecasting, including streamflow prediction. In streamflow forecasting, rather than the prediction another common application of K-NN is to join with other ML approaches (hybrid model) and error prediction of the models, which can boost the performance of the hybrid models (Ababaei et al., 2013; Akbari & Afshar, 2014).

7 | SYNTHETIC DATA AND STREAMFLOW FORECASTING

Stochastic data are synthetic and are randomly sampled from historical observed data. Stochastic data generation models can generate a long sequence of the required variable (e.g., streamflow) if they are calibrated (trained) well using observed data.

The use of observed historical data is proven to be inadequate for streamflow prediction due to the reasons such as (1) instrumental data does not necessarily capture the extreme conditions plausible to happen in the future since pre-instrumental extremes were more severe than what is observed in the instrumental record and (2) nonstationary relationship between climate and flow as a consequence of climate change, which causes using the climate data for runoff modeling prone to errors. On the other hand, there are uncertainties in climate change data created by GCMs. Therefore, syntactic data become required especially to tackle forecasting under climate change uncertainties (Kiem et al., 2021). Stochastic data have a few decades of history and one of the earliest works in hydrology is back in 1977 for predicting rainfall using an auto-regression algorithm (Dyer, 1977). The use of stochastic data in water resources/security models has become a standard practice of water resources/supply agencies in recent years since the stochastic data can preserve key statistics of the instrumental (recorded) inputs (e.g. rainfall, evaporation, flow) while covers the advantages briefed below. Figure 9 shows the range of possible flows created by stochastically generated data (red dot line and the blue line is the medium) versus observed data (black dot). The graph is generated by Wathnet (Kuczera, 1992, 2020). Subsection 10 (under Tools section) provides more details of the tool.

7.1 | Advantages

Stochastic data are a suitable replacement for recorded (instrumental) data because the generated data (if well-calibrated) can create long sequences of streamflow data, that is based on the observed data and in which extreme dry/wet sequences are captured. These extreme sequences do not necessarily occur in recorded data. Stochastically generated streamflow can tackle the problem of the non-stationary relationship between rainfall and streamflow. The nonstationary relationship is normally not considered in rainfall-runoff generated streamflow. The abovementioned advantages make the scholastic data more suitable for water resources assessments such as yield studies.

7.2 | Disadvantages

The data generation process is dependent on several parameters and fine-tuning these parameters is sometimes difficult. This may cause uncertainties on which way is better to follow.

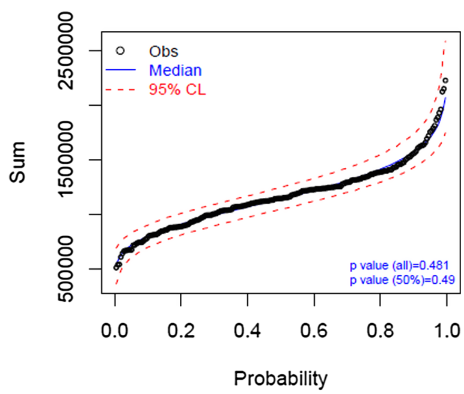

FIGURE 9 Comparison of range of possible flow by stochastic data versus observed flows (vertical axis flow mega liters) created by Wathnet5.

7.3 | Stochastic data generation

In the literature, a variety of alternative methods have been created and put to the test for generating stochastic data. These can be divided into two types in general:

1. Short to long time scale method: Methods for generating stochastic data at short time scales (daily or sub-daily) that aim to replicate the statistics for the short time scale at which the data are generated are known as short to long time scale methods. These methods typically rely on scaling relationships to reproduce the statistics at longer time scales (such as multiday, seasonal, and annual time scales).
2. Large to short time scale methods: first generate stochastic data at long time scales (seasonal, annual, or multi-year) using a parametric stochastic generator, and stochastically disaggregate the resulting data to shorter time scales (monthly, daily, or sub-daily) using a nonparametric approach.

The short to long time scale are often used for systems where the results are more strongly influenced by events that occur at shorter periods, such as floods or runoff events into systems with very little storage (e.g., household rainwater tanks). In contrast, long-to-short time scale approaches are frequently used in systems where results are determined by longer-term factors, such as water resources planning for metropolitan water supply systems, which can include extremely large multi-year reservoirs.

8 | PERFORMANCE FUNCTION

Performance functions are utilized for measuring the performance of the model by comparing the observed data with modeled results. Here is a list of the most common performance functions that have been used in streamflow forecasting.

- Coefficient of correlation (R^2).
- r: Pearson's correlation coefficient, Pearson (linear estimator) of correlation.
- Nash–Sutcliff efficiency (NSE): The NSE is calculated as one minus the ratio of the error variance of the modeled time series divided by the variance of the observed time series. The perfect match (zero error) has a NSE equal to 1.
- Kling–Gupta efficiency (KGE; Gupta et al., 2009): KGE, a normalized hydrologic metric, evaluates overall performance geared toward high flows (sensitive to outliers).
- Mean square error (MSE): MSE measures the average of the squares of the errors.
- Ratio of standard deviations (rSD): rSD indicates if flow variability is being over- or underestimated.
- Root mean square error (RMSE): RMSE (or root mean square deviation, RMSD) represents the square root of the second sample moment of the differences between predicted values.
- Normalized root mean square error (NRMSE): NRMSE normalizes the RMSE by relating the RMSE to the observed range of the variable.
- PBIAS: Percent bias indicates if total streamflow volume is being over- or underestimated.
- Taylor diagram (Taylor, 2001): It can be used to visually compare different performance functions and indicate which of them is more realistic.
- PBIAS_HF: Percent bias of flows equal or greater than Q98 (Yilmaz et al., 2008) characterizes response to large precipitation events.
- PBIAS_LF: Percent bias of flows equal or less than Q30 (Yilmaz et al., 2008) characterizes baseflow.

