Advances in Water Resources 152 (2021) 103920

Contents lists available at ScienceDirect

Advances in Water Resources

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

A workflow to address pitfalls and challenges in applying machine learning models to hydrology

Amr Gharib, Evan G.R. Davies*

Civil and Environmental Engineering Department, University of Alberta, Canada

ARTICLE INFO

Keywords:
Machine learning
Data driven modeling
Model workflow
Variable selection
Seasonal streamflow forecasting
Forecast lead time

ABSTRACT

Data-driven modeling with machine learning (ML) algorithms in hydrologic modeling and forecasting suffers a number of pitfalls and challenges, including variable selection bias, resubstitution validation, inconsistent validation processes for different algorithms and model selection using the test set. They lead to incorrect model development and biased, overly optimistic performance estimates – and thus, unreliable models. This study presents a novel model building and testing workflow that addresses common machine learning (ML) challenges and pitfalls. The presented workflow incorporates optional variable transformation and preprocessing techniques, and applies to various ML model types, variable selection algorithms and resampling techniques. We demonstrate its performance through streamflow forecasting for the Bow River Basin, Alberta, Canada, using four conventional ML algorithms (ANN, SVM, ELM and RBF networks) driven by local hydrometeorological conditions and large-scale climate indices. Using the cross-validation average out-of-sample results and a separate test set, the prediction accuracy estimate bias (the relative difference between the model performance estimated using the validation sets and a separate test set) was empirically estimated to be 5.6%, 4.4%, 2.5% and 3.0% for the Seasonal, May, June and July models, respectively. In addition, the streamflow forecasting models had an average coefficient of determination of 0.85. Preprocessing and dimensionality reduction through principal component analysis (PCA) was detrimental to prediction accuracy. Snow water equivalent from individual snow courses proved the most important predictor for Bow River streamflow, while global climate indices, including the PDO, AMO and PNA, increased the Nash–Sutcliffe efficiency by 6% to 50%. Finally, although forecasting skill decreased with increasing forecast lead time, satisfactory forecasts (NSE>0.5) could be obtained two months in advance of the spring melt, at the end of February. Extensions of this study should address the tendency of different variable selection algorithms to pick irrelevant or redundant predictors after changes in training data.

1. Introduction

Data-driven modeling with machine learning (ML) algorithms is one of the most rapidly developing topics in the field of hydrologic modeling and forecasting. In addition to streamflow forecasting, ML algorithms have been applied to areas such as urban water demand forecasting, estimation of flow duration curves at ungauged sites and rainfall-runoff modeling (Schmidt et al., 2020). However, both model development and performance estimation of data driven ML models for such applications suffer a number of challenges and pitfalls (Quilty and Adamowski, 2018; Schmidt et al., 2020; Solomatine and Ostfeld, 2008; Wu et al., 2013; Zhang et al., 2015; Zheng et al., 2018). Such common challenges and pitfalls include variable selection bias, resubstitution validation and inconsistent validation processes and resamples for different algorithms, as well as model selection by the test set.

A key challenge facing ML applications is variable selection bias, which results in the selection of irrelevant or redundant predictors. Selection bias occurs when the same data are used both to create a model from a training set and to select the associated input predictors. In practice, this means that the variable selection process was not considered part of the model building process (Kuhn and Johnson, 2013; Liu and Motoda, 2007; Quilty and Adamowski, 2020; Schmidt et al., 2020). It leads to improved performance for the same data set when compared with model development without selection bias; however, the results are sensitive and repeating the same variable selection process with a slightly different training set can lead to different results, because the uncertainty from the variable selection process can be far greater than the uncertainty of the model (Ambroise and McLachlan, 2002; Quilty and Adamowski, 2020). Even if the best predictors are selected, the performance estimates from this process are biased because they do not reflect the uncertainty involved in the variable selection process and its impact on this result (Kuhn and Johnson, 2013; Quilty et al., 2019; Quilty and Adamowski, 2020). Short historical records, large numbers of predictors and complex and powerful ML models increase the likeli-

* Corresponding author.
E-mail address: evan.davies@ualberta.ca (E.G.R. Davies).

https://doi.org/10.1016/j.advwatres.2021.103920
Received 15 November 2020; Received in revised form 12 April 2021; Accepted 14 April 2021
Available online 18 April 2021

hood of selection bias, and all are commonly encountered in applying ML algorithms to hydrologic modeling and forecasting.

Another common pitfall involves testing a model with the same data used to train it, which is called resubstitution validation (Raschka, 2020; Wu et al., 2013; Zheng et al., 2018). This second pitfall produces an overly optimistic performance estimation due to overfitting (Solomatine and Ostfeld, 2008; Wu et al., 2013), which causes a model to match the training data too well, such that it applies more poorly to "unseen data", or data that was not included in the training set. To train the model properly and avoid overfitting, an additional validation dataset is required. A popular approach uses k-fold cross-validation to build the model and ensure it generalizes to unseen data. However, using cross-validation to estimate the model performance could lead to biased estimates. The cross validation estimate is biased since the model was already exposed to the data in the validation set, which guided the selection of the variables, algorithm, model structure and hyperparameter values during model training and validation (Raschka, 2020; Solomatine and Ostfeld, 2008). If a test set cannot be used because of limited historical records, performance estimates from the cross-validation only are likely overly optimistic in terms of prediction accuracy. However, for comparing among a set of models and selecting the best model, the biased, optimistic estimates from cross-validation can be used, given that the performance estimations of all models are affected evenly (Raschka, 2020).

A third pitfall involves the use of different validation processes and resamples for different algorithms. Both cross-validation and resampling processes must be consistent across the different algorithms under consideration, in order to ensure the performance estimation bias affects them equally and that the model behaviors and error ranges are comparable between different algorithms. In partitioning data between training and validation datasets, any data transformation or preprocessing must be applied as part of the cross-validation process and not performed independently in advance (Hastie et al., 2009; Quilty and Adamowski, 2018), since it could otherwise 1) leak information about the mean and the variance of the validation set to the training data, if the full training/validation dataset was preprocessed independently in advance or 2) compromise the test data, if the entire dataset is used. Such leaks influence both model selection and the estimation of the model's generalization performance and prediction accuracy on unseen data. A fourth, related pitfall is using the test set for algorithm selection or for optimizing the model hyperparameters, which are the settings or tuning parameters of a machine learning algorithm that aid the control of algorithm behavior and dictate the flexibility or complexity of a model (Quilty and Adamowski, 2020; Raschka, 2020; Solomatine and Ostfeld, 2008). Choosing a model based on the test set leaks information from the test set to the model and produces an optimistic or biased estimate of the model performance. Test sets must be used only to estimate model performance. Because cross-validation results give a more robust perspective of model generalization to new datasets, they should be used instead for making model choices and final model selection.

Falling into the traps identified above in building data-driven ML models and in estimating their performance can lead to less interpretable ML models, since they may then include irrelevant or redundant predictors in the model. Worse, failing to avoid the pitfalls can lead to poorer generalization performance on unseen data and a biased estimate of the developed model's future prediction accuracy – and thus uncertain, unreliable model results.

