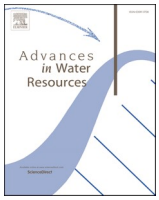
Advances in Water Resources 182 (2023) 104569

Contents lists available at ScienceDirect

Advances in Water Resources

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



Data reformation – A novel data processing technique enhancing machine learning applicability for predicting streamflow extremes

Vinh Ngoc Tran [a], Valeriy Y. Ivanov [a,*], Jongho Kim [b]

[a] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109, USA
[b] School of Civil and Environmental Engineering, University of Ulsan, South Korea

ARTICLE INFO

Keywords:
Machine learning
Extrapolation
Data reformation
Relative strength index
Streamflow predictions
Extreme events
Out-of-samples

ABSTRACT

Hydrologists have been actively exploring the utility of machine learning (ML) models for predicting streamflow. While ML methods have proven to be as accurate as conventional modeling techniques for streamflows well represented in the training set, they continue to lack satisfactory skills for extreme events. In this study, a novel 'data reformation' technique is proposed based on the Relative Strength Index (RSI) – a measure of speed and direction of changes in the time series. RSI homogenizes all observations to a constrained 0–100 range, and all 'out-of-sample' data in the testing set fall within the space of the training set. Long Short-Term Memory network with an attention mechanism is used to train three ML models using 55,055 events from the CAMELS dataset (670 basins, 1980–2014). Predictions are made for 12,424 events, of which 3,810 are significantly higher than streamflows in the training set. The ML model based on RSI-reformed data exhibits superior performance, as compared to other advanced ML models without data reformation. Peaks up to 15 times larger than those in the training events are accurately predicted, leading to an outperforming model skill for 433 out of 670 catchments. These findings indicate that incorporating a new data reformation technique into the data pre-processing step in ML modeling can enhance the utility of ML models for extreme events. This research encourages further exploration to identify better data reformation methods to enable confident ML predictions.

1. Introduction

Machine learning (ML) has gained considerable traction in the geophysical science community [Reichstein et al., 2019]. While still in their nascent state, many studies over the last decade have demonstrated that ML models can surpass existing state-of-the-art modeling techniques in complex problems, and hydrologists have started actively exploring ML in the domain of streamflow simulation [Ahn et al., 2022; Feng et al., 2021; Han et al., 2023; Kratzert et al., 2018]. ML algorithms do not rely on predetermined equations or assumptions (such as traditional process-based models), but rather learn from the data themselves, thereby enabling them to adapt to evolving conditions and uncover hidden insights [*Xu and Liang*, 2021]. This adaptability and capacity to handle intricate interactions make ML a compelling candidate for improving streamflow predictions [Alizadeh et al., 2021; Kratzert et al., 2018].

Specifically, multiple studies have implemented streamflow simulation and highlighted that the applicability outcomes of ML exceed those of simple lumped hydrological models [Arsenault et al., 2023; Frame et al., 2021; *Liu* et al., 2023a]. The application of ML is possible for predictions and simulations across various timeframes, ranging from hours, days, and months [*Cheng* et al., 2020; Dehghani et al., 2023; *Hunt* et al., 2022; *Xiang and Demir*, 2020]. ML applicability can be broad and this approach has been employed in numerous case studies including global datasets [Tang et al., 2023; Wilbrand et al., 2023]. "Prediction of ungauged basins" (PUB) has been one acute area of research interest in hydrology [Feng et al., 2021; Kratzert et al., 2019b; Le et al., 2022]. A number of studies have also focused on advancing ML methods such as data processing and model optimization, as well as designing new model types such as hybrid models [Ahmed et al., 2021; Konapala et al., 2020; Liu et al., 2023b; Nourani et al., 2014; Yu et al., 2023] or physically informed ML [Bhasme et al., 2022; Frame et al., 2021; Lu et al., 2021; Zhong et al., 2023].

Nevertheless, ML applications are not without inherent drawbacks. These techniques exhibit challenges related to data non-linearity and non-stationarity, model interpretability, uncertainty quantification, model selection, the need for high-quality training data, and "out-of-sample prediction" [Quilty et al., 2023; *Xu and Liang*, 2021]. The latter

* Corresponding author.
E-mail address: ivanov@umich.edu (V.Y. Ivanov).

https://doi.org/10.1016/j.advwatres.2023.104569
Received 23 June 2023; Received in revised form 22 October 2023; Accepted 31 October 2023
Available online 3 November 2023
0309-1708/Published by Elsevier Ltd.

challenge, which refers to the estimation of magnitude ranges to which ML training has not been exposed to, has been identified as one of the greatest challenges for ML models over the past decades [Frame et al., 2021; Kratzert et al., 2019a; *Todini*, 2007; *Tran and Kim*, 2022]. This has led to a debate between physical process-oriented modelers and data-driven modelers, with the former arguing that ML models lack an appreciation of physical characteristics and dynamics in their study domains, resulting in a lack of confidence in data-driven model outputs due to their heavy reliance on training sets [*Todini*, 2007]. Similarly, it has been argued that data-driven models may not be as effective in conditions that differ from the training data [*Kirchner*, 2006; Vaze et al., 2015]. Indeed, a conventional application of machine learning methodologies might be unable to accurately predict or extrapolate estimates outside of the training data space, despite their generally strong predictive capabilities for data within it [Kratzert et al., 2019a; Tran et al., 2020; *Tran and Kim*, 2022].

Research into the utilization of ML has sought to tackle the challenge of the "testing set" outside of space of the "training set" by broadening the data range of the training set to encompass a wider range of potential scenarios. However, this is often not feasible due to the difficulty of obtaining a sufficient number of observed extreme events for training – simply because they have not been observed. Climate change has the potential to create extreme events that have not been experienced before [Bao et al., 2017; *Bloschl* et al., 2020; *Doi and Kim*, 2020; 2021; *Prein* et al., 2016], and internal climate variability can also lead to extreme events that are different from those that have been recorded, even if climate remains stationarity [Bao et al., 2017; Beniston et al., 2007; *Doi and Kim*, 2020; Gao et al., 2020; Kim et al., 2018; Milly et al., 2008]. This limited capacity to extrapolate data is a major impediment for ML to be applied in real-world settings with increasingly frequent extreme events [Donat et al., 2016; *Dottori* et al., 2018; *Ivanov* et al., 2021; *Prein* et al., 2016], and thus an alternative solution is needed to ensure predictability of events beyond the available training data space.

A search of the Web of Science (accessed in January 2023) for two keywords "streamflow" and "machine learning" yielded a total of 466 research studies. This number attests to the burgeoning use of ML in streamflow modeling research. Surprisingly, no results were found when adding the keywords "extrapolation" or "out-of-sample prediction". A broader search was conducted for machine learning studies in all fields on time series applications related to "out-of-distribution" (OOD). Most of the identified studies mainly used OOD in data splitting (between training and testing datasets) for evaluating model performance [Ahmad et al., 2021; Boyer et al., 2021; Geiger et al., 2020; Moller et al., 2021; Olenskyj et al., 2022; Yeung et al., 2021], rather than focusing on proposing approaches to address this issue. The most relevant research that can be found is on the potential of ML in simulating "extreme events" [e.g., Frame et al., 2021]. They are defined as high-return-period (low-probability) streamflow events. It should be noted that "extreme events" may or may not indicate that the events are outside the scope of the training data. Recent research by Frame et al., [2021] has proposed a technique to improve the effectiveness of ML in predicting such extreme events. The authors hypothesized that incorporating physical (i.e., mass balance) constraints into the ML architecture would be advantageous. However, their evaluation results showed that pure ML was more effective than the physically-informed ML approach and that "adding mass balance constraints to the data-driven model reduced model skill during extreme events". Liu et al., [2023b] and Quilty et al., [2023] proposed the use of sophisticated neural networks to measure the confidence intervals of predictions. By considering the uncertain range, this approach can enhance the predictability for OOD events, but not substantially different from the training dataset. An alternative way to improve the efficacy of ML in predicting extreme events could be to combine it with a process-based model [Konapala et al., 2020; Tran et al., 2023b]. Our recent study has put forward three potential strategies for augmenting the skill of data-driven models for "out-of-sample prediction" by: (i) enriching information on physical phenomena in data-driven models through the utilization of high-fidelity samples generated by process-based models; (ii) broadening the training data space by considering additional input and parameter uncertainties; (iii) or constructing a hybrid model that combines a standard predictive model with a model that has extrapolation capabilities [*Tran and Kim*, 2022]. Nevertheless, these strategies have been proposed for the generation of surrogate models that replicate a computationally expensive model using "synthetic data", but not for *pure ML* applications that solely utilize observational data. It implies that, as of now, no successful method has been established to address the issue of "out-of-sample prediction".

In this study, we propose a strategy that has the capacity to overcome this long-standing challenge. We hypothesize that a novel data preprocessing step called "*data reformation*" can re-scale data to be within a restricted range, resulting in ML training that is effective for "out-of-sample prediction" problem. Specifically, both training and testing data after they have undergone a reformation process are contained within the same "homogenized" data space. The implication of this process is that out-of-sample data are located in the same data space as the training samples, and can be referred to as "in-of-reformed" samples. The reformed data should be used to train an ML model instead of the original data. If both the training and testing datasets are constrained to the same range, the trained ML model should be able to compute well for out-of-original samples in the testing set. In this study, we specifically focus on the high-return-period (low-probability) streamflow (flood) events and compare the proposed ML model (using data reformation) with two baseline models that use different data processing techniques: (i) a standard ML model with data normalization only and (ii) a ML model with standard data transformation. In method (ii), data transformation such as a wavelet transform is a popular data processing method used in ML research, as it can be used to solve problems related to the diagnosis, classification, and forecasts of extreme weather events [Nourani et al., 2014; *Sang*, 2013; *Tran* et al., 2021]. It involves decomposition of time series into multiple lower-resolution subseries, and extracting useful information from the original data [Nourani et al., 2014]. In this study, we do not compare the performance of ML with benchmarking models since this has been done in many previous studies [Feng et al., 2021; Frame et al., 2021; Kratzert et al., 2018, 2019a].

2. Methods

2.1. ML model: LSTM with attention mechanism

Long Short-Term Memory (LSTM) network is a type of recurrent neural network (RNN) that can learn long-term dependencies between input and output features by resolving gradients that are expanding or vanishing [*Hochreiter and Schmidhuber*, 1997; Kratzert et al., 2018]. LSTM adapts vanilla RNNs with three gates (forget gate f_t, input gate i_t, and output gate o_t) and preserves more useful information of input. In this study, we use an *attention*-based LSTM, a state-of-the-art LSTM variant for predicting streamflow [Ding et al., 2019; *Hunt* et al., 2022; Vaswani et al., 2017]. The attention mechanism assigns scores to each input feature, allowing the consideration of interdependency of input sequences at various time steps [Wang et al., 2016]. This enables LSTM not only to handle the long-term dependencies of driving sequences over historical time steps, but also to improve the ability of LSTM to capture high nonlinearity [Alizadeh et al., 2021; Ding et al., 2020; Li et al., 2019]. Specifically, given the k-th input time series $\boldsymbol{x}_k = \{x_{k,1}, x_{k,2}, \ldots, x_{k,T}\}$ with T that is the size of the time series, we can construct an input attention mechanism by referring to the previous hidden state h_{t-1} and the cell state c_{t-1} as:

$$e_{k,t} = \boldsymbol{v}^T \tanh(\boldsymbol{W}_e h_{t-1} + W_e c_{t-1} + \boldsymbol{U}_e \boldsymbol{x}_k) \quad (1)$$

$$\alpha_{k,t} = \frac{e_{k,t}}{\sum_{j=1}^{N_{in}} \exp(e_{j,t})} \quad (2)$$

where $\boldsymbol{v}$, $\boldsymbol{W}_e$,$\boldsymbol{U}_e$ are learnable parameters; factor $\alpha_{k,t}$ is the attention weight measuring the importance of k-th input at time t; and *tanh* is an activation function of LSTM layer. Note that all learnable parameters will be called $\boldsymbol{\theta}$ hereafter ($\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{b}, \boldsymbol{v}\}$). With $\alpha_{k,t}$, we can adaptively extract the input data series with:

$$\tilde{\boldsymbol{x}}_t = \left[\alpha_{1,t}x_{1,t}, \alpha_{2,t}x_{2,t}, \ldots, \alpha_{N_{in},t}x_{N_{in},t}\right], \tag{3}$$

where N_{in} denotes the number of inputs $\boldsymbol{x}_t = [x_{1,t}, x_{2,t}, \ldots, x_{N_{in},t}]$. The hidden state at time t can be updated as:

$$h_t = \mathbf{f}(h_{t-1}, \tilde{\boldsymbol{x}}_t), \tag{4}$$

where $\mathbf{f}$ is an LSTM unit that can be estimated with $\tilde{\boldsymbol{x}}_t$, as detailed in S.1 in the Supplementary Material 1 (SM1). For more details about the attention mechanism, readers are referred to Wang et al., [2016].