9 | TOOLS

Nowadays, Microsoft Excel is a common tool for mathematical and graphical usage. Excel has got several capabilities that satisfy most of the day to day needs. However, in cases like complex problems, problems that are repeated several times, and fast computations, Excel (even with Visual Basic programming) may not be enough or practical. Therefore, standalone packages or tools that are based on programming tools such as R and Python have been developed to help modelers. Below are some standalone tools or packages that have been used in streamflow forecasting.

9.1 | Stochastic data generators

9.1.1 | Stochastic Climate Library

With minor modifications in 2006 and 2007, the Stochastic Climate Library (SCL) software was initially made available in 2005. The eWater toolbox still offers SCL, which was first created by the Cooperative Research Centre for Catchment Hydrology. SCL (Srikanthan et al., 2007) contains the following options for stochastic models:

- Annual rainfall—first-order autoregressive model with parameter uncertainty.
- Monthly rainfall—modified method of fragments with annual data generated using the above annual rainfall model (long to short time scale).
- Daily rainfall—transition probability matrix (with Boughton's correction) (short to long time scale).
- Sub-daily rainfall—DRIP model (short to long time scale).
- Annual climate—first-order autoregressive multivariate model.
- Monthly climate—modified method of fragments (long to short time scale).
- Daily climate—first-order autoregressive multivariate model conditioned on rainfall state and nested in monthly and annual models (long to short time scale).
- Multi-site daily rainfall—multi-site two-part model nested in monthly and annual models (long to short time scale).

Srikanthan et al. (2007) and the references therein provide comprehensive descriptions for each of these models. Tools are available in the SCL user interface that enables the user to calibrate a stochastic data generation model and then produce stochastic replicates. The important statistics from the underlying (historical) data can be easily compared to the generated replicates using features in the SCL user interface. This capability is especially helpful for calibrating models. Also included in eWater Source software is the multi-site daily rainfall (long to short time scale) generator from SCL (Satheesh, 2017).

9.1.2 | MSSSCAR

The multi-site multi-season multi-state contemporaneous auto-regressive model (MSSSCAR) is a tool that is provided in WATHNET5 (Kuczera, 1992, 2020). These are the key features of MSSSCAR:

- Multi-site: It can be used in a system or catchment that contains numerous sites with data on temperature, streamflow, evaporation, and/or precipitation.
- Multi-season: It can be used to generate data for any number of seasons within a year or to generate data for a whole year (although between two and four seasons would be typical).
- Multi-state: Depending on one underlying climatic driver variable, such as the El Niño Southern Oscillation (ENSO) or the Interdecadal Pacific Oscillation, the yearly or seasonal generation model may (or may not) be conditioned (IPO).
- The annual/seasonal generation model maintains the autocorrelation structure of the annual or seasonal totals, that is, the degree to which a wet or dry season or year is followed by additional wet or dry seasons or years.
- Produces outputs at daily, monthly, seasonal, or annual time steps, by disaggregating the generated seasonal or annual totals using the method of fragments, with fragments selected using a kernel nearest neighbor approach.

In comparison to the models that are available in SCL, MSSSCAR offers many features that set it apart. In contrast to SCL's models, which only output data on an annual time-step at first, the multiseason feature in MSSSCAR may be especially helpful for systems where there are obvious shift between the wet and dry seasons. The stochastic data generation method may also be conditionally used with an underlying climate driver in the MSSSCAR's multi-state feature, which is not possible with any of the SCL models. The disaggregation strategy used in the majority of the models in SCL is far less tolerant of within-season (or within-year) variability in patterns than the kernel closest neighbor approach used in MSSSCAR.

9.1.3 | foreSIGHT package in R

The Systems Insights from Generation of Hydroclimatic Timeseries (foreSIGHT) package has been developed in the R statistical analysis system (Bennett et al., 2018). foreSIGHT uses the stochastic data generation technique from (Richardson, 1981), which generates daily rainfall occurrence and amount on a short to long period.

9.2 | ML tools

There are many tools and packages available for ML. Here is a list of the most common ones.