This study therefore presents a novel model building and testing workflow with an appropriate resampling and testing process, which 1) addresses the impact of uncertainty involved in the input variable selection on the bias in the estimated prediction performance, while considering the uncertainty in the model structure and input data; 2) addresses the variable selection bias by including the variable selection process in model building; 3) prevents falling into the other pits identified above and 4) includes optional variable transformation and preprocessing techniques. Recent studies have identified the need to consider both input variable selection and model structure decisions within the model building framework to ensure robustness of the results (Kim et al., 2020; Quilty et al., 2019; Quilty and Adamowski, 2020; Schmidt et al., 2020). To the best of our knowledge, this is the first paper in the field of hydrology to present a single, systematic workflow to address the uncertainty of input variable selection, model structure and input data simultaneously.

To demonstrate the performance of the proposed workflow, we developed a streamflow forecasting model for the cordilleran and snowmelt-dominated Bow River Basin in Alberta, Canada, by using ML algorithms with a combination of local hydrometeorological conditions and large-scale climate indices. The streamflow forecasts, at monthly to seasonal time scales, are themselves essential for optimal planning of water resources, particularly for seasonal reservoir operation and planning. Such forecasts can improve water use efficiency and provide early drought and flood warning (Danandeh Mehr et al., 2014). Their importance is only rising as climate change intensifies the hydrological cycle and causes more frequent and hazardous flood and drought events (National Academies of Sciences Engineering and Medicine, 2016; Schmidt et al., 2020).

Reliable forecasts are inherently complex and challenging because streamflow is generated by various processes with stochastic, non-linear and non-stationary characteristics (El-Shafie et al., 2009). There are two general categories of streamflow forecasting models: data-driven models and process-based models. Data-driven models are more practical than process-based models in cases where precise runoff forecasts are more important than an understanding of physical catchment processes (Liu et al., 2014). Such models are based on establishing statistical inference between inputs, in the form of antecedent conditions such as observed flow, snowpack and precipitation as well as large-scale climate signals, and seasonal or monthly streamflow over forecasting time horizon (Singh, 2016). We applied and tested four different ML algorithms: artificial neural networks (ANN), support vector machines (SVM), extreme learning machines (ELM) and radial basis function networks (RBF). Further, multiple linear regression (MLR) is also included as a simple and easy-to-understand reference model for comparison against the relatively more complex ML algorithms. Note that we have selected these particular ML algorithms because more complicated approaches are not suitable for our limited input data and short historical records – reliable results from a data-driven model require ten times the number of observations as the degrees of freedom (Abu-Mostafa et al., 2012). Because of its large set of potential input variables and relatively short historical record of snowpack values, which are one of the most important streamflow predictors in the area, the Bow River Basin provides a highly relevant case study location. Our focus is summer season streamflow forecasting to match seasonal irrigation needs in Southern Alberta, which accounts for over 70% of all irrigated land in Canada. After applying the presented workflow to our example, we then further analyzed the trained models to 1) select the best ML algorithm based on its generalization performance and compare the performance of nonlinear ML models to MLR; 2) test the impact of reducing dimensionality by principal component analysis (PCA) as an optional preprocessing technique; 3) determine the key predictors for streamflow in the Bow River Basin and 4) investigate the impact of increasing the lead time on prediction accuracy.

The remainder of this paper is organized as follows. In the next section we review the related ML algorithm applications in the literature, input variables along with their links to streamflow variability and streamflow drivers in the study area. Section 3 gives background on the study area. Section 4 presents the workflow and methodological details for data transformation, preprocessing, as well as variable selection and ML algorithms, hyperparameter tuning, prediction accuracy estimation and other evaluation measures. Empirical results and discussion are in Section 5, while summary and concluding remarks are provided in Section 6.

2. Background

This section reviews previous studies on input variable selection and model structure uncertainties as well as applications of the machine learning algorithms to streamflow forecasting. In addition, it explores local hydrometeorological variables and global climate indices that are associated with the variability in streamflow and/or the variability in other local variables that contribute to streamflow generation, such as temperature, precipitation and snow cover.

2.1. Uncertainty in input variable selection and model structure

Input variable selection is critical to the development of reliable data-driven models (Galelli et al., 2014; Kim et al., 2020; Quilty et al., 2016; Quilty and Adamowski, 2020). Recent studies show that careful variable selection can improve the model performance and that its impact on the performance is problem dependent (Quilty and Adamowski, 2020; Tyralis and Papacharalampous, 2017). The uncertainty in input variable selection stems from the dependence of the variable selection algorithm on both the input data and the model structure used to select the variables.

Quilty and Adamowski (2018) proposed the Wavelet Data-Driven Forecasting Framework to address the incorrect use of wavelet-based models when applied to real-world forecasting problems. Quilty et al. (2019) developed the Stochastic Data-Driven Forecasting Framework (SDDFF), its single wavelet-based counterparts (SWDDFF) and ensemble multiwavelet-based version (EW-SDDFF), which addressed input data, input variable selection, parameter, and model output uncertainty in the respective models. Finally, Quilty and Adamowski (2020) explored SWDDFF and SDDFF by varying the level of uncertainty in the two approaches: 1) none, 2) model parameters, 3) model parameters and output and 4) input variable selection, as well as model parameters and output. The inclusion of wavelet decomposition and input variable selection uncertainty led to significant gains in forecast accuracy and reliability compared to the alternatives. Kim et al. (2020) proposed an approach to deal with uncertainty in model structure and input variable selection using a modified backward elimination procedure for a ANN model; this approach generates a number of input layers where each layer represents a set of the potential input variables.

Our study addresses a gap in the literature by proposing a workflow that handles uncertainties in input variable selection, model structure and input data, and avoids introducing other sources of potential performance estimation bias.

2.2. Brief review of machine learning algorithms

Numerous data-driven conventional and machine learning models have been applied to streamflow forecasting. For decades, conventional time series models such as Linear Regression (LR), Multiple Linear Regression (MLR) and Auto-Regressive Integrated Moving Average (ARIMA) have been used to forecast streamflows; however, they are unable to capture the non-linearity of the streamflow response to the input variables. Hydrologists have therefore used machine learning techniques over the past two decades as flexible and robust techniques that overcome the drawbacks of conventional models and incorporate the inherent nonlinearity of hydrological datasets without their explicit definition (El-Shafie et al., 2009; He et al., 2014; Humphrey et al., 2016; Kuhn and Johnson, 2013; Maier and Dandy, 2000; Rieker and Labadie, 2012; Tan et al., 2018; Yaseen et al., 2016c).