2.2. Data reformation: relative strength index

After testing numerous data pre-processing techniques accepted from various disciplines such as engineering, economics, and social sciences, we have identified a suitable method that meets the needs of reformulating a time series data into a new data form that is constrained to a limited range. Specifically, the method relies on the Relative Strength Index (RSI) developed by *Wilder* [1978]. RSI is a momentum oscillator index that is widely used in the field of economics to measure the speed and change of price shifts. By reforming time-series data on prices into the RSI series, their values are restricted to be within a range between 0 and 100. Even if new extreme prices occur in the future, the new RSI values will remain within the range since the index only reflects the relative changes in the price, rather than their actual levels. This concept can also be applied for processing streamflow data.

Specifically, the RSI calculation for streamflow can be carried out in the following manner. To simplify its application, RSI is broken down into its basic components: average rise (*AR*), average fall (*AF*), and a duration of period (N_{RSI}) that contains the averaging intervals for *AR* and *AF* [*Wilder*, 1978]. The RSI is calculated as:

$$\mathrm{RSI} = 100 - \frac{100}{1 + AR/AF} \tag{5}$$

In this work, daily streamflow data are used, so one day is the considered data resolution. Consider that N_{AR} and N_{AF} are the number of total days corresponding to the rising and falling streamflow (as compared to the streamflow of the previous day), starting from the first record of daily flow (see Fig. 1 for illustration), such that $N_{AR} + N_{AF} = N_{RSI} - 1$. Variables *AR* and *AF* can be computed as:

$$AR = \frac{1}{N_{AR}} \sum_{i=1}^{N_{AR}} \Delta Q_i^{\uparrow}, \tag{6}$$

$$AF = \frac{1}{N_{AF}} \sum_{j=1}^{N_{AF}} \Delta Q_j^{\downarrow}, \tag{7}$$

where $\Delta Q_i^{\uparrow}$ and $\Delta Q_j^{\downarrow}$ represent the magnitudes of the increase or decrease in streamflow as compared to the previous day during the i-th rising and j-th falling days of streamflow series, respectively.

All components (i.e., N_{RSI}, N_{AR}, N_{AF}, $\Delta Q_i^{\uparrow}$ and $\Delta Q_j^{\downarrow}$) used in the computation of RSI are graphically visualized in Fig. 1. In application of Eq. (5), the duration N_{RSI} is in essence a window through which the original streamflow series are analyzed, allowing to obtain a single RSI value for each window location. Eq. (5) is applied on successive N_{RSI} periods shifted at the resolution of the original streamflow data (i.e., one day). As a result, upon data reformation in this manner, the total duration of the RSI series is $N_Q - (N_{RSI} - 1)$ days ($N_Q$is the total duration of streamflow data), as N_{RSI} first records of Q are not converted to RSI since they represent the size of the window. Thus, N_{RSI} should not exceed N_Q.

The utilization of RSI in practice is akin to any data processing

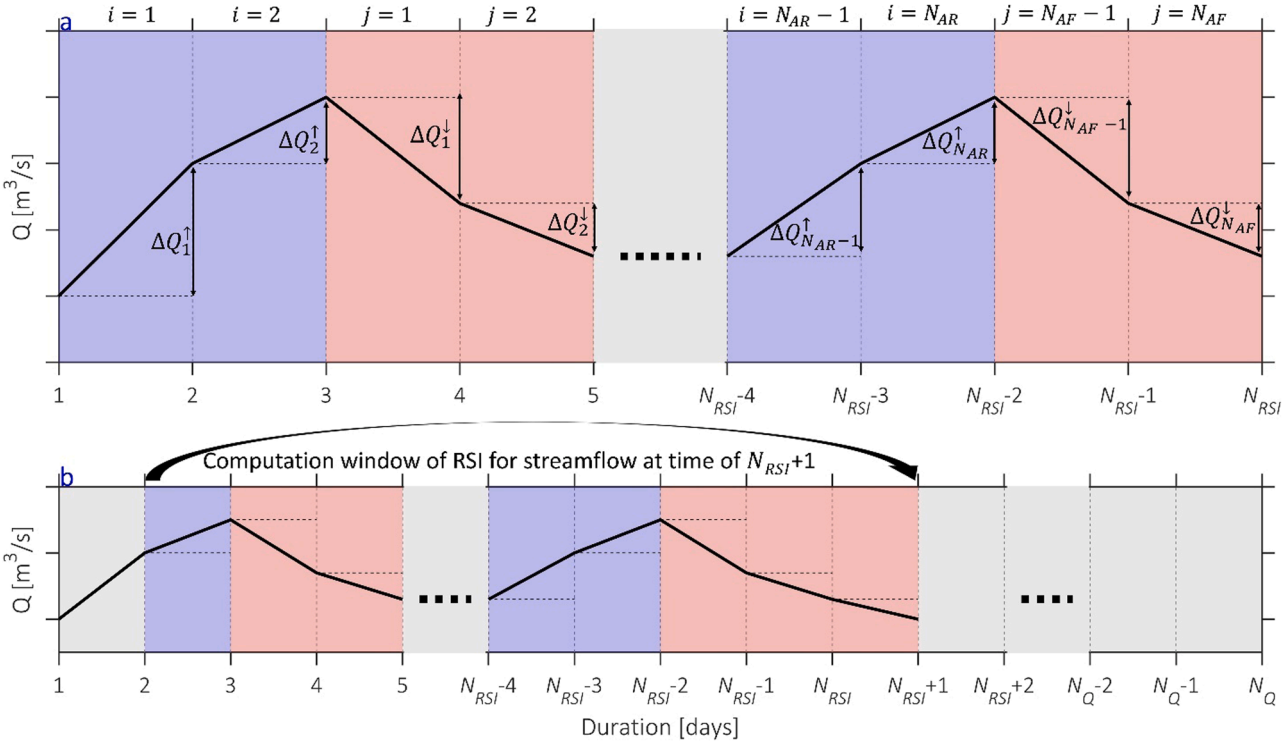


Fig. 1. A visualization illustrating essential components of computation of the RSI series described in Section 2.2. Subplots (a) and (b) describe the computation window of RSI for streamflows at times N_{RSI} and $N_{RSI} + 1$, respectively. Changes in streamflow are detected as positive $\Delta Q^{\uparrow}$or negative $\Delta Q^{\downarrow}$ (zero change is counted as either) and *AR* and *AF* can be computed as in Eqs. (6) – (7). The shaded areas in blue and red show the rising and falling periods (days) of streamflow, respectively.

technique (data normalization or transformation) [Ali et al., 2014; Liu et al., 2014; Zitnik et al., 2019]. Specifically, the measured streamflows are essential for the transformation/reformation process (i.e., RSI in this work). The reformed RSI time series then can be used for training of a ML model. When trained, the ML model will yield simulated RSI series as output in the testing/prediction phase without the need for actual streamflow. The predicted RSI values are subsequently converted into streamflows that have their original unit by using the inverse of Eqs. (5-7). Specifically, given the predicted RSI at times t and $t+1$ and streamflow (Q) data from $t - N_{RSI} + 1$ to t, the predicted Q at $t+1$ can be obtained by performing the inverse of Eqs. (5)-(7). Specifically, if $\mathrm{RSI}_{t+1} > \mathrm{RSI}_t$, the average fall ($AF$) can be computed as in Eq. (7), while the average rise (AR) can be computed inverse of Eq. (5) as follows:

$$AR = \left(\frac{100}{100 - \mathrm{RSI}_{t+1}} - 1\right) AF \tag{8}$$

Q_{t+1} is then computed by inverting Eq. (6) using the estimated AR:

$$Q_{t+1} = AR \times N_{AR} - \sum_{i=1}^{N_{AR}-1} \Delta Q_i^{\uparrow} + Q_t \tag{9}$$

The calculation of Q_{t+1} when $\mathrm{RSI}_{t+1} \leq \mathrm{RSI}_t$ is executed in a similar fashion as outlined above. Specifically, first, AR is computed as in Eq. (6), AF is then derived by inverting Eq. (5) using the estimated AR. Q_{t+1} is finally calculated by inverting Eq. (7) with the estimated AF. The computation of predicted Q for subsequent time steps is also performed in a similar manner, using Q obtained in prior steps.

The data reformation technique is distinct from other data preprocessing methods typically used in ML applications, such as data transformation. Conventional data transformation techniques, such as the Wavelet analysis and the Fourier analysis, are often applied to analyze non-stationary time series or transient phenomena in data. As highlighted earlier, when new out-of-sample data representing extreme events are included, the range of the transformed data is affected. Conversely, RSI data reformation converts the data into the same constrained space, even when new data with extreme values are added.

2.3. Discrete wavelet transform

The discrete wavelet transform (DWT) is a well-known statistical method that is used to decompose data series (i.e., climate forcings) into multiple sub-series with lower frequency by controlling the scaling and shifting factors of basic wavelets, also known as the 'mother' wavelets [*Kumar and Foufoula-Georgiou*, 1994; *Percival and Walden*, 2000]. DWT can be used to analyze non-stationary transitions such as breakdown points, discontinuities, and local minima and maxima [*Adamowski and Sun*, 2010]. Whilst there are numerous basic wavelets that can be chosen, this study applies the Daubechies wavelets [*Quilty and Adamowski*, 2018; Quilty et al., 2019] frequently used in ML applications.

One of the issues encountered with DWT in applications related to flow predictions is that it is not inherently shift-invariant, i.e., the values of the details and approximations do not change with the values of the original data series. This means DWT cannot be applied to problems related to singularity detection, forecasting, and nonparametric regression [*Maheswaran and Khosa*, 2012]. To overcome these problems, an à trous algorithm that uses redundant information attained from observational data has been suggested and used in this work [*Shensa*, 1992]. The decomposition formulas of an à trous algorithm are defined as [*Quilty and Adamowski*, 2018]:

$$D^j{}_t = A^{j-1}_t - A^j{}_t, \tag{10}$$

$$A^j{}_t = \sum_{l=0}^{L-1} g_l A^{j-1}_{t-2^{j-1} l \bmod N_t}, \tag{11}$$

where $D^j{}_t$ and $A^j{}_t$ represent the jth-level wavelet (detail) and scaling (approximation) coefficients of the original time series at time t; g_l is a scaling filter with $g_l = g_l^{\mathrm{DWT}}/\sqrt{2}$, where g_l^{DWT} is a scaling filter for DWT; L is the length of the scaling filter; l denotes index for L; and *mod* refers to the modulo operator. At $j = 0$, $A^0{}_t$ is equal to the original time series of $\boldsymbol{x}_t$. The latter can be obtained from the wavelet coefficients using additive reconstruction:

$$\boldsymbol{x}_t = \sum_{j=1}^{J} D^j{}_t + A^J{}_t. \tag{12}$$

An original signal is decomposed into D_t^1 and A_t^1 through the wavelet and scaling filters, and A_t^1 is further decomposed into D_t^2 and A_t^2 through the same process. This expansion is repeated until j reaches the maximum level J. The number of decomposed sub-series is $J+1$. For example, if $J=3$, the sub-series would be $[D_t^1, D_t^2, D_t^3, A_t^3]$ for each original time series. The total number of sub-series for N_{in} input variables is therefore $(J+1) \times N_{in}$. The approximation A^j becomes increasingly rough as j increases.

In this study, a level 3 of the decomposition was chosen to transform the data series. This choice is somewhat arbitrary, but it is consistent with what has been used in prior research [*Budu*, 2014; Nayak et al., 2013; Ni et al., 2020; Nourani et al., 2009; *Venkata Ramana* et al., 2013]. The input data series (i.e., climate forcings in this work) enters the wavelet transform model as input, and it is decomposed in the first level of decomposition into an approximate and a detailed one. In the next levels, the approximate signal is subsequently decomposed into a new approximate and a detailed one. After using the DWT to decompose a forcing data series, four sub-series are obtained, comprised of three detail parts and one approximation series. This combination of these four sub-series, alongside the original data, acts as the input for the ML.

2.4. Baseline models

In this study, two baseline models are designed for comparisons with the model that uses the RSI technique (referred to as $\mathrm{M_{RSI}}$). The first baseline model (referred to as $\mathrm{M_{Naive}}$) is a traditional model that relies solely on the original streamflow data (with a normalization technique employed). This approach is well-established in the literature and has been used in many studies involving the prediction of streamflow [Dehghani et al., 2023; Han et al., 2023; *Hunt* et al., 2022; Kratzert et al., 2018].

The second baseline ML model is developed using data transformation based on the discrete wavelet transforms (referred to as $\mathrm{M_{WT}}$). The rationale for constructing $\mathrm{M_{WT}}$ stems from prior research indicating that a ML model using this data transformation technique can more accurately detect and represent infrequent events, thereby enhancing accuracy of their simulation [*Adamowski and Sun*, 2010; Ni et al., 2019; Quilty et al., 2019; *Tran* et al., 2021].