- KNIME is a standalone software that provides a powerful capability and flexibility for data analysis. KNIME provides a platform where R, Python, Java, and several other resources can be used and linked to each other simultaneously. Therefore, the common benefits of the available resources can be summoned in this tool.
- ANNdotNET (https://github.com/bhrnjica/anndotnet) is an open source project for DL written in Csharp.
- R packages:
- NNbenchmark—a package to evaluate neural network R packages.
- rminer—R package for data mining, classification and regression.
- ANN2—R package for ANN.
- nnet—R package for ANN.
- Neuralnet R package: Training of neural networks using backpropagation, resilient backpropagation with or without weight backtracking or the modified globally convergent version. The package allows flexible settings through custom-choice of error and activation function. Furthermore, the calculation of generalized weights is implemented. The package can be found from here https://cran.r-project.org/web/packages/neuralnet/index.html.
- Caret R package: The caret package, also known as Classification and REgression Training, is a collection of tools that aims to make the process of developing predictive models more efficient. The package includes tools for data splitting, pre-processing, feature selection, model tuning using resampling, variable importance estimation, as well as other functionality.
- CAST R package that offers the ability to use geographical or spatial–temporal data to run caret.
- earth—R package for Fast MARS.
- stats—R package for the MLR and GLM.
- RF—R package.
- mgcv—R package for GAM.
- Cubist—R package for M5 models.
- Python packages:
- NumPy: NumPy is the fundamental library for scientific computing in Python. It provides high-performance multi-dimensional arrays and a wide range of mathematical functions for efficient data manipulation and analysis.
- Scikit-learn: Scikit-learn is a comprehensive ML library that provides a wide range of algorithms for supervised and unsupervised learning, including regression, classification, clustering, and dimensionality reduction. Its user-friendly interface and extensive documentation make it a popular choice for beginners and experts alike.
- TensorFlow: TensorFlow is a powerful open-source library for developing and deploying DL models. It offers a flexible and customizable architecture that allows developers to build complex neural networks for various tasks, including image recognition, natural language processing, and time series forecasting.
- PyTorch: PyTorch is another popular DL library known for its dynamic computational graph and ease of debugging. It provides a more intuitive and Pythonic interface compared to TensorFlow, making it a favorite among some developers.
- Keras: Keras is a high-level DL library that sits on top of TensorFlow or PyTorch, providing a simpler and more user-friendly API for building and training neural networks. Its modular design makes it easy to experiment with different architectures and hyperparameters.
- XGBoost: XGBoost is a powerful gradient-boosting library that excels at regression and classification tasks. It is known for its accuracy, efficiency, and ability to handle large datasets.

10 | CHALLENGES AND CONCLUSION

1. This paper reviewed the most recent studies on applications of DDAI-based approaches in streamflow forecasting and summarized in Table A1. Nominating the best single or hybrid model of ML and other aspects of AI (FS, EC, and WT) is not possible and practical. Since, as discussed in each section of ML algorithms, the authors have claimed the best performance approaches based on the performance criteria they have used. Although there are common criteria, sometimes they have used different criteria. Even with common criteria such as R^2, RMSE, and NSE, the comparison is not possible because of the different combinations of input data, the purpose of forecasting and lead-time, time steps, and the number of cases (case studies as one or many streamflow sites). Also, it can reveal that the tuning of the parameters may be more important than the network and structure of the predictive model (Tabas et al., 2023).
2. This paper shows the recent advanced DDAI-based techniques can improve streamflow forecasting; however, given that historical records of climate/streamflow data are temporarily limited (e.g., limited to 100 years ago), the key challenges related to future streamflow and are listed below.
   - How to assess the performance of the long-term water resources systems in the severe conditions (such as droughts) that are more severe than those included in the historical record.

- How best to account for the impact of climate change in assessing the adequacy of the supply to meet current and future needs.
- How to cost-effectively manage the water supply system to ensure the security of supply (even in the most severe drought).

There is only one paper (Kiem et al., 2021) that started to tackle this issue by using the future projection of temperature, as the most certain output of GCMs, to filter stochastic data and create a future projection of streamflow data under climate change. However, there is room for more improvements in their method such as the application of AI techniques and testing more climate variables such as maximum or minimum temperature. Although precipitation is among the less certain climate variables out of the GCMs, the AI techniques reviewed in this paper can be used to make use of it with more reliability.

3. Raw data are always a challenge in streamflow forecasting because of issues such as the availability of data within the required range that is suitable for calibration, missing observed data, outliers, and misleading noise. This paper reviewed a number of pre-processing methods such as standardization, normalization, DWT, and CWT wavelet signal processes to improve the raw data and shape them to better fit for modeling. However, there is room for the allocation of other techniques.

AUTHOR CONTRIBUTIONS

Khandakar Ahmed: Supervision. **Bruce Gu:** Supervision.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The tool and data of Appendix B are openly available through a public GitHub link.

ORCID

Heerbod Jahanbani https://orcid.org/0000-0003-1599-9147

REFERENCES

Ababaei, B., F. Mirzaei, T. Sohrabi, and S. Araghinejad. 2013. “Reservoir Daily Inflow Simulation Using Data Fusion Method.” *Irrigation and Drainage* 62(4): 468–76.

Adnan, R.M., R.R. Mostafa, A. Elbeltagi, Z.M. Yaseen, S. Shahid, and O. Kisi. 2021. “Development of New Machine Learning Model for Streamflow Prediction: Case Studies in Pakistan.” *Stochastic Environmental Research and Risk Assessment* 36: 999–1033.

Adnan, R.M., X. Yuan, O. Kisi, and Y. Yuan. 2017. “Streamflow Forecasting Using Artificial Neural Network and Support Vector Machine Models.” *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)* 29(1): 286–94.

Adnan, R.M., M. Zounemat-Kermani, A. Kuriqi, and O. Kisi. 2021. *Machine Learning Method in Prediction Streamflow Considering Periodicity Component*. Singapore: Springer.

Ahmadianfar, I., V. Demir, S. Heddam, A. Al-Areeq, S. Abba, M.L. Tan, B. Halder, H. Marhoon, and Z. Yaseen. 2023. “Daily Scale Streamflow Forecasting Based-Hybrid Gradient Boosting Machine Learning Model.” https://doi.org/10.21203/rs.3.rs-2486952/v1.