Artificial neural networks (ANN), support vector machines (SVM) and radial basis function networks (RBF) are the most widely used ML algorithms for streamflow modeling and forecasting (Yaseen et al., 2015). Shabri and Suhartono (2012) compared the performance of SVM, ANN and ARIMA models for the Kinta River in Perak, Peninsular Malaysia, and found the SVM model outperformed the others. Rasouli et al. (2012) found that SVM outperformed MLR. Lima et al. (2016) compared linear to nonlinear approaches and found that nonlinear flow forecasting models for Englishman River, Georgia, USA, and Stave River, BC, Canada had better skills than the linear models. Ghorbani et al. (2016) compared SVM, ANN and MLR models for Big Cypress River, Texas, USA, and reported that SVM and ANN performed better than MLR in estimating peak discharge values, which MLR tended to underestimate. They noted that ANN was flexible but that its training was time consuming. Kalra et al. (2013a,b) built SVM, ANN and MLR models to forecast spring-summer streamflow for North Platte River, USA; the SVM model outperformed the others for all tested cases. The same results were achieved for an SVM model for the Upper Colorado River Basin, USA (Kalra and Ahmad, 2009). Carrier et al. (2013) found better results for an SVM model than MLR model forecasts in the western United States. He et al. (2014) examined ANN, SVM and Adaptive Neural-Fuzzy Inference System (ANFIS) approaches for forecasting river flow in a semiarid environment, and similarly concluded that the SVM model outperformed the others. Yaseen et al. (2015) explored the literature on machine learning approaches in streamflow forecasting, and concluded that SVM produces accurate and robust classification results for forecasting streamflow and that it copes well with noisy data. Yaseen et al. (2016b) demonstrated superior performance of RBF over ANN. Kisi and Cigizoglu (2007) found that RBF was superior to ANN and GRNN for both short-term and long-term river streamflow forecasting.

In addition to ANN, SVM and RBF, extreme learning machines (ELM) offer excellent modeling capabilities, including a simple neural network structure, randomly assigned weights and a fast learning process (Yaseen et al., 2016a), and can produce similar results to SVM (Huang et al., 2012). ELM has also been applied in hydrological studies. Deo and Şahin (2015) applied ELM to model the monthly effective drought index. They demonstrated a significant improvement in forecasting performance over ANN, as well as a 32-times faster training speed. Yadav et al. (2016) found comparable performance among Online Sequential ELM, SVM and ANN, and that the ELM model had the minimum prediction error. Deo and Şahin (2016) found the ELM model for streamflow forecasting in Queensland, Australia, provided superior results to ANN models.

Importantly, all the studies described above found that nonlinear models outperformed linear models, and that the latter tended to underestimate peak values. Further, although several studies found that ELM, RBF and SVM outperformed ANN, no single model outperformed the others in all cases. Study conclusions from different locations can depend on the associated catchment hydrology and the nature of input data used, so it is important to try various models and select the best model for each case. Therefore, ANN, RBF, SVM and ELM, along with MLR, were investigated in this study.

2.3. Input variables

To develop a streamflow forecasting model, all possible predictors of streamflow generation processes must be considered for a given study area, especially in mountainous areas with high variability. For a data-driven forecasting model, the local hydrometeorological variables typically considered are streamflow (Behzad et al., 2009; Jain and Kumar, 2007; Kisi and Cimen, 2011; Noori et al., 2011; Nourani et al., 2011), precipitation (Behzad et al., 2009; Li et al., 2010; Noori et al., 2011; Nourani et al., 2011), temperature (Behzad et al., 2009; Noori et al., 2011) and solar radiation (Noori et al., 2011). Previous studies have also found that the use of large-scale climate indices as predictors improves streamflow forecasting skill, in combination with the local hydrometeorological observations. For example, Rasouli et al. (2012) reported best results for a daily streamflow forecasting model for a basin in British Columbia, Canada, that mixed local observations and climate indices, especially at longer lead times of 5-7 days, where the lead time represented the lag between release of the streamflow prediction and observation of the corresponding streamflow.

Makkeasorn et al. (2008) argued that large-scale climate indices incorporating sea surface temperature anomalies may capture some climate change impacts and improve the forecasting accuracy.

Because climate patterns are inter-connected and no single climate index explains all climatic variability within a river basin, climate indices are typically used in groups (Kalra et al., 2013b). Various large-scale climate indices are associated with snowpack, air temperature and precipitation variability during winter and summer seasons in North America, and, in particular, western Canada. The Pacific Decadal Oscillation (PDO), a decadal-scale oscillation characterized by the leading principal component of the North Pacific SST anomalies (Mantua et al., 1997), influences runoff timing in Canada (Burn, 2008). Gobena and Gan (2006) found that interdecadal streamflow signals were in good agreement with the PDO index, and that their variations dominated streamflow variability in the Bow River Basin. In the summer, positive PDO phases lead to drier conditions and a higher frequency of low streamflow events (Bonsal and Shabbar, 2011), and to below-normal winter precipitation and above-normal winter temperatures in western Canada, while negative PDO phases have opposite effects (Bonsal et al., 2001). The El Niño-Southern Oscillation (ENSO) represents the prominent interannual variability produced by equatorial Pacific sea surface temperature (SST) anomalies (Hsieh and Tang, 2001). During the summer, Shabbar and Skinner (2004) found that El Niño events led to drier conditions, while La Niña events conversely produced excess moisture over western Canada. Likewise, El Niño events have been shown to produce warmer winter weather, slightly higher than normal snowfall and increased spring streamflows, while La Niña produces the opposite effects, which are more concentrated in western Canada (Hsieh and Tang, 2001). The traditional index for ENSO in the Pacific Ocean is called the Southern Oscillation Index (SOI), which measures the difference in sea level pressure occurring between Tahiti and Darwin during El Niño or La Niña events (Ladd and Gajewski, 2009). Several other indices related to ENSO also monitor the SST anomalies averaged across a given region in the tropical Pacific: Niño1+2, Niño3, Niño3.4 and Niño4 (Gobena and Gan, 2009; Moradi et al., 2020). The Pacific-North American pattern (PNA) relates temperature and precipitation anomalies across North America to conditions over the Pacific Ocean, and is strongly related to snow cover variability over North America (Brown and Goodison, 1996). A higher PNA index is related to lower snow cover, and a positive PNA is associated with above-average temperatures over western Canada. In Alberta, positive PNA is associated with below average winter precipitation and reduced snow cover (Bonsal and Shabbar, 2011; Brown and Goodison, 1996). PNA effects on streamflow vary regionally over southwestern Canada (Gobena and Gan, 2009). Finally, the Atlantic Multidecadal Oscillation (AMO) represents SST variability in the North Atlantic Ocean. The positive phase of the winter AMO tends to be associated with dry summer conditions over the central and northern regions of the Canadian Prairies (Bonsal and Shabbar, 2011).

3. Study area

The Bow River watercourse is 587 km long stretching from the Bow Glacier in the Canadian Rockies eastward through the City of Calgary to its confluence with the Oldman River west of Medicine Hat, Alberta, and drains an area of 25,300 km^2. The annual precipitation exceeds 700 mm in its mountainous headwaters, but then decreases significantly eastward to give 330 mm of mean annual precipitation. The Bow has a higher natural flow yield than its three surrounding subbasins because of the large portion of its basin within the Rockies. Its streamflow typically peaks during the spring-summer melt freshet. However, variable topography, low average annual precipitation (600 mm) and high winds, particularly in the mountainous portion of the basin that generates most of the runoff, render the basin prone to water shortages and emphasize the importance of reliable runoff forecasts to support water managers' decisions.