3. Data

This research uses CAMELS (Catchment Attributes and Meteorology for Large-Sample Studies) dataset produced by the National Center for Atmospheric Research [*Newman* et al., 2015]. The dataset that has been extensively used in machine learning studies of streamflow simulation. The availability of this standardized dataset facilitates replication of published studies and comparisons with their results, which can maximize the effectiveness of new model development. The dataset is comprised of daily meteorological and discharge data from 671 catchments in the contiguous United States (CONUS) (Fig. 2a), with areas ranging from 4 to 25,000 km^2. These catchments typically exhibit natural flow patterns and have long-term streamflow gauge records spanning 1980 to 2014 [*Newman* et al., 2015]. In this study, 670 catchments were selected to conduct numerical experiments (basin 8202,700 was excluded due to its exceptionally short and discontinuous streamflow series). The Daymet dataset was used as the climate forcings [*Newman*

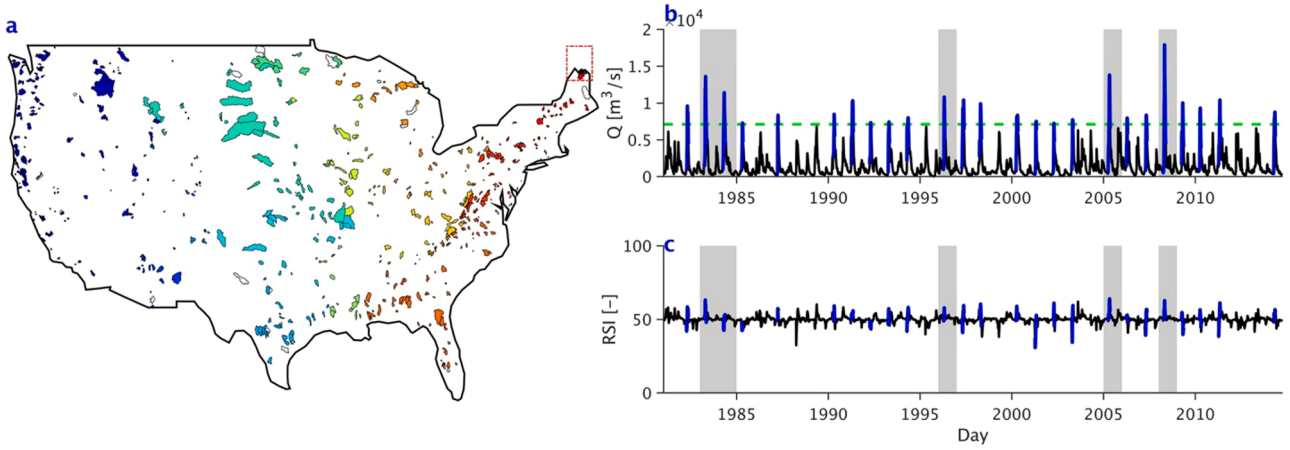


Fig. 2. Subplot (a) illustrates spatial locations of 670 selected catchments from the CAMELS database located in 18 Water Resources Council regions (indicated with basin shape colors) in the conterminous United States. Subplots (b) and (c) respectively present streamflow series (1981–2014) and the reformed RSI data for the first catchment 1013,500 (inside the red dash-line box in subplot (a)) in the database. The blue line indicates the selected streamflow (flood) events with peak flow higher than the threshold (green line), which is calculated as the 97.5% of daily flow using the entire hydrograph series (i.e., 2.5% of the flow duration curve). The areas shaded in gray represent time intervals containing flood events that are selected for model testing. The remaining flood events (i.e., no shading) are used for model training and validation.

et al., 2015; Thornton et al., 1840]. Meteorological data include the length of day-light (Dayl), daily total precipitation (Prcp), surface-incident solar radiation (Srad), snow water equivalent (Swe), 2-meter daily maximum and minimum air temperatures (Tmax and Tmin), and water vapor pressure (Vp) (hereafter referred to as "input variables"). This study focuses on constructing models for each individual basin, rather than aiming to develop a general model for PUB. Consequently, several static catchment attributes related to soils, climate, vegetation, topography, and geology were not used to construct "a universal ML model", as in previous PUB studies [Feng et al., 2021; Kratzert et al., 2019a, 2019b; Rahmani et al., 2021].

4. Experimental setup

4.1. Data processing

This study emphasizes the efficiency and robustness of data reformation in enhancing the capacity of ML to predict extreme flood events that are vastly dissimilar from the training events. To achieve this, we specifically designed a testing dataset to include events that are much larger than those used to train the ML algorithm. To illustrate the process, streamflow series for basin 1013,500 in the CAMELS dataset is used as exemplary (Fig. 2b-c). The data processing steps are carried out as follows.

(1) Determination of settings for data reformation. In order to reform the original streamflow data to RSI series, the first step is to ascertain the necessary period (window) durations N_{RSI}. We assessed the effects of N_{RSI} on the mutual relationship between the RSI series and the candidate inputs (i.e., Dayl, Prcp, Srad, Swe, Tmax, Tmin, and Vp), as evidenced by the Pearson correlation coefficient (Fig. A.1). Our tests indicate that with small N_{RSI}magnitudes, correlations between RSI series and most of the input variables are relatively low. Correlation values have the tendency to grow with larger N_{RSI} and reach a steady level with N_{RSI} larger than 200 days. A higher correlation between the input series and target variables is preferable in building a ML model [Hagen et al., 2021; Ren et al., 2020; Tran et al., 2021]. In this study, we opted for a sufficiently large N_{RSI} value (i.e., 365 days) and used it for all 670 catchments. It is important to be aware of the effect of N_{RSI} when selecting it, as an overly large N_{RSI} may result in a reduced amount of data following the reformation process. Given the choice of N_{RSI}, the streamflow series for station 1013,500 (Fig. 2b) is reformed into an RSI series depicted in Fig. 2c.

(2) Selection of extreme flood events. By incorporating a peak-over-threshold approach in this study, we filter event peak flows using a selected streamflow quantile as a threshold [Bačová-Mitková and Onderka, 2010; Lang et al., 1999; Solari and Losada, 2012]. Common choices of the threshold percentiles based on the flow duration curves used in prior studies are 75, 90, 95, 97.5, and 99 [Renard et al., 2006; Solari and Losada, 2012]. In this study, the 97.5 percentile threshold was chosen. This selection is somewhat arbitrary, but it allows for a sufficient number of flood events available for training and testing, also ensuring that the selected events are extreme. A fixed event duration of 30 days was applied for each included streamflow event, with the peak flow occurring on day 21 (the event duration does not have an effect on the model performance). The selected events for station 1013,500 are depicted in Fig. 2b (blue line), totaling in 25 flood events. The purpose of using an event-focused approach is to train a model that generates accurate predictions for peak flows. It is important to emphasize that ML is particularly apt at detecting patterns in data-rich areas (such as moderate and low flow conditions). However, ML models may fail in predicting extreme events due to the scarcity of peak data in the training dataset since their magnitudes are 'anomalous'. Therefore, we purposefully filter the data to enhance the fraction of data with high flows, while decreasing the level of representation of small and moderate flows. This might help improving the performance of the baseline models, particularly for predictions of flood peaks.

(3) Data partition into "training" and "testing" sets. Using the years with flood events selected in (2), an ordered set based on the annual maximum streamflow is developed. All events observed in the top 20% of the years with highest flows are selected as the "testing" set; the events observed in the remaining 80% of the years are considered as the "training/validation" set. One should keep in mind that the count of flood events per year may differ, and some peak flows in a given year of the testing set may have magnitudes that are smaller than those in the training/validation set. Nonetheless, the exemplary data set in Fig. 2b shows that streamflow maxima in the testing set have magnitudes that exceed those of the peak flows in the training/validation set. The corresponding RSI series are processed and partitioned in a

similar fashion, as seen in Fig. 2c. The non-chronological ordering of flood events between training and testing sets may raise concerns about the potential data leakage issues that arise when input-output pairs in the testing set are present in the training data. However, this problem did not occur in this case study, as the target outputs in the testing set are not included in the training/validation target sequences.

(4) Data normalization. Prior to the use of the data for the training and testing of ML models, all selected input variables and target outputs are standardized using min-max normalization, which is a standard process for training ML [Singh and Singh, 2020].

DWT is explicitly employed to decompose each forcing data for the M_{WT} model into four distinct subsets of series that are detailed in Section 2.3. In total, 35 data sub-series comprised of seven climate variables (i.e., Dayl, Prcp, Srad, Swe, Tmax, Tmin, and Vp) and the streamflow series is used as the inputs for the M_{WT}. In contrast, the M_{Naive} and M_{RSI} models employ 8 data series each, comprised of the same 7 climate variables and Q (for M_{Naive}) and the RSI (for M_{RSI}). The target outputs of M_{Naive}, M_{WT}, and M_{RSI} are Q, Q, and RSI series, respectively.

4.2. Model training

In this study, we design three models to predict streamflow for 30-day events with a lead time of 1 day. The configuration of candidate inputs and target outputs for the three models is presented as Eqs. (13)-(15).

$$Q_{t+1}^{Sim} = M_{Naive}\begin{pmatrix} \boldsymbol{x}_{1,t-L1} & \cdots & \boldsymbol{x}_{1,t+1} \\ \cdots & \cdots & \cdots \\ \boldsymbol{x}_{7,t-L7} & \cdots & \boldsymbol{x}_{7,t+1} \\ Q_{t-L_{RSI-365}}^{Obs} & \cdots & Q_t^{Obs} \end{pmatrix} \tag{13}$$

$$Q_{t+1}^{Sim} = M_{WT}\begin{pmatrix} \boldsymbol{x}_{1,t-L1} & \cdots & \boldsymbol{x}_{1,t+1} \\ \cdots & \cdots & \cdots \\ \boldsymbol{x}_{35,t-L35} & \cdots & \boldsymbol{x}_{35,t+1} \\ Q_{t-L_{RSI-365}}^{Obs} & \cdots & Q_t^{Obs} \end{pmatrix} \tag{14}$$

$$RSI_{t+1}^{Sim} = M_{RSI}\begin{pmatrix} \boldsymbol{x}_{1,t-L1} & \cdots & \boldsymbol{x}_{1,t+1} \\ \cdots & \cdots & \cdots \\ \boldsymbol{x}_{7,t-L7} & \cdots & \boldsymbol{x}_{7,t+1} \\ RSI_{t-L_{RSI}}^{Obs} & \cdots & RSI_t^{Obs} \end{pmatrix} \tag{15}$$

where L denotes the lookback window of each candidate input. To ensure fairness in comparing the models, the lookback windows of observed streamflow (Q^{Obs}) for M_{Naive} and M_{WT} models were determined to be $365+L_{RSI}$. That is, the number of historical values Q^{Obs} used to predict Q^{Sim} at $t+1$ is equal to the number of Q^{Obs} used to estimate RSI^{Obs} from $t - L_{RSI}$ to t.

Prior to training the ML model, a fundamental challenge in ML application studies is ascertaining the most suitable input variables ($\boldsymbol{x}$) and their lookback windows (L). Research has demonstrated that a variety of techniques can effectively tackle this issue [*Ahmad and Hossain*, 2019; Alizadeh et al., 2021; May et al., 2008; Thanh et al., 2022; *Tran* et al., 2021; Xu et al., 2022] and in this study, we implemented a most up-to-date method. Specifically, to determine important input variables and their lookback windows that have the greatest impact on the target outputs, we employed mutual information criterion in conjunction with the Hampel test [May et al., 2008] as a stopping criterion to select important input variables for ML training (see Text S.2 in SM for further details on variable selection). The prior range of the lookback window are from 1 to 365. The input time series for each event is processed separately and then stacked in order to form the training and testing dataset. Since ML model was constructed for each individual catchment, the input variables were also selected separately for each catchment. Note that the selected 30-day streamflow events from Section 4.1 are used as the target output, while the input data can be linger than 30 days and can extend beyond that time series.

In the next step, it is essential to tune the ML hyper-parameters, such as the number of hidden layers, the number of hidden units, the dropout rate, and the batch size [Kratzert et al., 2019b; *Yang and Shami*, 2020]. To accomplish this, we rely on the Bayesian optimization with a Gaussian process, a widely-utilized and highly-efficient approach (see Text S.3 in SM). The initial ranges for the values of the above four hyper-parameters were assigned to [1–4], [10–512], [0–0.9], and [8–512], respectively. The hidden states were initialized as zeros, which are default states in Tensorflow [Abadi et al., 2016]. The mean square error was utilized as the loss function, and the ADAM optimizer (with the learning rate of 0.0001) was employed to facilitate the training of the model [*Kingma and Ba*, 2014]. The learning rate was chosen somewhat arbitrarily; however, it aligns well with a number of previous studies that have used and recommended this value due to its proven effectiveness [*Cho and Kim*, 2022; Frame et al., 2021; *Hunt* et al., 2022; Le et al., 2019].