Ahmed, K., D.A. Sachindra, S. Shahid, Z. Iqbal, N. Nawaz, and N. Khan. 2020. “Multi-Model Ensemble Predictions of Precipitation and Temperature Using Machine Learning Algorithms.” *Atmospheric Research* 236: 104806.

Akbari, M., and A. Afshar. 2014. “Similarity-Based Error Prediction Approach for Real-Time Inflow Forecasting.” *Hydrology Research* 45(4–5): 589–602.

Akbarian, M., B. Saghafian, and S. Golian. 2023. “Monthly Streamflow Forecasting by Machine Learning Methods Using Dynamic Weather Prediction Model Outputs over Iran.” *Journal of Hydrology* 620: 129480.

Anshuka, A., F.F. van Ogtrop, and R.W. Vervoort. 2019. “Drought Forecasting through Statistical Models Using Standardised Precipitation Index: A Systematic Review and Meta-Regression Analysis.” *Natural Hazards* 97: 955–77.

Ashrafi, M., L.H.C. Chua, C. Quek, and X. Qin. 2017. “A Fully-Online Neuro-Fuzzy Model for Flow Forecasting in Basins with Limited Data.” *Journal of Hydrology* 545: 424–35.

Bahram Saghafian, A.H., and M. Dehghani. 2017. “Effect of Enso on Annual Maximum Floods and Volume over Threshold in the Southwestern Region of Iran.” *Hydrological Sciences Journal* 62(7): 1039–49.

Belayneh, A., J. Adamowski, B. Khalil, and B. Ozga-Zielinski. 2014. “Long-Term SPI Drought Forecasting in the Awash River Basin in Ethiopia Using Wavelet Neural Network and Wavelet Support Vector Regression Models.” *Journal of Hydrology* 508: 418–29.

Belayneh, A., J. Adamowski, B. Khalil, and J. Quilty. 2016. “Coupling Machine Learning Methods with Wavelet Transforms and the Bootstrap and Boosting Ensemble Approaches for Drought Prediction.” *Atmospheric Research* 172–173: 37–47.

Bennett, B., S. Culley, S. Westra, and H.R. Maier. 2018. “An R tool for scenario-neutral climateimpact analysis of water resource systems.” International Congress on Environmental Modellingand Software. 22.

Bonaccorso, G. 2018. *Data Science and Machine Learning*, 2nd ed. Birmingham, UK: Packt Publishing, Limited.

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45(1): 5–32.

Chang, F.J., P.A. Chen, Y.R. Lu, E. Huang, and K.Y. Chang. 2014. “Real-Time Multi-Step-Ahead Water Level Forecasting by Recurrent Neural Networks for Urban Flood Control.” *Journal of Hydrology* 517: 836–46.

Chen, Y., and D.S. Oliver. 2012. “Ensemble Randomized Maximum Likelihood Method as an Iterative Ensemble Smoother.” *Mathematical Geosciences* 44: 1–26.

Cortes, C., and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20(3): 273–97.

Dehghani, M., B. Saghafian, F. Nasiri Saleh, A. Farokhnia, and R. Noori. 2014. "Uncertainty Analysis of Streamflow Drought Forecast Using Artificial Neural Networks and Monte-Carlo Simulation." *International Journal of Climatology* 34(4): 1169–80.

Dyer, T.G.J. 1977. "On the Application of Some Stochastic Models to Precipitation Forecasting." *Quarterly Journal of the Royal Meteorological Society* 103(435): 177–89.

Egidi, L., F. Pauli, and N. Torelli. 2022. "Avoiding Prior–Data Conflict in Regression Models via Mixture Priors." *Canadian Journal of Statistics* 50: 491–510.

Ehteram, M., H.A. Afan, M. Dianatikhah, A.N. Ahmed, C.M. Fai, M.S. Hossain, M.F. Allawi, and A. Elshafie. 2019. "Assessing the Predictability of an Improved ANFIS Model for Monthly Streamflow Using Lagged Climate Indices as Predictors." *Water* 11(6): 1130.

Evans, M., and H. Moshonov. 2006. "Checking for Prior-Data Conflict." *Bayesian Analysis* 1(4): 893–914.

Ganguli, P., and M.J. Reddy. 2014. "Ensemble Prediction of Regional Droughts Using Climate Inputs and the SVM–Copula Approach." *Hydrological Processes* 28(19): 4989–5009.

Gizaw, M.S., and T.Y. Gan. 2016. "Regional Flood Frequency Analysis Using Support Vector Regression under Historical and Future Climate." *Journal of Hydrology* 538: 387–98.

Gupta, H.V., H. Kling, K.K. Yilmaz, and G.F. Martinez. 2009. "Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling." *Journal of Hydrology* 377: 80–91.

Hayder, G., M.I. Solihin, and H.M. Mustafa. 2020. "Modelling of River Flow Using Particle Swarm Optimized Cascade-Forward Neural Networks: A Case Study of Kelantan River in Malaysia." *Applied Sciences* 10: 8670. https://doi.org/10.3390/app10238670.

He, Z., X. Wen, H. Liu, and J. Du. 2014. "A Comparative Study of Artificial Neural Network, Adaptive Neuro Fuzzy Inference System and Support Vector Machine for Forecasting River Flow in the Semiarid Mountain Region." *Journal of Hydrology* 509: 379–86.

Hecht-Nielsen, R. 1987. "Counterpropagation Networks." *Applied Optics* 26(23): 4979–84.