The Bow River Basin is important for irrigation. Of the total water allocations in the Bow River Basin, 77% or the total sum of water license of 1,160,000 dam^3/year is allocated to three irrigation districts that comprise a total of 259,600 hectares, while about 10% is allocated to the city of Calgary (Alberta Agriculture and Forestry, 2020). Fig. 1 shows the study location. The Bow River streamflow in the summer months clearly exhibits high natural variability (Fig. 2).

Water managers in Alberta currently use Rocky Mountain snowpack survey values, reservoir winter storage levels, modeled soil moisture content in the spring and normal seasonal rainfall volumes and temperatures to estimate the water supply and to plan for the next irrigation season. However, the forecasts are uncertain and relying on them for planning poses risks for irrigators and other users (Jean and Davies, 2016). As shown below, machine learning models can significantly improve streamflow forecasting accuracy.

4. Methods

To address the variable selection bias and other pitfalls identified above, we introduce here a novel workflow with five main stages:

1. Data collection and preparation,
2. Data partitioning for training and validation, and for testing and performance evaluation,
3. Model building, with optional data transformation and preprocessing, variable selection and hyperparameter tuning,
4. Model training with the best identified ML algorithm, selected variables and tuned hyperparameters on the entire training-validation set, and
5. Final model testing using the test dataset.

The workflow features two resampling or cross-validation loops, an outer loop for application of the variable selection algorithm to identify the predictor set with the minimum residual sum of squares (RSS), and an inner loop for application of the exhaustive search, or grid search, algorithm to identify the optimal hyperparameters corresponding to the minimum RSS. Note that the workflow can accommodate other hyperparameter tuning algorithms, such as random search algorithms. Fig. 3 shows the complete model building and testing workflow, which is flexible and can be applied to any ML algorithm, variable selection algorithm and resampling method. Sections of Fig. 3 associated with the five stages of the workflow listed above are indicated in the figure, while further details on these stages are provided below.

4.1. Data collection and preparation

The first stage of model building is data collection for both local hydro-meteorological variables and large-scale climate indices. For the Bow River Basin, local-scale historical records of monthly average precipitation, minimum, maximum and average temperatures, snow water equivalent, average humidity and estimated solar radiation data, as well as streamflow data generated as monthly-averaged naturalized historical flow records, were collected for the Upper Bow River watershed in southern Alberta. Data from mountainous, upstream areas of the Bow River Basin is critical because a substantial portion of the runoff comes from this area. However, relatively fewer stations exist in this mountainous area compared to the foothills, and records are short and sometimes incomplete. Of the 21 active snow courses in the area, nine records begin in 1970, while the remaining stations date from 1981 (with some missing records).

Snow water equivalent (SWE) data were extracted from snow course data obtained from Alberta Environment and Parks (Darcy Talma, personal communication 2020). Monthly average SWE for individual stations with 20 years of data or more were used.

Township data from Alberta Agriculture and Forestry (AAF) was used for historical precipitation, temperature, humidity and estimated

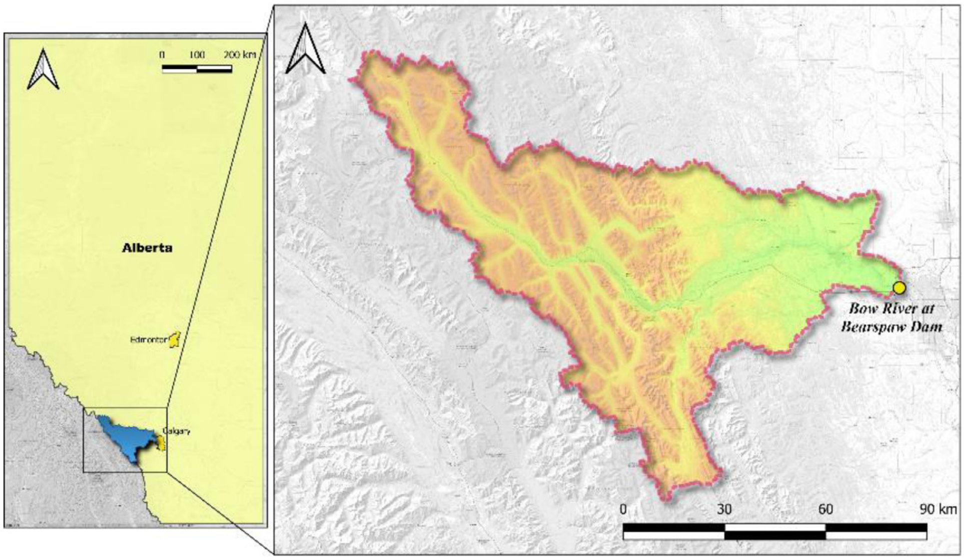

Fig. 1. Location of the Bow River Basin upstream of Bearspaw Dam surrounded by red dashed border

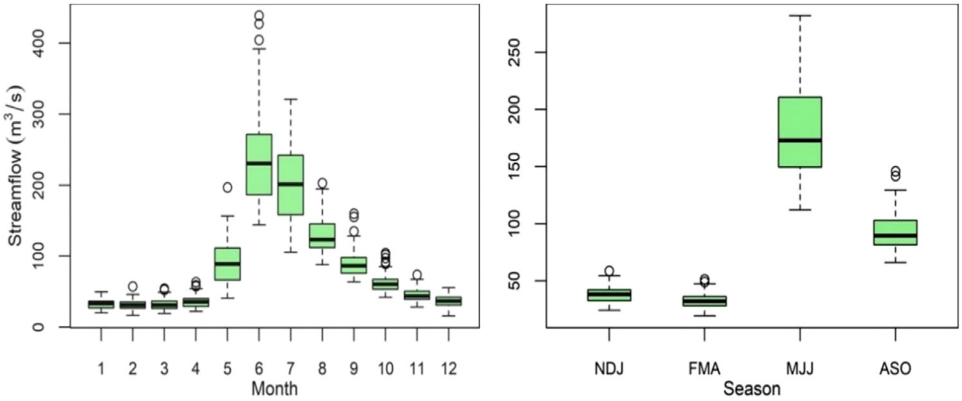

Fig. 2. Variability in monthly and seasonal streamflow values. For each box, the mid-line is the median value and the top and bottom lines are the 1st and the 3rd quartiles, respectively. The whiskers extend to the minimum of the most extreme streamflow or 1.5 times the difference between the 1st and 3rd quartiles, or the interquartile range.

Table 1
List of potential local hydrometeorological input variables.

Variable	Record Start Year	No. stations	Frequency	Type	Source
Naturalized historical streamflow of the Bow River at Bearspaw Dam	1920	1	Weekly	Full set of annual values	Alberta Environment & Parks and Ilich et al. (2018)
Snow Water Equivalent (SWE)	1970	21	Monthly	October – April values	Alberta Environment & Parks
Precipitation	1956	50	Monthly	Full set of annual values	Alberta Agriculture & Forestry
Average daily maximum and minimum temperatures, and monthly maximum and minimum temperatures	1956				
Estimated solar radiation	1956				
Average humidity	1956				

solar radiation records (https://agriculture.alberta.ca/acis/alberta-weather-data-viewer.jsp). AAF interpolates and averages township data from up to eight nearby meteorological stations for each township in Alberta. For the 50 townships within the basin boundaries, the interpolated data were downloaded and averaged over the basin. Table 1 lists the local hydrometeorological input variables used in the study and their sources.