To maximize efficiency, an early stopping technique [Zhang et al., 2021] was implemented to expedite the training of the model. Specifically, this early stopping approach permits for an indefinite number of training epochs and terminates training when the model's performance ceases to advance on the validation dataset [*Liu and Mehta*, 2019]. The maximum number of epochs was predetermined to be 500 for all case study basins. Additionally, K-fold cross-validation (validation is not fixed to a particular subset in the training/validation set, [*Stone*, 1974]) with the K-fold number of 10 (as had been preferred in many prior studies) was used to ascertain whether the model has been sufficiently optimized [*Wong and Yeh*, 2019]. Specifically, the training/validation set is partitioned into $K = 10$ distinct, equitable subsets, or "folds". A ML model is then trained on K-1 folds of the data and subsequently validated on the leftover fold. This approach is cycled K times with K models, with each fold being used in turn as validation dataset.

4.3. Evaluation metrics

The testing set is used to evaluate ML model performance. Since this study is particularly concerned with the capability to predict peak flows, so the Exact Peak Error (EPE) metric [*Cunderlik and Simonovic*, 2004] was used to measure the accuracy of the model's predictions in comparison to the observed streamflow data:

$$EPE = \frac{Q_{peak}^{Sim} - Q_{peak}^{Obs}}{Q_{peak}^{Obs}} \times 100\%, \tag{16}$$

where Q_{peak}^{Sim} and Q_{peak}^{Obs} denote the simulated and observed streamflow peak for a given event, respectively. The unit of EPE is a percentage with a theoretical range of $(-\infty, +\infty)$. A negative value of EPE implies that the simulated peak is lower than the observed one, and vice versa.

Additionally, the traditional Nash–Sutcliffe efficiency (NSE) coefficient is utilized as a metric to assess the overall predictive skill of trained ML models.

$$NSE = 1 - \frac{\sum_{t=1}^{T}\left(Q_t^{Obs} - Q_t^{Sim}\right)^2}{\sum_{t=1}^{T}\left(Q_t^{Obs} - Q_{mean}^{Obs}\right)^2}, \tag{17}$$

where Q_t^{Obs} and Q_t^{Sim} are the actual observation and predicted streamflow outputs at time t; Q_{mean}^{Obs} is the mean of the observation over the entire event; T is the total number of time steps of the event. The value of NSE ranging from $-\infty$ to 1 and NSE = 1 indicates a perfect model with an estimation error equal to zero.

5. Results

5.1. Event selection and RSI reformation results

The reformation of the streamflow data and selection of the events were carried out for 670 catchments, resulting in the total of 55,055 streamflow events. Of these, 42,631 were used for the purpose of training and validating ML models (blue lines in Fig. 3e), while the remaining 12,424 events were reserved for testing the model (gray lines). In the testing set, 3810 events (or out-of-sample events) had peak flow magnitudes that exceeded the largest event in the training set, while the remaining 8614 events had peak flow magnitudes that are lower to the highest peak flows in the training set. The out-of-sample events are represented in Fig. 3e with normalized streamflow (NormQ) values greater than 1.0. The normalized series is obtained by dividing original streamflow by the maximum value in the training/validation dataset for each basin, resulting in the highest NormQ value of the training/validation set equal to 1. It is evident that peak flows of these events are significantly larger than peaks in the training set data, up to 5–14 times. The use of NormQ is to ensure the uniformity in data size and to highlight the differences between the data used for training the model and the data used for testing the model for all basins.

The RSI series (Fig. 3b) exhibit an appreciably lower range of variability than the original streamflow series (Fig. 3a). For example, Figs. 3c-d display examples for five events of an exemplary basin. Despite the apparent differences between the five events in the original series (e.g., the event shown in magenta has a peak 1.8–2 times higher than peak streamflow of the other events), when transformed into the RSI form, the time series of these events become nearly indistinguishable. This is precisely what is expected when converting out-of-sample into in-of-reformed samples.

Upon examination of the RSI series for 55,055 events (Fig. 3f), out-of-sample events cannot be clearly discerned in the testing set – this is an expected outcome of the application of the reformation data preprocessing. It can be seen that the RSI values only fluctuate within a limited range, and the data space of both the training/validation and testing sets are substantially similar (as demonstrated in the inset of Fig. 3f). The results illustrate the efficacy of the data reformation technique in bringing what would be out-of-samples in the physical streamflow range – into the in-of-reformed range for RSI values.

5.2. Model performance results

A comparison of the performance of three models in simulating streamflow for events in the test set is illustrated in Figs. 4, 5, and 6. The primary outcome is that the M_{RSI} model using the reformed data remarkably outperforms the baseline models in predicting out-of-sample events. Fig. 4 presents the results of three models' predictions for 4 first catchments in the CAMELS dataset as examples (i.e., 1013,500, 1022,500, 1030,500, and 1031,500). Results for other catchments can

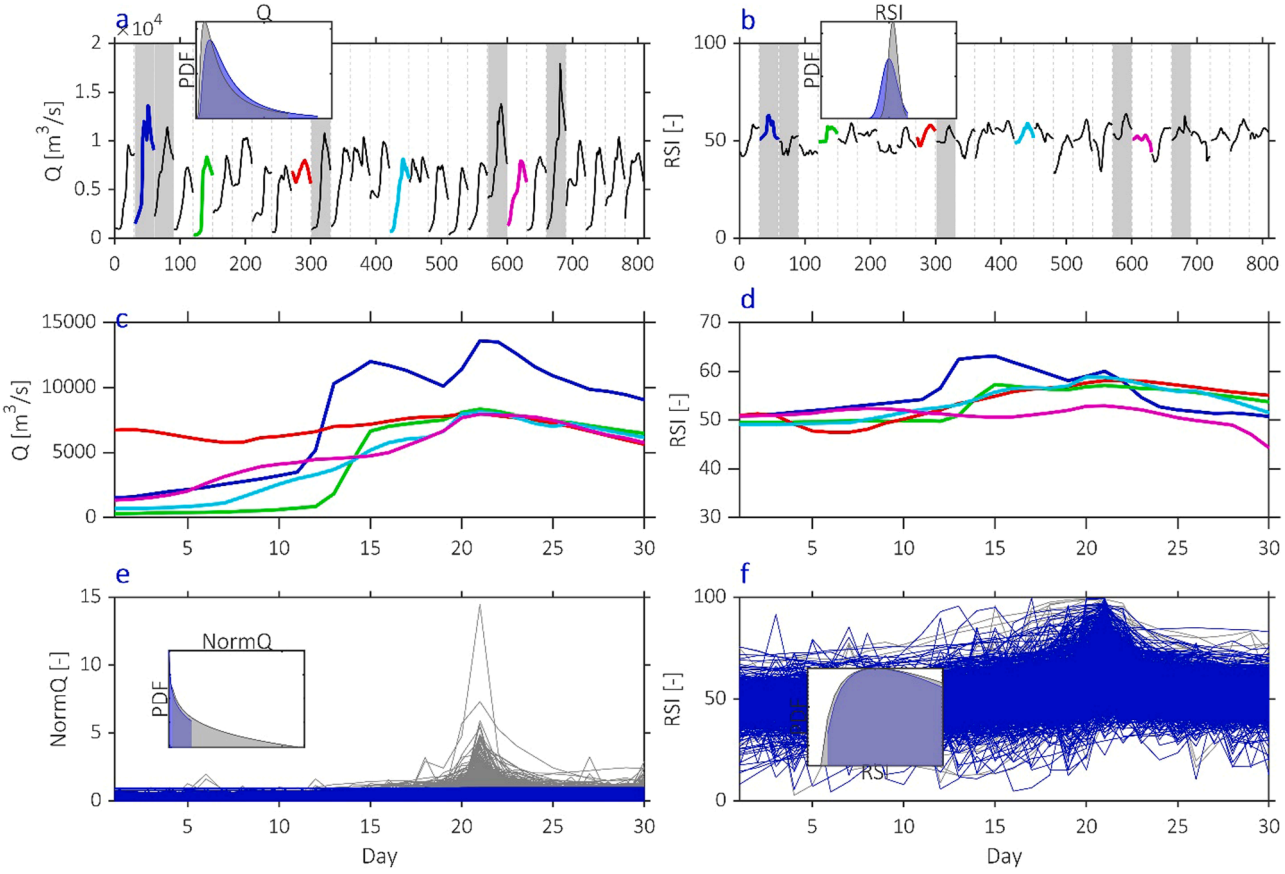


Fig. 3. (a) Concatenated hydrographs that are flood events selected as a subset of those shown in Fig. 2b (basin 1013,500), with 20% of the streamflow data (events highlighted by areas shaded in gray) used to test the ML models, and the remaining 80% used for training/validation. Colors of the time series illustrate different exemplary events in (c). The inset depicts the probability distribution function (PDF) of streamflow from the complete training and validation datasets (the blue shaded area) and the testing dataset (gray). (b) The reformed streamflow series in (a) in the form of Relative Strength Index (RSI). The inset plot shows PDF of the RSI data used for training (blue) and testing (grey). Subplots (c) and (d) show the original streamflow and reformed RSI series, respectively, for five events (colored lines in (a) and (b)) over the period of 30 days with peak flow occurring on day 21. (e) The normalized streamflow (NormQ) and (f) the RSI series for 55,055 streamflow events from 1980 to 2014 across 670 basins. The PDF based on the inverse Gaussian distribution of the two (training/validation vs. testing) sets are shown in (e) and (f) with both *x*-axis and *y*-axis in $\log_{10}$ scale, with the limits for *x*-axis in (a), (b), (e), and (f) being [0–20,000], [0–100], [0–15] and [0–100], respectively.

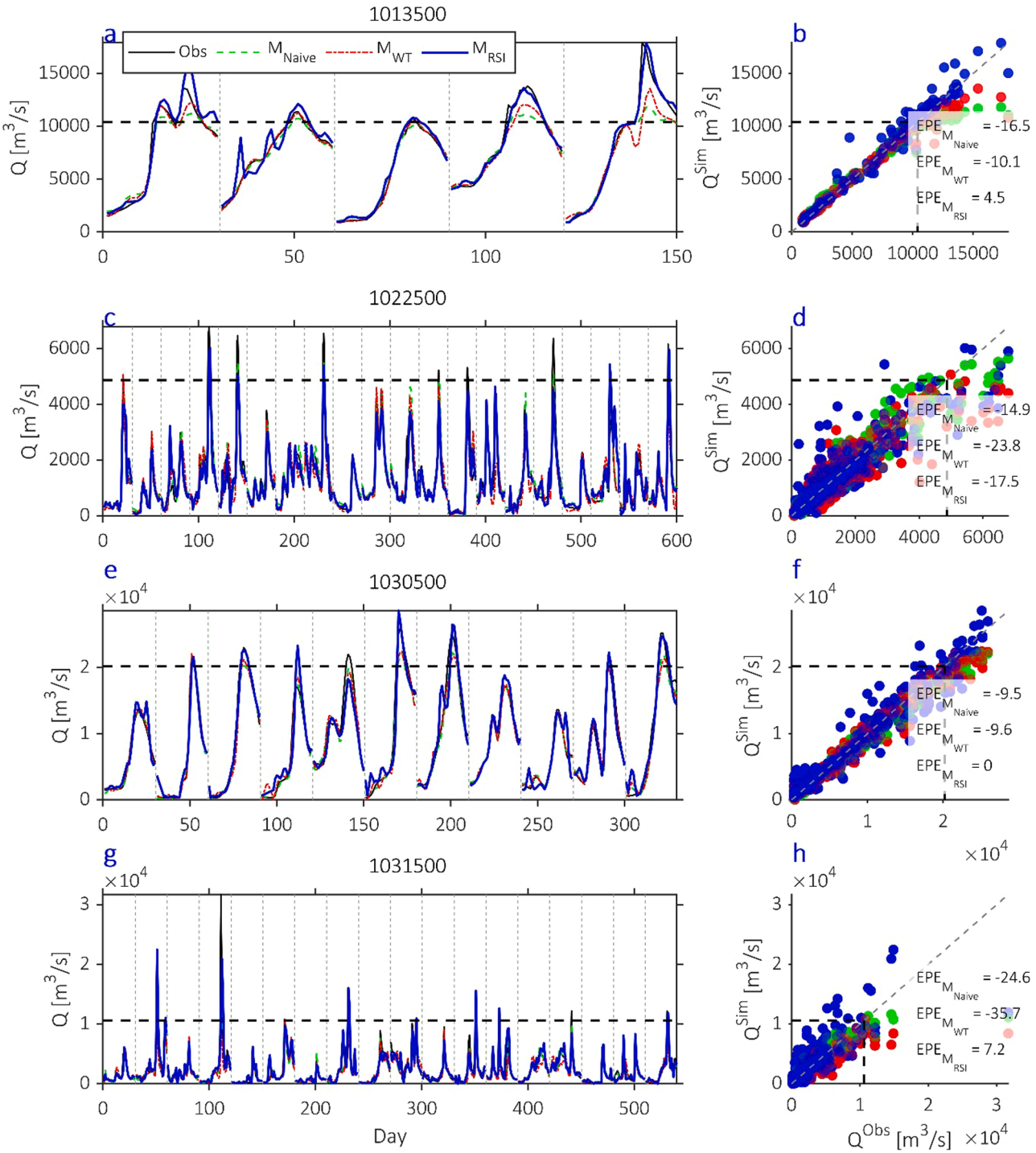


Fig. 4. (a, c, d, e) Comparisons of hydrographs simulated using the three ML models in the testing set and observations for four catchments (i.e., 1013,500, 1022,500, 1030,500, and 1031,500). (b, d, f, h) Scatter plots of simulated streamflow using the three models (*y*-axis) versus observations (*x*-axis). The black-dashed line in both subplots delineates the maximum streamflow in the training/validation set. The average EPE (exact peak error) values for the three models were computed for out-of-sample events that fall in the upper/left area of the plot demarcated by the black-dashed line.

be found in the dataset shared as Tran et al., [2023a]. Fig. 4 demonstrates that only the M_{RSI} model can provide accurate streamflow simulation for the events outside of the range of the training/validation set, except for results for the catchment of 1022,500. Both the M_{Naive} and the M_{WT} models were only able to generate satisfactory results for streamflow magnitudes up to near the highest limit of streamflows in the training set (the first event). The average EPE results reveal that with data outside the training/validation set, only the M_{RSI} model can yield highly accurate results (with EPE value of 4.5, 0, and 7.2%, respectively for catchment of 1013,500, 1030,500, and 1031,500). The M_{Naive} and M_{WT} models predict peak flows below the observed peaks, with EPE values below zero and the average EPE values ranging from about −10% (catchment 1030,500) to 35.7% (catchment 1031,500). It is not unexpected that for the events that have streamflows close to the range used in the training/validation set, all three models exhibit very good performance with the predicted peak flows that are very similar to the observations.