Hipni, A., A. El-shafie, A. Najah, O.A. Karim, A. Hussain, and M. Mukhlisin. 2013. "Daily Forecasting of Dam Water Levels: Comparing a Support Vector Machine (SVM) Model with Adaptive Neuro Fuzzy Inference System (ANFIS)." *Water Resources Management* 27(10): 3803–23.

Holland, J.H. 1975. *Adaptation in Natural and Artificial Systems*. Cambridge, MA: University of Michigan Press.

Hrnjica, B., and A.D. Mehr. 2020. "Energy Demand Forecasting Using Deep Learning." In *Smart Cities Performability, Cognition, & Security. EAI/Springer Innovations in Communication and Computing*, edited by F. Al-Turjman. Cham: Springer. https://doi.org/10.1007/978-3-030-14718-1_4.

Huang, N.E., and Z. Wu. 2008. "A review on Hilbert-Huang transform: Method and its applications to geophysical studies." *Reviews of Geophysics* 46: RG2006. https://doi.org/10.1029/2007RG000228.

Ian, G., B. Yoshua, and C. Aaron. 2016. *Deep Learning Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press.

Ibrahim, K.S.M.H., Y.F. Huang, A.N. Ahmed, C.H. Koo, and A. El-Shafie. 2022. "A Review of the Hybrid Artificial Intelligence and Optimization Modelling of Hydrological Streamflow Forecasting." *Alexandria Engineering Journal* 61(1): 279–303.

Jahanbani, H., and A.H. El-Shafie. 2011. "Application of Artificial Neural Network in Estimating Monthly Time Series Reference Evapotranspiration with Minimum and Maximum Temperatures." *Paddy and Water Environment* 9(2): 207–20.

Jahanbani, H., L.T. Shui, A. Massah Bavani, and A.H. Ghazali. 2011. "Uncertainty of Climate Change and Its Impact on Reference Evapotranspiration in Rasht City, Iran." *Journal of Water and Climate Change* 2(1): 72–83.

Jimeno-Sáez, P., J. Senent-Aparicio, J. Pérez-Sánchez, D. Pulido-Velazquez, and J.M. Cecilia. 2017. "Estimation of Instantaneous Peak Flow Using Machine-Learning Models and Empirical Formula in Peninsular Spain." *Water* 9(5): 347. https://doi.org/10.3390/w9050347.

Kalteh, A.M. 2013. "Monthly River Flow Forecasting Using Artificial Neural Network and Support Vector Regression Models Coupled with Wavelet Transform." *Computers and Geosciences* 54: 1–8.

Kambalimath, S., and P.C. Deka. 2021. "Performance Enhancement of SVM Model Using Discrete Wavelet Transform for Daily Streamflow Forecasting." *Environmental Earth Sciences* 80(3): 101.

Kecman, V. 2005. *Support Vector Machines—An Introduction*. Berlin, Heidelberg: Springer.

Kelley, P., A. O'Brien, and eWater CRC (Australia). 2012. *eWater Source: Scientific Reference Guide*. Canbera, Australia: eWater Cooperative Research Centre.

Khan, N., D.A. Sachindra, S. Shahid, K. Ahmed, M.S. Shiru, and N. Nawaz. 2020. "Prediction of Droughts over Pakistan Using Machine Learning Algorithms." *Advances in Water Resources* 139: 103562.

Khosravi, K., A. Golkarian, and J.P. Tiefenbacher. 2022. "Using Optimized Deep Learning to Predict Daily Streamflow: A Comparison to Common Machine Learning Algorithms." *Water Resources Management* 36: 699–716.

Kiem, A.S., G. Kuczera, P. Kozarovski, L. Zhang, and G. Willgoose. 2021. "Stochastic generation of future hydroclimate using temperature as a climate change covariate." *Water Resources Research* 57: 2020WR027331.

Kisi, O., B. Choubin, R.C. Deo, and Z.M. Yaseen. 2019. "Incorporating Synoptic-Scale Climate Signals for Streamflow Modelling over the Mediterranean Region Using Machine Learning Models." *Hydrological Sciences Journal* 64(10): 1240–52.

Kitchenham, B. 2004. *Procedures for Performing Systematic Reviews*, Vol 33. Keele, UK: Keele University.

Koza, J.R. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Vol 1. Cambridge, MA: MIT Press.

Kuczera, G. 1992. "Water Supply Headworks Simulation Using Network Linear Programming." *Advances in Engineering Software* 14(1): 55–60.

Kuczera, G. 2020. "Multi-Site Multi-Season Multi-State Contemporaneous Auto-Regressive Model (MSSSCAR)." In *Wathnet5 User Manual*. NSW, Australia: WMAWater.

Lian, Y., J. Luo, W. Xue, G. Zuo, and S. Zhang. 2022. "Cause-Driven Streamflow Forecasting Framework Based on Linear Correlation Reconstruction and Long Short-Term Memory." *Water Resources Management* 36: 1661–78.

Lin, J.Y., C.T. Cheng, and K.W. Chau. 2006. "Using Support Vector Machines for Long-Term Discharge Prediction." *Hydrological Sciences Journal* 51(4): 599–612.

Liu, J., X. Yuan, J. Zeng, Y. Jiao, Y. Li, L. Zhong, and L. Yao. 2022. "Ensemble Streamflow Forecasting over a Cascade Reservoir Catchment with Integrated Hydrometeorological Modeling and Machine Learning." *Hydrology and Earth System Sciences* 26: 265–78.

Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, et al. (Eds.), 2021. *IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution OF Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK, Newyork, USA: Cambridge University Press. In Press. https://doi.org/10.1017/9781009157896

Meng, E., S. Huang, Q. Huang, W. Fang, L. Wu, and L. Wang. 2019. "A Robust Method for Non-Stationary Streamflow Prediction Based on Improved EMD-SVM Model." *Journal of Hydrology* 568: 462–78.

Mohamadi, S., S.S. Sammen, F. Panahi, M. Ehteram, O. Kisi, A. Mosavi, A.N. Ahmed, A. El-Shafie, and N. Al-Ansari. 2020. "Zoning Map for Drought Prediction Using Integrated Machine Learning Models with a Nomadic People Optimization Algorithm." *Natural Hazards* 104(1): 537–79.

Mohammed, M., M.B. Khan, and E.B.M. Bashier. 2016. *Machine Learning: Algorithms and Applications*. Milton, UK: Taylor & Francis Group. http://ebookcentral.proquest.com/lib/vu/detail.action?docID=4683300.

Mohri, M., A. Rostamizadeh, and A. Talwalkar. 2012. *Foundations of Machine Learning. Adaptive Computation and Machine Learning Series*. Cambridge, MA: The MIT Press. https://search.ebscohost.com/login.aspx?direct=trueAuthType=shibdb=nlebkAN=478737site=ehost-livecustid=s1145751.

Mutsvari, T., D. Tytgat, and R. Walley. 2016. "Addressing Potential Prior-Data Conflict when Using Informative Priors in Proof-of-Concept Studies." *Pharmaceutical Statistics* 15: 28–36.

Ni, L., D. Wang, J. Wu, Y. Wang, Y. Tao, J. Zhang, and J. Liu. 2020. "Streamflow Forecasting Using Extreme Gradient Boosting Model Coupled with Gaussian Mixture Model." *Journal of Hydrology* 586: 124901.

Noori, R., A.R. Karbassi, A. Moghaddamnia, D. Han, M.H. Zokaei-Ashtiani, A. Farokhnia, and M.G. Gousheh. 2011. "Assessment of Input Variables Determination on the SVM Model Performance Using PCA, Gamma Test, and Forward Selection Techniques for Monthly Stream Flow Prediction." *Journal of Hydrology* 401(3–4): 177–89.

Perera, B.J.C. 2013. "Water Supply Planning and Simulation Software Package-Realm." In *Proceedings of International Forestry and Environment Symposium*. Nugegoda, Sri Lanka: Department of Forestry and Environmental Science University of Sri Jayewardenepura.

Phan, T.T.H., and X.H. Nguyen. 2020. "Combining Statistical Machine Learning Models with ARIMA for Water Level Forecasting: The Case of the Red River." *Advances in Water Resources* 142: 103656.

Rahimzad, M., A. Moghaddam Nia, H. Zolfonoon, J. Soltani, A. Danandeh Mehr, and H.-H. Kwon. 2021. "Performance Comparison of an LSTM-Based Deep Learning Model versus Conventional Machine Learning Algorithms for Streamflow Forecasting." *Water Resources Management* 35: 4167–87.

Reis, G.B., D.D. da Silva, E.I. Fernandes Filho, M.C. Moreira, G.V. Veloso, M.S. Fraga, and S.A.R. Pinheiro. 2021. "Effect of environmental covariable selection in the hydrological modeling using machine learning models to predict daily streamflow." *Journal of Environmental Management* 290: 112625. https://doi.org/10.1016/j.jenvman.2021.112625.

Richardson, C.W. 1981. "Stochastic Simulation of Daily Precipitation, Temperature, and Solar Radiation." *Water Resources Research* 17: 182–90.

Sachindra, D.A., K. Ahmed, M.M. Rashid, S. Shahid, and B.J. Perera. 2018. "Statistical Downscaling of Precipitation Using Machine Learning Techniques." *Atmospheric Research* 212: 240–58.

Sapitang, M., W.M. Ridwan, K.F. Kushiar, A.N. Ahmed, and A. El-Shafie. 2020. "Machine Learning Application in Reservoir Water Level Forecasting for Sustainable Hydropower Generation Strategy." *Sustainability* 12(15): 6121.

Satheesh, A. 2017. Stochastic Analysis Tool. In eWater Source User Guide 5.0. https://wiki.ewater.org.au/display/SD50/Stochastic+Analysis+Tool.

Schwefel, H.-P. 1981. *Numerical Optimization of Computer Models*. Chichester: John Wiley & Sons, Inc.

Shamshirband, S., S. Hashemi, H. Salimi, S. Samadianfard, E. Asadi, S. Shadkani, K. Kargar, A. Mosavi, N. Nabipour, and K.-W. Chau. 2020. "Predicting Standardized Streamflow Index for Hydrological Drought Using Machine Learning Models." *Engineering Applications of Computational Fluid Mechanics* 14: 339–50.

Sharma, S., P. Srivastava, X. Fang, and L. Kalin. 2015. "Performance Comparison of Adoptive Neuro Fuzzy Inference System (ANFIS) with Loading Simulation Program C++ (LSPC) Model for Streamflow Simulation in El Niño Southern Oscillation (ENSO)-Affected Watershed." *Expert Systems with Applications* 42(4): 2213–23.