Eleven potential large-scale climate indices were also tested as potential input variables. Five of the selected indices represent the El Niño-Southern Oscillation (ENSO): Niño1+2, Niño3, Niño3.4 and Niño4, and the Southern Oscillation Index (SOI). The other six indices include the Atlantic Multidecadal Oscillation (AMO), the Arctic Oscillation (AO), the North Atlantic Oscillation (NAO), the North Pacific Index (NP), the Pacific Decadal Oscillation (PDO) and the Pacific-North American pattern (PNA). Because previous studies found that using new indices constructed by spatially averaging climate data offered no obvious advantage (Gobena and Gan, 2009), we used only standard climate indices available in the literature. Table 2 lists the eleven large-scale predictors and their sources. Monthly time series of the abovementioned indices were retrieved from the listed websites.

After collection of all the data, local and global predictor variables were grouped into three different input datasets for forecasting seasonal runoff: S1 included only climate index predictors from 1950 to 2019 (i.e. no local predictor variables); S2 included a combination of local observations and climate indices from 1970 to 2019; and S3 built on S2, with more comprehensive and complete SWE records but a shorter record of 1981 to 2019. Table 3 summarizes the statistical characteristics of the summer streamflow for the full period of the input dataset.

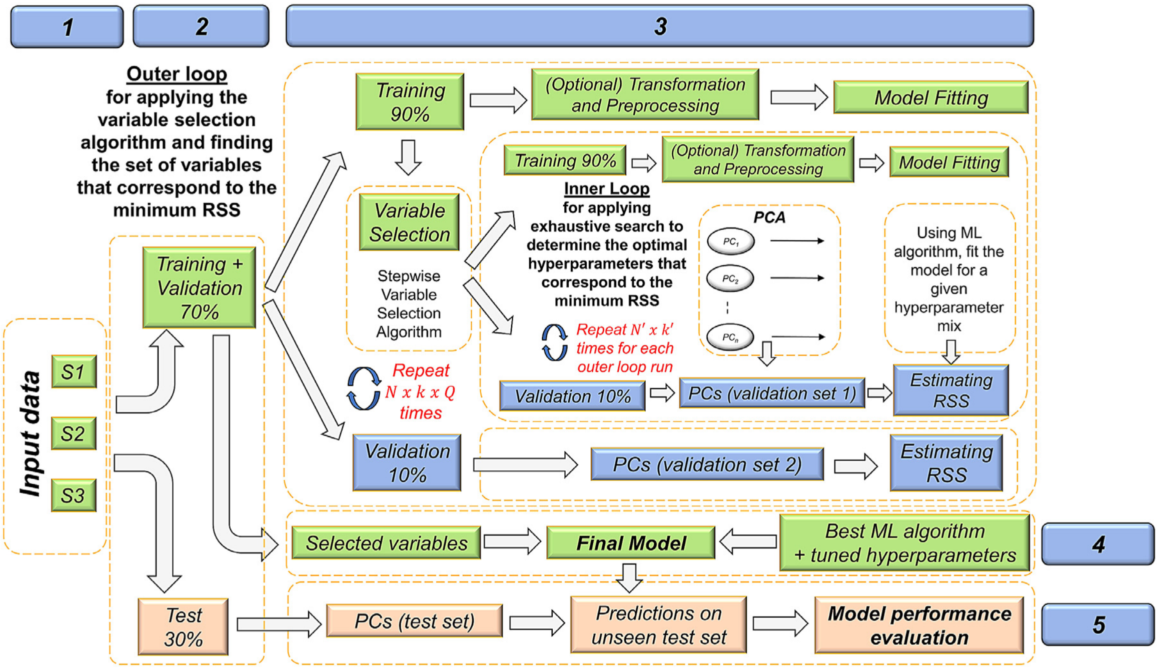
1
2
3
Outer loop
for applying the variable selection algorithm and finding the set of variables that correspond to the minimum RSS
Input data
S1
S2
S3
Training + Validation 70%
Test 30%
Training 90%
(Optional) Transformation and Preprocessing
Model Fitting
Variable Selection
Stepwise Variable Selection Algorithm
Repeat N x k x Q times
Training 90%
(Optional) Transformation and Preprocessing
Model Fitting
Inner Loop
for applying exhaustive search to determine the optimal hyperparameters that correspond to the minimum RSS
PCA
PC1
PC2
PCn
Using ML algorithm, fit the model for a given hyperparameter mix
Repeat N′ x k′ times for each outer loop run
Validation 10%
PCs (validation set 1)
Estimating RSS
Validation 10%
PCs (validation set 2)
Estimating RSS
Selected variables
Final Model
Best ML algorithm + tuned hyperparameters
4
PCs (test set)
Predictions on unseen test set
Model performance evaluation
5

Fig. 3. Flowchart of the complete model building and testing workflow with proper resampling using two loops, an inner loop for the hyperparameter optimization and an outer loop for variable selection. S1, S2, S3 are three partitions of the input dataset considered in the study. PCA denotes the customary principal component analysis, PC denotes principal component, and RSS denotes the residual sum of squares. **N, k**, N′, **k′**, and **Q** denote the number of repetitions and number of folds for the outer loop, the number of repetitions and number of folds for the inner loop, and the number of times the variable selection algorithm is run until stopping, respectively.

Table 2
List of potential large-scale climate indices.

Variable	Record Start Year	Source	Website
Niño 1+2, Niño 3, Niño 3.4, Niño 4	1870	Earth System Research Lab of National Oceanic and Atmospheric Administration (NOAA)	www.esrl.noaa.gov/psd/gcos_wgsb/Timeseries/
Pacific/North American pattern (PNA)	1950		https://www.esrl.noaa.gov/psd/data/correlation/pna.data
Atlantic Multidecadal Oscillation (AMO)	1856		http://www.esrl.noaa.gov/psd/data/timeseries/AMO/
North Pacific Index (NP)	1899		www.esrl.noaa.gov/psd/gcos_wgsb/Timeseries/
Arctic Oscillation (AO)	1950		www.esrl.noaa.gov/psd/gcos_wgsb/Timeseries/
Pacific Decadal Oscillation (PDO)	1854	National Climate Data Center of NOAA	https://www.ncdc.noaa.gov/teleconnections/pdo/data.csv
Southern Oscillation Index (SOI)	1866	Climate Research Unit - University of East Anglia	https://crudata.uea.ac.uk/cru/data/soi/soi_3dp.dat
North Atlantic Oscillation (NAO)	1821		https://crudata.uea.ac.uk/cru/data/nao/nao_3dp.dat

Table 3
Descriptive statistics for Bow River summer season streamflow.

Partition	Period	No. input variables (*)	Q (m^3/s)				
			Mean	Standard Deviation	Median	Minimum	Maximum
S1	1951-2019	133	186.0	40.31	178.5	109.9	294.5
S2	1970-2019	234	183.8	43.26	175.9	109.9	294.5
S3	1981-2019	239	187.9	44.73	176.7	125.6	294.5

(*) Number of input variables for data partition S1 = 11 monthly global climate indices (as listed in Table 2) * 12 months + 1 average annual SOI. Partition S2 = S1 input variables + 17 monthly snow course values + 7 other local hydrometeorological variables (as listed in Table 1) * 12 months, and S3 = S2 input variables + 5 additional monthly snow course values.