A comparison of results of peak flows for the three models across

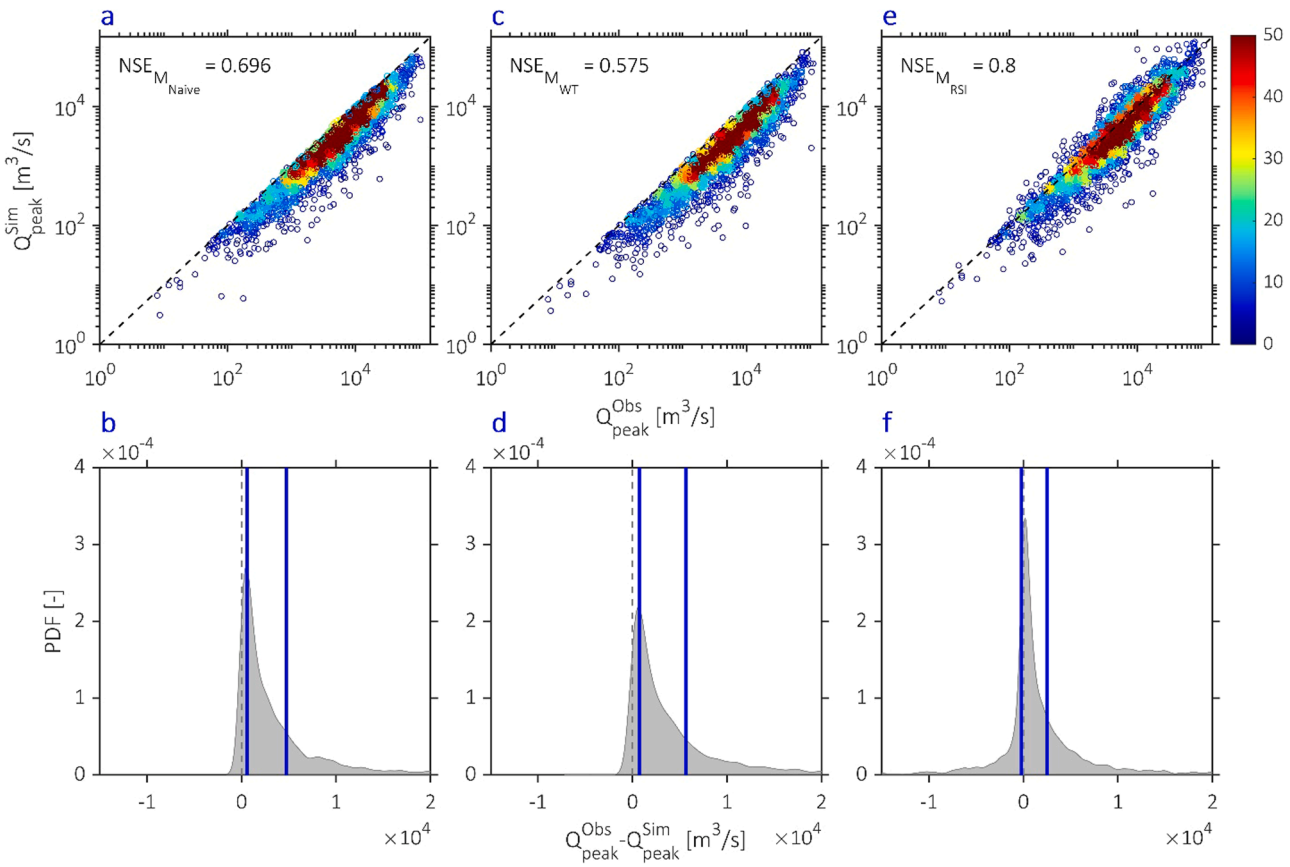


Fig. 5. Heatscatter plots (subplots a, c, and e) for 3810 simulated flow peaks (Q^{Sim}_{peak}) using the three ML models vs. observed flow peaks (Q^{Obs}_{peak}) in the testing set (all 670 basins). It should be noted that all 3810 peak streamflows shown in this figure exceed peak flows in the training/validation set (i.e., they have $NormQ_{peak}$ greater than 1.0). Both *x*- and *y*-axis are plotted using the log-scale. The colors reflect the density of data points (shown by the color bar). Three probability distribution function (PDF) plots (b, d, and f) depict the distribution of streamflow difference ($Q^{Obs}_{peak} - Q^{Sim}_{peak}$) computed for the three models, M_{Naive} (left), M_{WT} (center), M_{RSI}(right). The vertical blue lines denote the interquartile range (25% - 75%) of the PDFs.

3810 out-of-sample events in 670 catchments is shown in Figs. 5 and 6. The graphical inspection of results in Fig. 5 demonstrates that the M_{RSI} model yields the most satisfactory performance with an average NSE of 0.8 computed for the predicted versus observed peak flows. The M_{Naive} and M_{WT} models result in NSE values of 0.696 and 0.575, respectively. By analyzing the scatter plots in relation to 1:1 line, it is evident that the M_{RSI} model generates results that are more closely related to the observations (i.e., the dark red color in Fig. 5e indicates a high density of data points more evenly distributed along the 1:1 line). In contrast, Figs. 5a-c suggest that the two remaining models are less successful in accurately predicting the peak flows: the simulated magnitudes are generally lower than the measured values, and the absolute majority of the predicted flow peak values is below the actual observations. The difference between the observed and predicted peak flows of the three models ($Q^{Obs}_{peak} - Q^{Sim}_{peak}$) for 3810 flood events is illustrated in Figs. 5b-d-f. The results are consistent with the insights provided by the heatscatter plots: the M_{RSI} model produces mostly reliable results with the median streamflow difference close to zero, evenly spaced distance between the 25th and 75th percentiles on both sides of zero (Fig. 5f). In contrast, the peak differences for the M_{Naive} and M_{WT} are generally higher than zero, thus indicating that the predicted flood peaks are usually lower than observations. The PDFs for these two models place the 25–75th percentiles entirely in the positive region (Figs. 5b-d). Conversely, for the 8614 events that fall in the same range as the training events, the simulation results for all three models are comparable, all exhibiting satisfactory performance in predicting peak flows with a mean NSE value exceeding 0.9 (Fig. A.2).

Fig. 6a illustrates the spatial distribution of models with the highest performance. Here, the model performance is determined based on the mean of absolute EPE for all events for each catchment. Of the three ML models, the model with the smallest mean absolute EPE (i.e., closest to the theoretically ideal value of 0) is considered as the best one. By counting the number of watersheds in which a given ML model resulted in the best EPE, it can be seen that M_{RSI} surpasses M_{Naive} and M_{WT}, and is the most efficacious model for 433 catchments (accounting for 64.6% of basins). This can be compared to 190 (28.4%) and 47 (7%) watersheds in which the other two models, M_{Naive} and M_{WT}, respectively, result in the smallest EPE. It follows that the data reformation-based RSI method has a high degree of universality and is appropriate for application to various types of watersheds with varying hydrological, meteorological, or flow characteristics.

The EPE results for 3810 events are further demonstrated using the boxplots in Fig. 6b, with the mean EPE values obtained for the three models −41.25, −47.7, and −14.7%, respectively. Generally, these results are in line with the visual assessment in Fig. 5, showing that the M_{RSI} model can generate peaks that are closer to observations, as compared with the other two models. Particularly, for events that are more dissimilar from the training events (with NormQ ranging 3 to 15), the efficacy of M_{RSI} is further highlighted by the simulated peak magnitudes that are much closer to observations. Specifically, for 33.8% of all out-of-samples events in the testing set, the M_{RSI} simulations were found to be in excellent agreement with observations (EPE is within [−20%, +20%]) (see Fig. 7), even when events were more than five times higher than the upper limit of the range for training/validation events. In contrast, the majority of the EPE results (more than 99%)

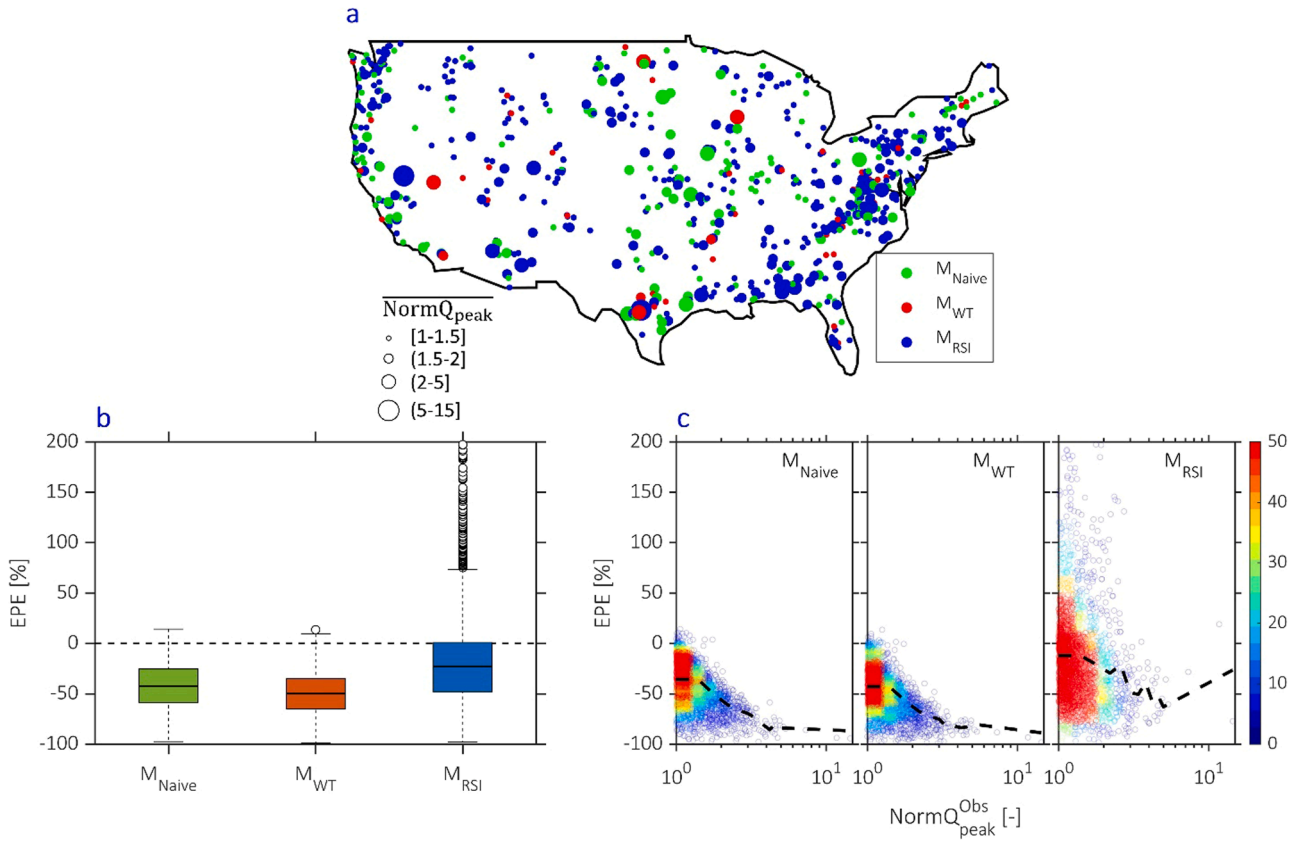


Fig. 6. The spatial map (a) illustrates distribution of the best model for 670 catchments in predicting peak flow based on the mean absolute Exact Peak Error (EPE). The three models M_{Naive}, M_{WT}, and M_{RSI} have the best performance for 190, 47, and 433 catchments, respectively. The circle size presents the average $NormQ_{peak}$ ($\overline{NormQ_{peak}}$) for all out-of-sample events (with NormQ > 1) for each catchment in the testing set. Boxplots in (b) are used to illustrate the EPE distribution for 3810 events across 670 catchments. The median (central mark), the 25th and 75th percentiles (edges of the box), and the maximum and minimum values excluding outliers (whiskers) are illustrated. Subplots in (c) display the heatscatter between EPE and $NormQ_{peak}$ (normalized peak flow) for the same events as in (a) and (b). The color of the dots reflects the concentration of data points (shown by the color bar). All events represented by this figure have peak magnitudes higher than the maximum peak flows in the training/validation set. The black dashed lines illustrate the average EPE value calculated according to the $NormQ_{peak}$ values with 0.5 and 3-sized bins between the intervals of [1–5] and (5–15], respectively.