Shortridge, J.E., S.D. Guikema, and B.F. Zaitchik. 2016. "Machine Learning Methods for Empirical Streamflow Simulation: A Comparison of Model Accuracy, Interpretability, and Uncertainty in Seasonal Watersheds." *Hydrology and Earth System Sciences* 20(7): 2611–28.

Sit, M., B.Z. Demiray, Z. Xiang, G.J. Ewing, Y. Sermet, and I. Demir. 2020. "A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources." *Water Science and Technology* 82(12): 2635–70.

Sivakumar, B., and R. Berndtsson. 2010. *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*. Singapore: World Scientific Publishing Company. http://ebookcentral.proquest.com/lib/vu/detail.action?docID=731125.

Srikanthan, R., F. Chiew, and A. Frost. 2007. *SCL Stochastic Climate Library, User Manual*. Melbourne, Australia: CSIRO.

Tabas, S.S., N. Humaira, S. Samadi, and N. Hubig. 2023. "Flowdyn: A Daily Streamflow Prediction Pipeline for Dynamical Deep Neural Network Applications." *Environmental Modelling and Software* 170: 105854.

Tan, W.Y., S.H. Lai, F.Y. Teo, and A. El-Shafie. 2022. "State-of-the-Art Development of Two-Waves Artificial Intelligence Modeling Techniques for River Streamflow Forecasting." *Archives of Computational Methods in Engineering* 2022: 1–27.

Taylor, K.E. 2001. "Summarizing Multiple Aspects of Model Performance in a Single Diagram." *Journal of Geophysical Research* 106: 7183–92 (also see PCMDI Report 55, http://wwwpcmdi.llnl.gov/publications/ab55.html).

Tirivarombo, S., D. Osupile, and P. Eliasson. 2018. "Drought Monitoring and Analysis: Standardised Precipitation Evapotranspiration Index (SPEI) and Standardised Precipitation Index (SPI)." *Physics and Chemistry of the Earth, Parts A/B/C* 106: 1–10.

Tyralis, H., G. Papacharalampous, and A. Langousis. 2021. "Super Ensemble Learning for Daily Streamflow Forecasting: Large-Scale Demonstration and Comparison with Multiple Machine Learning Algorithms." *Neural Computing and Applications* 33(8): 3053–68.

Vandal, T., E. Kodra, and A.R. Ganguly. 2018. "Intercomparison of Machine Learning Methods for Statistical Downscaling: The Case of Daily and Extreme Precipitation." *Theoretical and Applied Climatology* 137(1): 557–70.

Yaseen, Z.M., I. Ebtehaj, H. Bonakdari, R.C. Deo, A. Danandeh Mehr, W.H.M.W. Mohtar, L. Diop, A. El-shafie, and V.P. Singh. 2017. "Novel Approach for Streamflow Forecasting Using a Hybrid ANFIS-FFA Model." *Journal of Hydrology* 554: 263–76.

Yaseen, Z.M., A. El-shafie, O. Jaafar, H.A. Afan, and K.N. Sayl. 2015. "Artificial Intelligence Based Models for Stream-Flow Forecasting: 2000–2015." *Journal of Hydrology* 530: 829–44.

Yilmaz, K.K., H.V. Gupta, and T. Wagener. 2008. "A Process-Based Diagnostic Approach to Model Evaluation: Application to the NWS Distributed Hydrologic Model." *Water Resources Research* 44: 9417.

Zadeh, L.A. 1965. "Fuzzy sets." *Information and Control* 8(3): 338–53.

Zhang, Y., B. Ait-El-Fquih, K. Katterbauer, A. Shehri, and I. Hoteit. 2024. "Iterative Ensemble Smoothing Scheme for Reservoir History Matching with Unknown Observation Error Covariance." *Geoenergy Science and Engineering* 234: 212640.

Zhang, Z., Q. Zhang, and V.P. Singh. 2018. "Univariate Streamflow Forecasting Using Commonly Used Data-Driven Models: Literature Review and Case Study." *Hydrological Sciences Journal* 63(7): 1091–111.

Zhu, R., L. Yang, T. Liu, X. Wen, L. Zhang, and Y. Chang. 2019. "Hydrological Responses to the Future Climate Change in a Data Scarce Region, Northwest China: Application of Machine Learning Models." *Water* 11: 1588. https://doi.org/10.3390/w11081588.

Zhu, S., B. Hrnjica, M. Ptak, A. Choiński, and B. Sivakumar. 2020. "Forecasting of Water Level in Multiple Temperate Lakes Using Machine Learning Models." *Journal of Hydrology* 585: 124819.

How to cite this article: Jahanbani, Heerbod, Khandakar Ahmed and Bruce Gu. 2024. "Data-driven Artificial Intelligence-based Streamflow Forecasting, a Review of Methods, Applications, and Tools." *JAWRA Journal of the American Water Resources Association* 60(6): 1095–1119. https://doi.org/10.1111/1752-1688.13229.

APPENDIX A

TABLE A1 Summary of data-driven AI-based methods in streamflow forecasting.