4.2. Data partitioning training/validation and testing sets

Before any transformation or preprocessing of the input data, the entire dataset was partitioned into training and test sets. The ratio of the test set to the entire dataset is (somewhat) arbitrary and depends on the number of observations in the input dataset and the type of application. Typical values range from 20% to 40%. We used 30% of the entire dataset for testing, which falls within the typical range identified, and 70% for training/validation.

4.3. Model building

4.3.1. Data transformation and preprocessing

To ensure the input variables in each dataset shared a common scale that avoided distorting variations in the values, each of the three input datasets was 1) normalized to a mean of zero by subtracting the average value of each variable, and 2) adjusted to a standard deviation of one by dividing each variable by its standard deviation. Normalization improves the numerical stability of calculations (Kuhn and Johnson, 2013) and may also speed up ML model tuning.

As an optional data preprocessing step, principal component analysis (PCA) transforms input variable sets into smaller groups, or principal components (PC), that capture most of the information in the original set by finding linear combinations of them. By reducing the dimensionality of the modeling problem, PCA can improve model handling of the input variables (Dehghani et al., 2014; Noori et al., 2011; Schmidt et al., 2020; Wilks, 2019). In addition, it eliminates dependencies between input variables by creating uncorrelated predictors. To study its effect on prediction accuracy, input variables were transformed in this study into principal components, which were then ordered and a scree plot used to retain the components that represented 95% of the variability in the data (Schmidt et al., 2020).

Finally, data transformation, and preprocessing where applicable, was used as a part of the cross-validation process (Hastie et al., 2009; Quilty and Adamowski, 2018), to avoid leaking information about the mean and the variance of the validation set to the training data.

4.3.2. *Selection of input variables*

Although it is important to consider many potential predictors as input to an ML model, a smaller number of input predictors improves the chances of finding the best model for the output, or target, variable by reducing the number of possible candidate models (Liu and Motoda, 2007). The goal of variable selection is to find the optimal combination of predictors that maximizes the overall prediction accuracy. The method for best predictor selection removes irrelevant and redundant predictors and generates more reliable forecasts; further, a reduced set of predictors simplifies the model and reduces data collection costs.

To select predictors, we adopted a stepwise variable selection algorithm that starts by choosing the single best variable, identified as the single-variable model that has the minimum residual sum of squares. The model accuracy is then improved in a stepwise manner by adding the variable that reduces the residual sum of squares by the next-largest amount, with subsequent additions of input variables one at a time. The forward stepwise selection is a greedy algorithm. Once it includes a predictor in the model, it cannot remove it later. Therefore, a better algorithm must also account for the fact that some variables become redundant: predictors that improved prediction accuracy at an earlier stage of the selection process may subsequently reduce the overall prediction accuracy. Therefore, we coupled it with a backward elimination algorithm that reexamined all the previously selected variables for inclusion or exclusion after each new variable was selected. This addition and removal of predictors continued until all those that improved the overall prediction accuracy were added and no further addition to improve performance was possible. The advantage of using the stepwise variable selection algorithm, a wrapper approach, over a filter type variable selection algorithm is that the stepwise approach directly evaluates model performance when the input variables are used in combination rather than considering only highly influential input variables that can have their potency negated when used in combination (Kuhn and Johnson, 2019; Liu and Motoda, 2007). Although a third variable selection methods class, called embedded methods, can be more efficient and less computationally expensive than the wrapper methods, their application is restricted to specific ML algorithms. If the modelled process can be better fit by a ML algorithm that does not have an embedded method, then the prediction performance of the developed model is sub-optimal, whereas the application of the wrapper methods is generic to any ML supervised learning algorithm (Kuhn and Johnson, 2019). Embedded methods can perform variable selection as a part of the model building process and avoid variable selection bias. However, it is important to exercise caution when comparing the performance across different machine learning algorithms and their respective embedded methods (or wrapper/filter methods if different feature selection algorithm types were used) to avoid introducing other biases into the comparison, such as resubstitution validation, inconsistent validation sets across different algorithms, and model selection using the testing set.

4.3.3. *ML algorithms and hyperparameter tuning*

The streamflow forecasting models were based on antecedent conditions represented by input variables in monthly forms and lagged by 1-12 months to represent conditions up to one year ahead of the forecast issue date. Further, to study the effect of lead time on the forecasting accuracy, an additional six datasets were prepared, containing only input variables available at the time of prediction. The tested lead times ranged from one to six months. Shorter lead times made more variables available to the model and hence improved accuracy. The longest lead time, six months, corresponded to October of the previous year. October is the start of the water year in Canada and the obtained forecasting accuracy was low enough that longer lead times than six months were unnecessary. The lead times of 1-6 months are the same for the Seasonal, May, June, and July models. Further, predicting average Seasonal, June and July streamflows after the spring melt has no practical value. Therefore, forecasts were not produced after the end of April for the four cases, and longer lead times were calculated at monthly timesteps for each month from October of the previous year to March of the current year. For SWE, we used both basin averages (Section 4.1), and values from individual snow courses for which the model was allowed to assign weights separately.

Five different model types were assessed for seasonal and monthly streamflow forecasting: multiple linear regression (MLR), artificial neural network (ANN), support vector machine (SVM), extreme learning machine (ELM) and radial basis function neural network (RBF). A total of 76 models was developed. First, two sets of twenty models were developed, including seasonal, May, June and July models for each of the five ML algorithms (4 models x 5 algorithms). One of these sets of 20 models included preprocessing with PCA, while the other set of 20 models did not. (The models with PCA are called MLR-PC, ANN-PC, SVM-PC, ELM-PC and RBF-PC below.) Further, to investigate the impacts of longer lead times and compare between datasets S1, S2 and S3, an additional 36 RBF models were developed, including,

- 24 RBF models for seasonal and monthly forecasted streamflows with lead times up to six months,
- Four RBF models to compare between datasets S2 and S3, and,
- Eight RBF models to compare datasets with only local hydrometeorological variables (part of dataset S2) or global climate indices (dataset S1).

4.3.3.1. Multiple linear regression (MLR). Multiple linear regression (MLR) can produce a simple, fast and easy-to-understand reference model for comparison against the machine learning models. Linear regression assumes the streamflow is a linear combination of the input variables. The general form of the MLR model is (Bishop, 1996):

$$f(x) = c_o + c_1 x_1 + c_2 x_2 + \ldots + c_n x_n + \in \quad (1)$$

where $f(x)$ is the forecasted streamflow, $x_1;\ x_2;\ \ldots;\ x_n$ are n-numbers of independent variables, $c_1;\ c_2;\ \ldots;\ c_n$, are the unknown values of the coefficients, and $\in$ is the error term. These values represent the local behavior and are estimated by the least square method.