calculated from M_{Naive} and M_{WT} have a value of less than 0, with respectively more than 82% and 90% of out-of-samples events below EPE = −20% (Figs. 6b and 7). For these two models, most of the simulation results with EPE close to 0 were obtained for events that were similar or not substantially different from the training events, i.e., their normalized peak flow ($NormQ_{peak}$) was close to 1 (Fig. 6c). Expectedly, the inadequacy of the M_{Naive} and M_{WT} models is more evidently demonstrated for more extreme events, i.e., those that have peak magnitudes further away from peak flows in the training set. The mean EPE for three models were computed for events whose $NormQ_{peak}$ values were situated within the same range depicted as the black-dashed lines in Fig. 6c. Specifically, these mean EPE values clearly exhibit the correlation between EPE and the size of the event, with events that are further away and dissimilar from the training events, the mean of EPE values for the two models (M_{Naive} and M_{WT}) are lower and closer to −100%. On the other hand, the relationship between EPE derived from

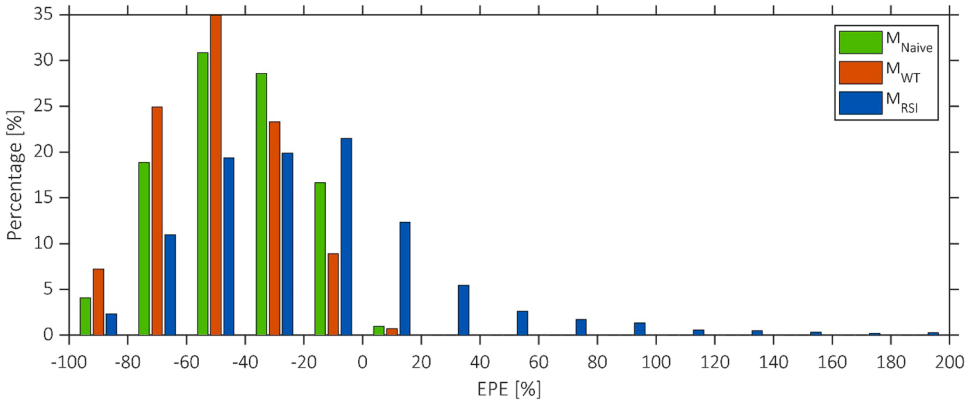


Fig. 7. The partitioning (as a percentage) of the estimated EPE values for the three models (see the legend) estimated for 3810 out-of-sample events across 670 catchments, with a 20% bin size employed.

M_{RSI} and the magnitude of events varies considerably compared to the other two models. The mean EPE obtained from M_{RSI} decreases from −10% to −55%, corresponding to $NormQ_{peak}$ values in the range of 1 to 5. For $NormQ_{peak}$ values exceeding 5, the behavior of the prediction skill becomes less apparent, but can reach a mean EPE value close to −20%, when the NormQ reaches 15 (black-dashed line in the last subplot in Fig. 6c).

6. Discussion

Although the M_{RSI} model has the capacity to predict beyond the magnitudes of events used in the training/validation set, the simulation results still remain far from ideal. Specifically, M_{RSI} predictions may still contain considerable errors with an absolute EPE greater than 50% for 20.7% of the out-of-sample events (789 out of 3810). While the M_{RSI} model surpasses the skill of the other two models, it is necessary to explore why it is not superior for all 670 watersheds (e.g., predicted results for catchment of 1022,500 in Fig. 4 and four catchments in Fig. S1 in SM).

The fundamental difference between the ML models lies in the use of RSI instead of Q as the target output. Our analysis reveals that while incorporating RSI for particular watersheds is beneficial for the "scaling" issue (i.e., to bringing training and testing datasets to the same range), it doesn't resolve many of the other inherent issues, such as strong non-linearity and non-stationarity of the watershed behaviors, or "hidden/unknown" uncertainties [*Gharib and Davies*, 2021; Prodhan et al., 2022; *Xu and Liang*, 2021]. Any of such features in the used dataset can lead to a poorer model performance, especially for events that have flow peaks near the training data distribution (highlighted in Fig. 6c, where the EPE values show an immense variation for small NormQ values). The proposed approach is viable for tackling one of the major challenges in machine learning applications, which is making out-of-sample predictions. It is particularly designed for extreme flood prediction. Therefore, caution is still required when selecting suitable data processing approaches, depending on the specific use cases (e.g., flood or drought prediction) and varying time frames, in order to achieve the most efficient model.

Further, it is logical to conclude that confidence intervals of ML models' predictions need to be assessed in their typical applications. Most of the current ML applications for streamflow prediction attempt to construct a "deterministic" model that fits optimally (without overfitting) the major fraction of training data, yet they do not attempt to provide uncertainty estimates [Alizadeh et al., 2021; Kratzert et al., 2018]. It is evident that the level of uncertainty in ML predictions can be significantly underestimated due to the lack of utilization of relevant information, such as data stochasticity [Kim et al., 2016a; b] or input and output noise [*Kendall and Gal*, 2017]. Uncertainty is inherent to all aspects of hydrological modeling, and it is generally accepted that predictions should account for it [*Beven and Freer*, 2001; Dwelle et al., 2019]. In many applications, such as streamflow predictions, it is as important to obtain the confidence of a prediction as the prediction itself [*Beven and Binley*, 1992]. Significant efforts have been undertaken to explore the uncertainties associated with physical-based models [*Beven and Binley*, 2014; Dwelle et al., 2019; Kim et al., 2015; *Moradkhani and Sorooshian*, 2008; Tran et al., 2020], while comparatively few efforts so far have been devoted to ML models, despite their recent surge in popularity [*Abdar* et al., 2021; Fang et al., 2020; Klotz et al., 2022; Liu et al., 2023b; Lu et al., 2021; *McDermott and Wikle*, 2019]. This presents an opportunity for future research to concentrate on evaluating the uncertainty of ML models, and inventing solutions to reduce it, while bolstering confidence in ML applications.

This study yields an unexpected outcome, that the M_{WT} model demonstrates a less successful performance than the M_{Naive} in predicting out-of-sample events. This is despite the usual assumption that extreme events can be more effectively discerned and therefore predicted through wavelet transformation and decomposition. This result demonstrates the efficacy of employing an attention mechanism in conjunction with an LSTM, allowing the LSTM to manage intricate relationships between inputs and outputs. It also allows to hone in on significant input variables that have a direct influence on the target output (streamflow), while disregarding any other input variables of lesser relevance. With the enhanced capability to recognize information for extreme events, the trained LSTM can yield accurate results, eliminating the need for additional methods, such as wavelet transform. This finding is also consistent with the conclusion of *Hunt* et al. [2022] and Han et al., [2023]. Conversely, the utilization of WL model type can have an adverse effect on the results, as the quantity of inputs fed into the training model will drastically grow after decomposition, resulting in a corresponding increase in the number of learnable parameters that need to be trained in LSTM. This implies that the model will be harder to train and may be more challenging to optimize, resulting in a potential decrease in model performance due to the issues of high dimensionality and convergence issues in model training [*Tran* et al., 2021]. Such results imply that with the introduction of advanced machine learning models that have the capacity to self-process information, the transformation of data using, for example, wavelet or Fourier transforms, may not be really necessary. Furthermore, this underlines the critical needs for ML progression in the coming years, such as, for example, the need to enable the creation of assisting mechanisms for ML to self- reform the data and bolster its extrapolation capabilities.

This study was designed to evaluate the efficacy of the proposed method, using information that is assumed to be available beforehand such as the forcing data and observed streamflow. In real-world settings, additional complexities arise regarding the inability to measure or collect observed flow data in a timely manner during extreme events (e.g., hurricane) due to, for example, infrastructure failures. To predict future flows, the proposed method requires only estimated inputs and observed streamflow for the prior time intervals (see Eq. (15)), enabling the methodology viability whenever such data are available. Flows predicted at preceding time intervals can be used to extend predictions for future time intervals, when observations are temporarily unavailable or require additional vetting. Real-world implementation of the proposed reformation technique warrants further investigation of its performance issues under various data availability scenarios, which is beyond the scope of this study.

Relatively straightforward data requirements may hint the feasibility of using this method for PUB studies that lack streamflow observations. Over the last decades, numerous studies have proposed various approaches and models in conjunction with regionalization or data assimilation techniques with the objective to reconstruct the past streamflow at high accuracy in the PUB context [Kratzert et al., 2019a; *Luce*, 2014; Sivapalan et al., 2003]. Consequently, we are of the opinion that the utilization of the RSI for PUB studies is feasible. Furthermore, a universal (or attribute-aware) model, trained with more diverse datasets drawn from different research areas, has the potential to simulate extreme peak flows even more accurately than the traditional M_{Naive} [Frame et al., 2021]. The application of such a model type combined with the RSI reformation would be a potential follow-on study in predicting the flow of extreme events for regions with insufficient data. Additionally, the potential application of the suggested approach can be extended to research utilizing ML for multidimensional (e.g., gridded fields) data predictions [Kotsiantis et al., 2007; Talukdar et al., 2020], even though this study concentrates on using ML for a univariate time series predictions. In geophysical sciences, not only state and flux variables (e.g., atmospheric temperature and humidity, soil moisture, canopy biomass, etc.) fluctuate with time but they also vary in space and that extending the method to reformation of multidimensional (2 or 3 dimensions) data has a theoretical potential. Computation of the relative change of quantities of interest between different locations in a multi-dimensional series can be accomplished analogously to what has been demonstrated for a univariate series.

7. Conclusion

The results of this study demonstrate practical usefulness of machine learning models for predicting extreme events that are much different from those that are used for training ML models. The central premise of the proposed method is that all data should be brought into a more homogenized data space, so that all out-of-samples can be converted into in-of-reformed samples. By reforming the data used for training/validation and testing to be in a homogenized data space, the difficulty of extrapolating out-of-samples goes away, and instead interpolation of in-of-reformed samples occurs. A noteworthy methodological point is that instead of using actual streamflow data, a different data kind is employed to train the model – namely, the relative change of streamflow derived using Relative Strength Index (RSI). The prediction of this relative change is then reversed to actual streamflow values. Overall, the results demonstrate that the prediction skill of the trained ML-RSI is remarkable, even for events that exceed magnitudes of the training/validation events by a factor of 3–15. Further research is necessary to construct better data reformation methodologies for training ML models to enhance their accuracy and ability to produce uncertainty quantification, when predicting extreme events that have not been encountered in the past.

CRediT authorship contribution statement

Vinh Ngoc Tran: Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Valeriy Y. Ivanov:** Supervision, Writing – review & editing, Funding acquisition. **Jongho Kim:** Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

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

Data availability

CAMELS data is available at https://ral.ucar.edu/solutions/products/camels availability. All machine learning models were trained using the Keras-Tensorflow library (https://github.com/leriomaggio/deep-learning-keras-tensorflow). Code for the Bayesian optimization, wavelet transform, mutual information, and Relative Strength Index were obtain from available open sources, including https://github.com/thuijskens/bayesian-optimization, https://github.com/PyWavelets/pywt, https://github.com/scikit-learn/scikit-learn, https://technical-analysis-library-in-python.readthedocs.io, respectively. The prediction results for all testing events over 670 catchments can be accessed at https://zenodo.org/record/7737,960#.ZBJkW3bMKUn.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C2008584). V. Ivanov acknowledges the support of the U.S. National Science Foundation CMMI program award # 2053429.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2023.104569.