AI method	Application	Author and year
ANN	Rainfall-runoff modeling	Shortridge et al. (2016)
ELM	Streamflow forecast under climate change	Zhu et al. (2019)
ANN	Monthly long-term streamflow prediction based on 12 antecedent flow	Lin et al. (2006)
LSTM	Daily streamflow forecasting	Rahimzad et al. (2021)
LSTM	Cascade streamflow forecasting	Liu et al. (2022)
VIF-PLC-LSTM	Daily runoff forecasting	Lian et al. (2022)
CNN-BAT	Daily streamflow forecasting rainfall and runoff with antecedent events	Khosravi et al. (2022)
ANFIS	Streamflow simulation using climate indices and ENSO effects	Sharma et al. (2015)
ANFIS	Daily streamflow prediction with limited data	Ashrafi et al. (2017)
ANFIS-FFA	Monthly streamflow forecasting	Yaseen et al. (2017)
ANFIS-BA	Monthly streamflow forecasting using predictors of lagged climate indices	Ehteram et al. (2019)
ANFIS-PSO	Monthly streamflow forecasting using predictors of lagged climate indices	Ehteram et al. (2019)
ANFIS-GA	Monthly streamflow forecasting using predictors of lagged climate indices	Ehteram et al. (2019)
ANFIS-GBO	Monthly streamflow forecasting using climate data and antecedent flow	Adnan, Zounemat-Kermani, et al. (2021)
ANFIS-GWO	Monthly streamflow forecasting using climate data and antecedent flow	Adnan, Zounemat-Kermani, et al. (2021)
ANFIS-PSO	Monthly streamflow forecasting using climate data and antecedent flow	Adnan, Zounemat-Kermani, et al. (2021)
ANFIS-ACO	Monthly streamflow forecasting using climate data and antecedent flow	Adnan, Zounemat-Kermani, et al. (2021)
ANFIS-GA	Monthly streamflow forecasting using climate data and antecedent flow	Adnan, Zounemat-Kermani, et al. (2021)
ANFIS-DE	Monthly streamflow forecasting using climate data and antecedent flow	Adnan, Zounemat-Kermani, et al. (2021)
CFNN	Monthly river flow prediction using flow and climate data	Hayder et al. (2020)
SVM	Daily streamflow forecasting	Hipni et al. (2013)
SVR	Streamflow forecast under climate change	Zhu et al. (2019)
SVR	Monthly long-term streamflow prediction based on 12 antecedent flow	Lin et al. (2006)
LSSVM	Predict streamflow based on antecedent flow	Kisi et al. (2019)
SVM-WT	Daily streamflow forecasting using antecedent flow data and lead-time 3–5 days	Kambalimath and Deka (2021)
SVM-DWT	Monthly streamflow forecasting	Kalteh (2013)
SVM-HHT	Monthly streamflow forecasting	Kalteh (2013)
SVM-GA	Monthly streamflow forecasting	Kalteh (2013)
PCA-SVM	Monthly streamflow forecasting with maximum of 3 months lead-time	Noori et al. (2011)
GT-SVM	Monthly streamflow forecasting with maximum of 3 months lead-time	Noori et al. (2011)
FS-SVM	Monthly streamflow forecasting with maximum of 3 months lead-time	Noori et al. (2011)
MARS	Predict streamflow based on antecedent flow	Kisi et al. (2019)
MLR	Daily streamflow prediction with covariable selection	Reis et al. (2021)
MLR	Daily streamflow prediction with 1–5 days lead time	Zhang et al. (2024)
RF	Daily streamflow forecasting using different co-variables	Adnan, Zounemat-Kermani, et al. (2021)
M5tree	Monthly streamflow prediction using climate signals and antecedent flow	Kisi et al. (2019)
GB	Various types of streamflow forecasting	Ni et al. (2020), Ahmadianfar et al. (2023)

TABLE A1 (Continued)

AI method	Application	Author and year
K-NN	Daily streamflow forecasting for reservoir	Akbari and Afshar (2014), Ababaei et al. (2013)
Stochastic	Streamflow selection (prediction) under climate change	Kiem et al. (2021)

Abbreviations: ACO, ant colony optimization; AI, artificial intelligence; ANFIS, adaptive neuro-fuzzy inference system; ANN, artificial neural network; BA, bat algorithm; BAT, bat algorithm; CFNN, cascade-forward neural networks; CNN, convolutional neural network; DE, differential equation; DWT, discrete wavelet transform; ELM, extreme learning machine; ENSO, El Niño Southern Oscillation; FFA, firefly optimization algorithm; FS, fuzzy set; GA, genetic algorithm; GB, gradient boosting; GBO, gradient-based optimization; GT, gamma test; GWO, Gray Wolf optimizer; HHT, Hilbert–Huang transform; K-NN, K-nearest neighbors; LSSVM, least square support vector machine; LSTM, long short-term memory; MARS, multivariate adaptive regression Splines; MLR, multiple linear regression; PCA, principal component analysis; PLC, pairwise linear correlation; PSO, particle swarm optimization; RF, random forest; SVM, support vector machine; SVR, support vector regression; VIF, variance inflation factor; WT, wavelet transform.

APPENDIX B

PRACTICAL EXAMPLE

This manuscript reviewed the recent data-driven artificial intelligence-based methods and tools. However, using these techniques and tools in practice may not be easy and straightforward, especially for beginners. This section provides a simple example of a streamflow forecasting problem and shows how to use different techniques to predict. A copy of the code and the input data are provided on GitHub: https://github.com/HeerbodJ/Streamflow_Forecasting.

Historical streamflows are available from Jan 1913 to Nov 2021 for a particular site (Monthy_Stf.csv). The question is to predict Dec 2021. In other words, the question here is how to predict Dec flow based on the antecedent flows (Jan to Nov). The R code is Data-driven_AI_Streamflow.R which is self-explanatory with text before each line of the code.