4.3.3.2. Artificial neural network (ANN). Neural networks are a class of flexible nonlinear models. With an appropriate number of nonlinear processing units, neural networks can learn from historical data and estimate any complex functional relationship with high accuracy. The most popular ANN for time series forecasting is the feed-forward model (Dehghani et al., 2014; Noori et al., 2011; Yaseen et al., 2015). Structurally, an ANN consists of an input layer that receives external information, an output layer, and one or more intermediate “hidden” layers that separate the input and output layers (Miller and Forte, 2017).

In this study, each feed-forward ANN was trained with the standard backpropagation algorithm to minimize the mean squared error, using gradient-descent to calculate the error function gradient iteratively, and with sigmoidal and linear type activation functions for the hidden and output layers, respectively. A single hidden layer feedforward networks (SLFNs) with a nonlinear activation function and a number of hidden neurons (Nh) less than or equal to the number of distinct samples (Ns) can approximate any function and learn with zero error (Huang and Babri, 1998). An exhaustive, or grid, search was used to find the optimum structure of the ANN, which involved trials of all possible hidden neuron numbers between 2 and Ns and selection of Nh associated with the least error across all the samples. In addition, values from zero to 40 were investigated for the regularization, or weight decay, parameter. To avoid under-fitting, we set the maximum number of epochs, or iterations, to 1.0e+6 to ensure that this number is not reached before the error is below the absolute tolerance (1.0e-4) or the optimizer becomes unable to reduce the error by 1.0e-8.

4.3.3.3. Support vector machine (SVM). Originally developed by Vapnik (2000), SVM is one of the most effective prediction models. It uses kernel functions to transform complex nonlinear input variables into linear form by mapping them into a higher-dimensional space, to allow for a linear solution in the higher space that matches the nonlinear solution in the initial nonlinear space. SVM is expressed by,

$$f(x) = \left\{ \sum_{i=1}^{Ns} w_i K(x, x_i) \right\} + \mathrm{B}_b \tag{2}$$

where w_i and B_b are the model parameters, *Ns* is the number of training samples to be estimated, $K(x, x_i)$ is the kernel function, and $f(x)$ is the forecasted streamflow (Çimen, 2008). The model parameters are determined by minimizing the regularized empirical risk function (Çimen, 2008),

$$R_{reg}[f(\vec{x})] = C \sum_{x_i \in X} l_\varepsilon(y_i - f(\vec{x})) + \frac{1}{2}\|\vec{w}\|^2 \tag{3}$$

where C is the cost parameter, a regularization parameter that influences the tradeoff between model complexity and approximation error, and l_ε is the loss function, which is the standard Vapnik's - ε intensive loss function. In this study, the radial basis function kernel was used (Çimen, 2008),

$$K(x, x_i) = e^{\frac{-\|x - x_i\|^2}{2\sigma^2}} \tag{4}$$

and an exhaustive search was conducted to find the cost and sigma tuning parameter.

4.3.3.4. Extreme learning machine (ELM). An ELM is structured as a single layer feed-forward neural network (SLFN) with a faster, more computationally-efficient algorithm than the backpropagation and support vector machine algorithms, among others (Ali et al., 2018). The enhanced performance is linked to the randomization of the input weights and biases (Lima et al., 2015), such that the output weights have unique least-squares solutions obtained by the Moore-Penrose generalized inverse function (Huang and Siew, 2004).

Considering Ns training data samples, the SLFN can be expressed by (Huang et al., 2004),

$$f(x) = \sum_{i=1}^{A} B_i g(w_i . x_i + b_i) \tag{5}$$

where B_i represents output weights to be estimated for the hidden layers with A hidden nodes, g is the activation function, w_i and b_i are the weights and biases, and $f(x)$ is the forecasted streamflow. An exhaustive search was used to determine the optimum structure and activation function of the ELM. Numbers of hidden nodes between 1 and 10000 were tested, as well as sigmoid, sine, radial basis, hard-limit, symmetric hard-limit, satlins, tan-sigmoid, triangular basis and positive linear activation functions. The number of hidden neurons and the activation functions were optimized for the minimum residual sum of squares and the associated optimal setting was selected. In the output layer, which generated the streamflow forecast, a linear transfer function was adopted – a common practice in hydrological time-series forecasting (Deo and Şahin, 2015).

4.3.3.5. Radial basis function network (RBF). RBF is a class of adaptive networks that was first introduced by Broomhead and Lowe (1988). An RBF shares the same SLFN structure as ELM and ANN, with three layers: input, hidden and output. The hidden layer uses a Gaussian (radial basis) activation function rather than the sigmoid function used by ANN. A radial basis function is any real-valued function whose value only depends on an arbitrary distance measure between an input value and a center, where the input value is the data point at which the function is evaluated according to another data point which is the center, and the absolute difference between the two points is the distance which represents the effect of the center on the neighboring points and is inversely proportional to its square root. In our study, a linear type function was used for the output layer. An exhaustive search was used to find the optimum number of hidden nodes of the RBF, with values between 2 and Ns investigated. The number of hidden neurons (Nh) associated with the minimum residual sum of squares was selected.

4.3.4. Model calibration and performance evaluation

A cross-validation scheme with observations split randomly into 10 subsets of equal size (folds) was employed to investigate the forecasting skill of the models. One subset was treated as the validation set (10%) and the remaining subsets (90%) formed the training set. The model was then trained with the training set and evaluated with the validation set. The process was repeated until each subset had been used as a validation set. Comparison of model predictions and test set observations formed the basis of performance measure estimations. This 10-fold cross-validation approach was used because it balances the error bias with the variance and is commonly used in practical machine learning applications (Abu-Mostafa et al., 2012; Kuhn and Johnson, 2013; Solomatine and Ostfeld, 2008). The cross-validation was repeated 3 times to estimate model performance for 30 different validation sets, which can efficiently increase the precision of the estimated cross-validation performance while retaining a small bias (Kuhn and Johnson, 2013) and give a robust estimate of the generalization performance and the variability that might affect the results if different resamples were considered. Finally, all transformations and preprocessing of training subsets were implemented within the cross-validation scheme to avoid bias in the performance estimation through information potentially leaked from the training to the validation set.

We used two resampling loops to properly resample the variable selection process. An outer loop encompassed the entire training/validation process and applied the variable selection algorithm within, which split the training/validation data to 90% for training and 10% for validation. An inner loop then further split the training data (90%) from the outer loop to 90% for training and 10% for validation for use in optimizing and tuning model hyperparameters. Using two resampling loops can significantly reduce the chances of over-fitting the predictors (Boulesteix and Strobl, 2009; Kuhn and Johnson, 2019, 2013). The outer loop was repeated $Q \times N \times k$ times, where Q is the number of steps required for the variable selection algorithm to stop, and N and k are the number of repetitions and folds of the outer loop, respectively. The inner loop was repeated $N' \times k'$ times each time the outer loop was called, where N' is the number of repetitions and k' is the number of folds of the inner loop. It should be noted that each time a validation set (stage 3) or test set (stage 5) is passed to the model, it is transformed using the transformation parameters calculated from the respective training set. Further, ppredictions generated from the developed models where the

Table 4
Average test set performance for Seasonal, May, June and July streamflow forecasting RBF models.