Appendix

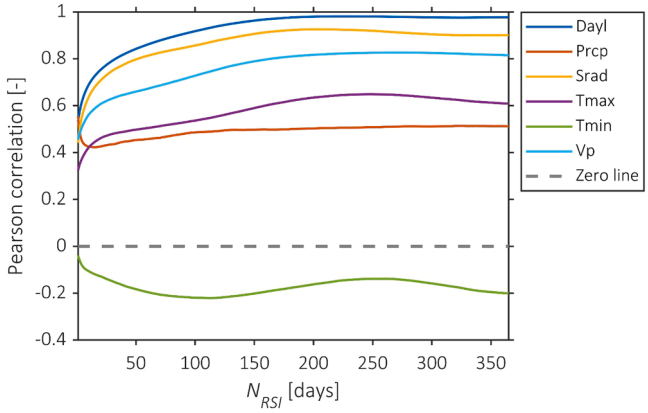


Fig. A.1. Demonstration of the influence of period duration, N_{RSI}, on the Pearson correlation between the RSI and individual input variables for catchment 1013,500. The value of N_{RSI} ranges from 1 to 365 days. The input variables include the length of day-light (Dayl [seconds]), daily total precipitation (Prcp [mm]), surface-incident solar radiation (Srad [W/m^2]), 2-meter daily maximum air temperature (Tmax [°C]), 2-meter daily minimum air temperature (Tmin [°C]), and water vapor pressure (Vp [Pa]). The snow water equivalent (Swe) values in this watershed are all equal to 0; therefore, the relationship between N_{RSI} and Swe is not depicted in this figure.

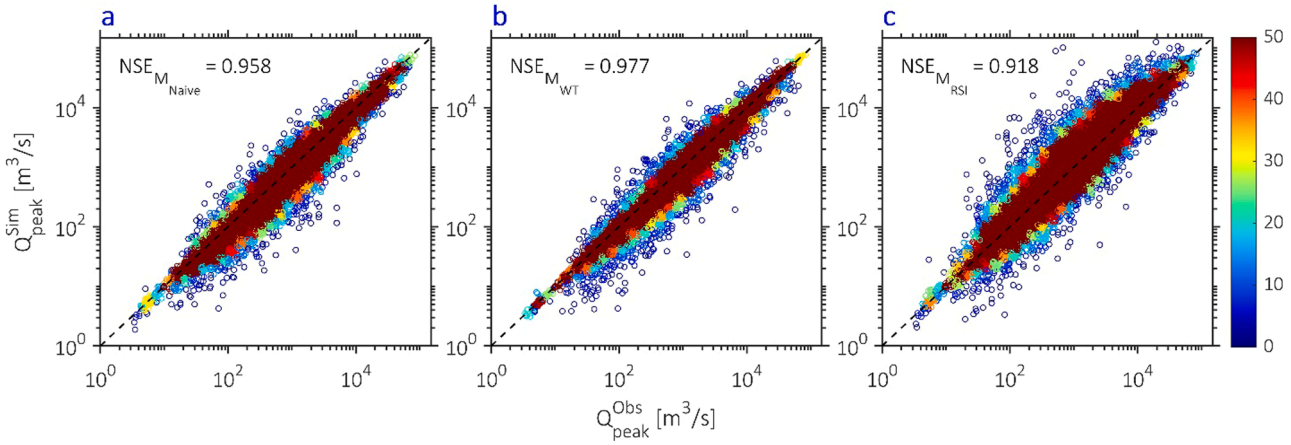


Fig. A.2. Heatscatter plots of 8614 simulated flow peaks by the three ML models versus observations for the testing set over 670 basins. The flood peaks shown in the subplots are smaller than the largest peak in the training/validation set, i.e., with $NormQ_{peak}$ smaller than 1.0. Both *x*- and *y*-axis are plotted using the log-scale. The colors represent the density of data points (shown by the color bar).

References

Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard (2016), {TensorFlow}: a system for {Large-Scale} machine learning, paper presented at 12th USENIX symposium on operating systems design and implementation (OSDI 16).

Abdar, M., et al., 2021. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Info. Fusion 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008.

Adamowski, J., Sun, K., 2010. Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. J. Hydrol. 390 (1–2), 85–91. https://doi.org/10.1016/j.jhydrol.2010.06.033.

Ahmad, S.K., Hossain, F., 2019. A generic data-driven technique for forecasting of reservoir inflow: application for hydropower maximization. Environ. Modell. Software 119, 147–165. https://doi.org/10.1016/j.envsoft.2019.06.008.

Ahmad, W., Shadaydeh, M., Denzler, J., 2021. Causal inference in non-linear time-series using deep networks and knockoff counterfactuals. paper presented at. In: Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE.

Ahmed, A.M., Deo, R.C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., Yang, L., 2021. Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. J. Hydrol. 599, 126350.

Ahn, S., Tran, T.D., Kim, J., 2022. Systematization of short-term forecasts of regional wave heights using a machine learning technique and long-term wave hindcast. Ocean Eng. 264, 112593 https://doi.org/10.1016/j.oceaneng.2022.112593.

Ali, P.J.M., Faraj, R.H., Koya, E., Ali, P.J.M., Faraj, R.H., 2014. Data normalization and standardization: a technical report. Mach Learn Tech Rep 1 (1), 1–6.

Alizadeh, B., Ghaderi Bafti, A., Kamangir, H., Zhang, Y., Wright, D.B., Franz, K.J., 2021. A novel attention-based LSTM cell post-processor coupled with Bayesian optimization for streamflow prediction. J. Hydrol. 601, 126526 https://doi.org/10.1016/j.jhydrol.2021.126526.

Arsenault, R., Martel, J.L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27 (1), 139–157.

Bačová-Mitková, V., Onderka, M., 2010. Analysis of extreme hydrological events on the Danube using the peak over threshold method. J. Hydrol. Hydromech 58 (2), 88–101.

Bao, J., Sherwood, S.C., Alexander, L.V., Evans, J.P., 2017. Future increases in extreme precipitation exceed observed scaling rates. Nat. Clim. Change 7 (2), 128–132. https://doi.org/10.1038/nclimate3201.

Beniston, M., Stephenson, D.B., Christensen, O.B., Ferro, C.A., Frei, C., Goyette, S., Halsnaes, K., Holt, T., Jylhä, K., Koffi, B., 2007. Future extreme events in European climate: an exploration of regional climate model projections. Clim. Change 81, 71–95.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Processes 6 (3), 279–298. https://doi.org/10.1002/hyp.3360060305.

Beven, K., and A. Binley (2014), GLUE: 20 years on, Hydrol. Processes, 28(24), 5897–5918, doi:10.1002/hyp.10082.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249 (1–4), 11–29. https://doi.org/10.1016/s0022-1694(01)00421-8.

Bhasme, P., Vagadiya, J., Bhatia, U., 2022. Enhancing predictive skills in physically-consistent way: physics informed machine learning for hydrological processes. J. Hydrol. 615, 128618.

Bloschl, G., et al., 2020. Current European flood-rich period exceptional compared with past 500 years. Nature 583 (7817), 560–566. https://doi.org/10.1038/s41586-020-2478-3.

Boyer, P., Burns, D., Whyne, C., 2021. Out-of-distribution detection of human activity recognition with smartwatch inertial sensors. Sensors 21 (5), 1669.

Budu, K., 2014. Comparison of wavelet-based ANN and regression models for reservoir inflow forecasting. J. Hydrol. Eng. 19 (7), 1385–1400. https://doi.org/10.1061/(asce)he.1943-5584.0000892.

Cheng, M., Fang, F., Kinouchi, T., Navon, I., Pain, C., 2020. Long lead-time daily and monthly streamflow forecasting using machine learning methods. J. Hydrol. 590, 125376.

Cho, K., Kim, Y., 2022. Improving streamflow prediction in the WRF-Hydro model with LSTM networks. J. Hydrol. 605, 127297 https://doi.org/10.1016/j.jhydrol.2021.127297.

Cunderlik, J., Simonovic, S.P., 2004. Calibration, Verification and Sensitivity Analysis of the HEC-HMS Hydrologic Model. Department of Civil and Environmental Engineering, The University of Western.

Dehghani, A., Moazam, H.M.Z.H., Mortazavizadeh, F., Ranjbar, V., Mirzaei, M., Mortezavi, S., Ng, J.L., Dehghani, A., 2023. Comparative evaluation of LSTM, CNN, and ConvLSTM for hourly short-term streamflow forecasting using deep learning approaches. Ecological Informatics 75, 102119.

Ding, Y., Zhu, Y., Feng, J., Zhang, P., Cheng, Z., 2020. Interpretable spatio-temporal attention LSTM model for flood forecasting. Neurocomputing 403, 348–359. https://doi.org/10.1016/j.neucom.2020.04.110.

Ding, Y., Y. Zhu, Y. Wu, F. Jun, and Z. Cheng (2019), Spatio-Temporal Attention LSTM Model for Flood Forecasting, paper presented at 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 14–17 July 2019.

Doi, M.V., Kim, J., 2020. Projections on climate internal variability and climatological mean at fine scales over South Korea. Stochastic Environmental Res. Risk Assessment 34 (7), 1037–1058. https://doi.org/10.1007/s00477-020-01807-y.

Doi, M.V., Kim, J., 2021. Addressing climate internal variability on future intensity-duration-frequency curves at fine scales across South Korea. Water 13 (20), 2828.

Donat, M.G., Lowry, A.L., Alexander, L.V., O'Gorman, P.A., Maher, N., 2016. More extreme precipitation in the world's dry and wet regions. Nat. Clim. Change 6 (5), 508–513. https://doi.org/10.1038/nclimate2941.

Dottori, F., et al., 2018. Increased human and economic losses from river flooding with anthropogenic warming. Nat. Clim. Change 8 (9), 781–786. https://doi.org/10.1038/s41558-018-0257-z.

Dwelle, M.C., Kim, J., Sargsyan, K., Ivanov, V.Y., 2019. Streamflow, stomata, and soil pits: sources of inference for complex models with fast, robust uncertainty quantification. Adv. Water Res. https://doi.org/10.1016/j.advwatres.2019.01.002.

Fang, K., Kifer, D., Lawson, K., Shen, C., 2020. Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. Water Resour. Res. https://doi.org/10.1029/2020wr028095.

Feng, D., Lawson, K., Shen, C., 2021. Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. Geophys. Res. Lett. https://doi.org/10.1029/2021 gl092999.

Frame, J., F. Kratzert, D. Klotz, M. Gauch, G. Shelev, O. Gilon, L.M. Qualls, H.V. Gupta, and G.S. Nearing (2021), Deep learning rainfall-runoff predictions of extreme events, doi:10.5194/hess-2021-423.

Gao, C., Booij, M.J., Xu, Y.P., 2020. Assessment of extreme flows and uncertainty under climate change: disentangling the uncertainty contribution of representative

concentration pathways, global climate models and internal climate variability. Hydrol. Earth Syst. Sci. 24 (6), 3251–3269.
Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., Veeramachaneni, K., 2020. Tadgan: time series anomaly detection using generative adversarial networks. paper presented at. In: Proceedings of the 2020 IEEE International Conference on Big Data (Big Data). IEEE.
Gharib, A., Davies, E.G., 2021. A workflow to address pitfalls and challenges in applying machine learning models to hydrology. Adv. Water Res. 152, 103920.
Hagen, J.S., Leblois, E., Lawrence, D., Solomatine, D., Sorteberg, A., 2021. Identifying major drivers of daily streamflow from large-scale atmospheric circulation with machine learning. J. Hydrol. 596, 126086.
Han, D., Liu, P., Xie, K., Li, H., Xia, Q., Zhang, Y., Xia, J., 2023. An attention-based LSTM model for long-term runoff forecasting and factor recognition. Environ. Res. Lett.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Hunt, K.M.R., Matthews, G.R., Pappenberger, F., Prudhomme, C., 2022. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. Hydrol. Earth Syst. Sci. 26 (21), 5449–5472. https://doi.org/10.5194/hess-26-5449-2022.
Ivanov, V.Y., et al., 2021. Breaking down the computational barriers to real-time urban flood forecasting. Geophys. Res. Lett. https://doi.org/10.1029/2021 gl093585.
Kendall, A., and Y. Gal (2017), What Uncertainties Do We Need in Bayesian Deep Learning for Computer, paper presented at Thirsty-first Conference on Neural Information Processing Systems.
Kim, J., Ivanov, V.Y., Fatichi, S., 2015. Climate change and uncertainty assessment over a hydroclimatic transect of Michigan. Stochastic Environmental Research and Risk Assessment 30 (3), 923–944. https://doi.org/10.1007/s00477-015-1097-2.
Kim, J., Ivanov, V.Y., Fatichi, S., 2016a. Environmental stochasticity controls soil erosion variability. Sci. Rep. 6 (1), 22065. https://doi.org/10.1038/srep22065.
Kim, J., Ivanov, V.Y., Fatichi, S., 2016b. Soil erosion assessment-Mind the gap. Geophys. Res. Lett. 43 (24), 12. https://doi.org/10.1002/2016 gl071480, 446-412,456.
Kim, J., Tanveer, M.E., Bae, D.H., 2018. Quantifying climate internal variability using an hourly ensemble generator over South Korea. Stochastic Environmental Research and Risk Assessment 32 (11), 3037–3051. https://doi.org/10.1007/s00477-018-1607-0.
Kingma, D.P., and J. Ba (2014), Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980.
Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42 (3).
Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 26 (6), 1673–1693. https://doi.org/10.5194/hess-26-1673-2022.
Konapala, G., Kao, S.C., Painter, S.L., Lu, D., 2020. Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. Environ. Res. Lett. 15 (10), 104022 https://doi.org/10.1088/1748-9326/aba927.
Kotsiantis, S.B., I. Zaharakis, and P. Pintelas (2007), Supervised machine learning: a review of classification techniques, Emerging artificial intelligence applications in computer engineering, 160(1), 3–24.
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22 (11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019a. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55 (12), 11344–11354. https://doi.org/10.1029/2019WR026065.
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019b. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23 (12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.
Kumar, P., Foufoula-Georgiou, E., 1994. Wavelet analysis in geophysics: an introduction. Wavelets in geophysics 4, 1–43.
Lang, M., Ouarda, T.B.M.J., Bobée, B., 1999. Towards operational guidelines for over-threshold modeling. J. Hydrol. 225 (3), 103–117. https://doi.org/10.1016/S0022-1694(99)00167-5.
Le, M.H., Kim, H., Adam, S., Do, H.X., Beling, P., Lakshmi, V., 2022. Streamflow Estimation in Ungauged Regions using Machine Learning: quantifying Uncertainties in Geographic Extrapolation. Hydrol. Earth Syst. Sci. Discuss. 1–24.
Le, X.H., Ho, H.V., Lee, G., Jung, S., 2019. Application of Long Short-Term Memory (LSTM) neural network for flood forecasting. Water 11 (7), 1387.
Li, Y., Zhu, Z., Kong, D., Han, H., Zhao, Y., 2019. EA-LSTM: evolutionary attention-based LSTM for time series prediction. Knowledge-Based Systems 181, 104785. https://doi.org/10.1016/j.knosys.2019.05.028.
Liu, L., Liu, X., Bai, P., Liang, K., Liu, C., 2023a. Comparison of flood simulation capabilities of a hydrologic model and a machine learning model. Int. J. Climatol. 43 (1), 123–133.
Liu, S., Lu, D., Painter, S.L., Griffiths, N.A., Pierce, E.M., 2023b. Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. Front. Water 5, 1150126.
Liu, Y.H., Mehta, S., 2019. Hands-On Deep Learning Architectures with Python: Create deep Neural Networks to Solve Computational Problems Using TensorFlow and Keras. Packt Publishing Ltd.
Liu, Z., Zhou, P., Chen, G., Guo, L., 2014. Evaluating a coupled discrete wavelet transform and support vector regression for daily and monthly streamflow forecasting. J. Hydrol. 519, 2822–2831.
Lu, D., Konapala, G., Painter, S.L., Kao, S.C., Gangrade, S., 2021. Streamflow simulation in data-scarce basins using bayesian and physics-informed machine learning models. J. Hydrometeorol. 22 (6), 1421–1438. https://doi.org/10.1175/JHM-D-20-0082.1.
Luce, C. (2014), Runoff Prediction in Ungauged Basins: synthesis Across Processes, Places and Scales, Eos, Transactions American Geophysical Union, 95(2), 22-22, doi: https://doi.org/10.1002/2014EO020025.
Maheswaran, R., Khosa, R., 2012. Comparative study of different wavelets for hydrologic forecasting. Comput. Geosci. 46, 284–295. https://doi.org/10.1016/j.cageo.2011.12.015.
May, R.J., Maier, H.R., Dandy, G.C., Fernando, T.G., 2008. Non-linear variable selection for artificial neural networks using partial mutual information. Environ. Modell. Software 23 (10–11), 1312–1326.
McDermott, P., Wikle, C., 2019. Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. Entropy 21 (2), 184. https://doi.org/10.3390/e21020184.
Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity is dead: whither water management? Science 319 (5863), 573–574.
Moller, F., D. Botache, D. Huseljic, F. Heidecker, M. Bieshaar, and B. Sick (2021), Out-of-distribution detection and generation using soft brownian offset sampling and autoencoders, paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
edited by Moradkhani, H., Sorooshian, S., 2008. General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis. In: Sorooshian, S., Hsu, K.-L., Coppola, E., Tomassetti, B., Verdecchia, M., Visconti, G. (Eds.), Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrologic Models. Springer, Berlin, pp. 1–24. https://doi.org/10.1007/978-3-540-77843-1_1. edited by.
Nayak, P.C., Venkatesh, B., Krishna, B., Jain, S.K., 2013. Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach. J. Hydrol. 493, 57–67. https://doi.org/10.1016/j.jhydrol.2013.04.016.
Newman, A.J., et al., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19 (1), 209–223. https://doi.org/10.5194/hess-19-209-2015.
Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., Zhang, J., 2019. Streamflow and rainfall forecasting by two long short-term memory-based models. J. Hydrol., 124296 https://doi.org/10.1016/j.jhydrol.2019.124296.
Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., Zhang, J., 2020. Streamflow and rainfall forecasting by two long short-term memory-based models. J. Hydrol. 583, 124296 https://doi.org/10.1016/j.jhydrol.2019.124296.
Nourani, V., Baghanam, A.H., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. J. Hydrol. 514, 358–377.
Nourani, V., Komasi, M., Mano, A., 2009. A multivariate ANN-wavelet approach for rainfall–runoff modeling. Water Resour. Manage. 23 (14), 2877–2894. https://doi.org/10.1007/s11269-009-9414-5.
Olenskyj, A.G., Donis-González, I.R., Earles, J.M., Bornhorst, G.M., 2022. End-to-end prediction of uniaxial compression profiles of apples during in vitro digestion using time-series micro-computed tomography and deep learning. J. Food Eng. 325, 111014.
Percival, D.B., Walden, A.T., 2000. Wavelet Methods For Time Series Analysis. Cambridge university press.
Prein, A.F., Rasmussen, R.M., Ikeda, K., Liu, C., Clark, M.P., Holland, G.J., 2016. The future intensification of hourly precipitation extremes. Nat. Clim. Change 7 (1), 48–52. https://doi.org/10.1038/nclimate3168.
Prodhan, F.A., Zhang, J., Hasan, S.S., Sharma, T.P.P., Mohana, H.P., 2022. A review of machine learning methods for drought hazard monitoring and forecasting: current research trends, challenges, and future research directions. Environ. Modell. Software 149, 105327.
Quilty, J., Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. J. Hydrol. 563, 336–353. https://doi.org/10.1016/j.jhydrol.2018.05.003.
Quilty, J., Adamowski, J., Boucher, M.A., 2019. A Stochastic data-driven ensemble forecasting framework for water resources: a case study using ensemble members derived from a database of deterministic wavelet-based models. Water Resour. Res. 55 (1), 175–202. https://doi.org/10.1029/2018wr023205.
Quilty, J., Jahangir, M.S., You, J., Hughes, H., Hah, D., Tzoganakis, I., 2023. Bayesian extreme learning machines for hydrological prediction uncertainty. J. Hydrol. 626, 130138 https://doi.org/10.1016/j.jhydrol.2023.130138.
Rahmani, F., Shen, C., Oliver, S., Lawson, K., Appling, A., 2021. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. Hydrol. Processes 35 (11), e14400. https://doi.org/10.1002/hyp.14400.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204.
Ren, K., Fang, W., Qu, J., Zhang, X., Shi, X., 2020. Comparison of eight filter-based feature selection methods for monthly streamflow forecasting – Three case studies on CAMELS data sets. J. Hydrol. 586, 124897 https://doi.org/10.1016/j.jhydrol.2020.124897.

Renard, B., Lang, M., Bois, P., 2006. Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data. Stochastic environmental research and risk assessment 21 (2), 97–112.
Sang, Y.F., 2013. A review on the applications of wavelet transform in hydrology time series analysis. Atmos. Res. 122, 8–15. https://doi.org/10.1016/j.atmosres.2012.11.003.
Shensa, M.J., 1992. The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Trans. Signal Process. 40 (10), 2464–2482. https://doi.org/10.1109/78.157290.
Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 97, 105524.
Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., O'connell, P., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48 (6), 857–880.
Solari, S., Losada, M., 2012. A unified statistical model for hydrological variables including the selection of threshold for the peak over threshold method. Water Resour. Res. (10), 48.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Series B Stat. Methodol. 36 (2), 111–133.
Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.A., Rahman, A., 2020. Land-use land-cover classification by machine learning classifiers for satellite observations—A review. Remote. Sens. 12 (7), 1135.
Tang, S., Sun, F., Liu, W., Wang, H., Feng, Y., Li, Z., 2023. Optimal postprocessing strategies with LSTM for global streamflow prediction in ungauged basins. Water Resour. Res., e2022WR034352
Thanh, H.V., Binh, D.V., Kantoush, S.A., Nourani, V., Saber, M., Lee, K.K., Sumi, T., 2022. Reconstructing daily discharge in a megadelta using machine learning techniques. Water Resour. Res. 58 (5), e2021WR031048.
Thornton, M., R. Shrestha, Y. Wei, P. Thornton, S. Kao, and B. Wilson (1840), Daymet: daily surface weather data on a 1-km grid for North America, Version 4. ORNL DAAC, Oak Ridge, Tennessee, USA, edited.
Todini, E., 2007. Hydrological catchment modelling: past, present and future. Hydrol. Earth Syst. Sci. 11 (1), 468–482. https://doi.org/10.5194/hess-11-468-2007.
Tran, T.D., Tran, V.N., Kim, J., 2021. Improving the accuracy of dam inflow predictions using a long short-term memory network coupled with wavelet transform and predictor selection. Mathematics 9 (5), 551. https://doi.org/10.3390/math9050551.
Tran, V.N., Dwelle, M.C., Sargsyan, K., Ivanov, V.Y., Kim, J., 2020. A novel modeling framework for computationally efficient and accurate real-time ensemble flood forecasting with uncertainty quantification. Water Resour. Res. https://doi.org/10.1029/2019WR025727.
Tran, V.N., V.Y. Ivanov, and J. Kim (2023a), Streamflow Predictions using Machine Learning with Data Reformation, edited, Zenodo, doi:https://doi.org/10.5281/zenodo.8309631.
Tran, V.N., Ivanov, V.Y., Xu, D., Kim, J., 2023b. Closing in on hydrologic predictive accuracy: combining the strengths of high-fidelity and physics-agnostic models. Geophys. Res. Lett. 50 (17), e2023GL104464 https://doi.org/10.1029/2023GL104464.
Tran, V.N., Kim, J., 2022. Robust and efficient uncertainty quantification for extreme events that deviate significantly from the training dataset using polynomial chaos-kriging. J. Hydrol., 127716 https://doi.org/10.1016/j.jhydrol.2022.127716.
Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin (2017), Attention is all you need, Advances in Neural Information Processing Systems, 30.
Vaze, J., Chiew, F., Hughes, D., Andréassian, V., 2015. Preface: hs02–hydrologic non-stationarity and extrapolating models to predict the future. Proc. Int. Assoc. Hydrol. Sci. 371, 1–2.
Venkata Ramana, R., Krishna, B., Kumar, S.R., Pandey, N.G., 2013. Monthly rainfall prediction using wavelet neural network analysis. Water Resour. Manage. 27 (10), 3697–3711. https://doi.org/10.1007/s11269-013-0374-4.
Wang, Y., M. Huang, X. Zhu, and L. Zhao (2016), Attention-based LSTM for aspect-level sentiment classification, paper presented at Proceedings of the 2016 conference on empirical methods in natural language processing.
Wilbrand, K., Taormina, R., ten Veldhuis, M.C., Visser, M., Hrachowitz, M., Nuttall, J., Dahm, R., 2023. Predicting streamflow with LSTM networks using global datasets. Front. Water 5, 1166124.
Wilder, J.W., 1978. New Concepts in Technical Trading Systems. Trend Research.
Wong, T.T., Yeh, P.Y., 2019. Reliable accuracy estimates from k-fold cross validation. IEEE Trans. Knowl. Data Eng. 32 (8), 1586–1594.
Xiang, Z., Demir, I., 2020. Distributed long-term hourly streamflow predictions using deep learning–A case study for State of Iowa. Environ. Modell. Software 131, 104761.
Xu, T., Liang, F., 2021. Machine learning for hydrologic sciences: an introductory overview. Wiley Interdisciplinary Rev. 8 (5), e1533.
Xu, W., Chen, J., Zhang, X.J., Xiong, L., Chen, H., 2022. A framework of integrating heterogeneous data sources for monthly streamflow prediction using a state-of-the-art deep learning model. J. Hydrol. 614, 128599.
Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415, 295–316.
Yeung, A.Y., Roewer-Despres, F., Rosella, L., Rudzicz, F., 2021. Machine learning–based prediction of growth in confirmed COVID-19 infection cases in 114 countries using metrics of nonpharmaceutical interventions and cultural dimensions: model development and validation. J. Med. Internet Res. 23 (4), e26628.
Yu, Q., Jiang, L., Wang, Y., Liu, J., 2023. Enhancing streamflow simulation using hybridized machine learning models in a semi-arid basin of the Chinese loess Plateau. J. Hydrol. 617, 129115.
Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115.
Zhong, L., Lei, H., Gao, B., 2023. Developing a physics-informed deep learning model to simulate runoff response to climate change in alpine catchments. Water Resour. Res. 59 (6), e2022WR034118.
Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., Hoffman, M.M., 2019. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Info. Fusion 50, 71–91.