Dataset	Model	Mean absolute error (MAE) m^3/s	Root mean square error (RMSE) m^3/s	Coefficient of determination (R^2)	Nash-Sutcliffe coefficient (NSE)
S1	Seasonal	26.83	32.79	0.43	0.24
	May	28.37	34.13	0.35	0.14
	June	36.06	43.95	0.43	0.38
	July	26.82	34.73	0.46	0.47
S2	Seasonal	14.10	16.79	0.81	0.77
	May	20.33	22.07	0.60	0.39
	June	33.25	38.48	0.51	0.44
	July	15.57	18.41	0.84	0.85
S3	Seasonal	22.05	25.81	0.72	0.45
	May	19.59	26.56	0.58	0.11
	June	35.69	54.21	0.29	0.05
	July	21.62	27.29	0.66	0.67

Table 5
Average performance evaluation measures for dataset S2 for the Seasonal, May, June and July streamflow forecasting models.

Model	Mean absolute error (MAE) m^3/s	Root mean square error (RMSE) m^3/s	Coefficient of determination (R^2)	Nash-Sutcliffe coefficient (NSE)
MLR				
Seasonal	22.25	27.04	0.73	0.48
May	25.44	31.21	0.56	0.21
June	42.46	49.89	0.68	0.37
July	27.68	32.56	0.62	0.37
ANN				
Seasonal	22.75	18.54	0.82	0.67
May	21.74	25.80	0.64	0.47
June	37.98	45.76	0.69	0.46
July	20.19	23.88	0.78	0.62
SVM				
Seasonal	15.96	18.35	0.87	0.78
May	16.53	20.66	0.75	0.66
June	26.73	33.17	0.82	0.71
July	16.24	19.76	0.85	0.75
ELM				
Seasonal	15.95	19.14	0.85	0.76
May	15.38	19.59	0.74	0.64
June	27.28	34.82	0.82	0.65
July	21.52	25.85	0.73	0.59
RBF				
Seasonal	13.51	16.23	0.90	0.82
May	14.79	18.35	0.80	0.72
June	24.63	31.26	0.82	0.75
July	15.28	18.36	0.88	0.80

target variable was transformed must be back-transformed to their original scale before calculating the RSS. This back-transformation must be applied multiple times when estimating the RSS during validation inside the inner and outer loops as well as when testing the performance of the final model.

The trained and validated models were evaluated with statistical measures of goodness of fit and absolute error, as recommended by Legates and McCabe (2005). Models with higher goodness of fit and minimum values of RMSE and mean absolute error (MAE) are recognized to perform better than other models. Our study applied the MAE, RMSE, mean absolute percentage error (MAPE), coefficient of determination (R^2) and Nash–Sutcliffe coefficient of efficiency (NSE) measures.

4.4. Final model training and testing

The model with the best generalization performance in the training/validation process was then selected and trained on the training/validation dataset, and the prediction accuracy over the test set was estimated. Because of the short length historical record available, we implemented preliminary runs using the outlined model building process, shown in Fig. 3. In the preliminary runs, we used 70% of the entire datasets for training/validation and 30% for testing. We applied stepwise variable selection and compared the performance of the five ML algorithms. The combination of input variables and ML with the best generalization performance was selected as the final model and tested with the test set. Then, we empirically estimated the bias in the prediction accuracy, which is the relative difference between the model performance estimated using the validation sets and a separate test set, on future data using only the cross-validation results. Further, to compare between datasets S1, S2 and S3, we used 30% of dataset S3 as the test set to ensure similar test set size and target variable values across the three datasets. Variable selection algorithms and machine learning methods were applied using the R programming language (R Core Team, 2020) – specifically, the caret (Kuhn, 2020), RSNNS (Bergmeir and Benítez, 2012), kernlab (Karatzoglou et al., 2004) and dplyr (Wickham et al., 2020) packages.

5. Results and discussion

The results and discussion focus on dataset S2, which outperformed both S1 and S3. Forecasts obtained using dataset S3 were less reliable because of its shorter training record. Additionally, dataset S1 had a longer training record but a smaller number of input variables, which affected the forecasting accuracy. To illustrate this point, Table 4 compares results among input datasets S1, S2 and S3 for their four best performing models, which were RBF models in all cases. Dataset S3 is omitted from further discussion, while S1 will be discussed later in the comparison between global and local input variables.

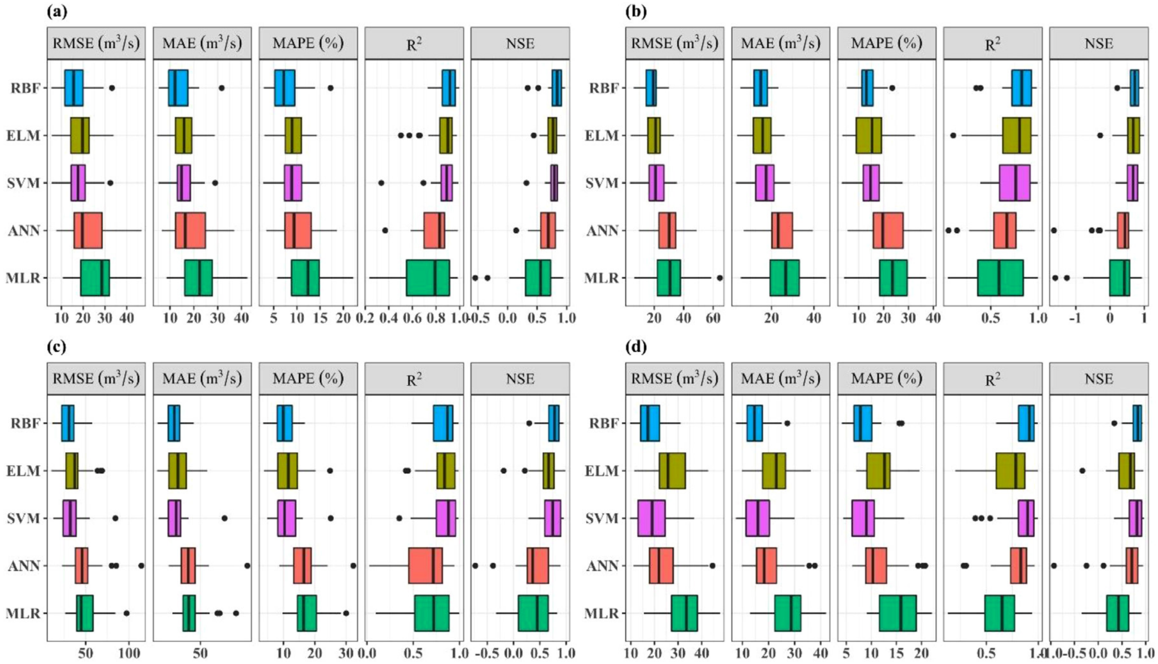

Fig. 4. Boxplots of performance evaluation measures for the out-of-sample forecasts for the period 1970-2019 for (a) seasonal, (b) May, (c) June and (d) July streamflow forecasting models in terms of root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), coefficient of determination (R^2) and Nash-Sutcliffe efficiency (NSE)

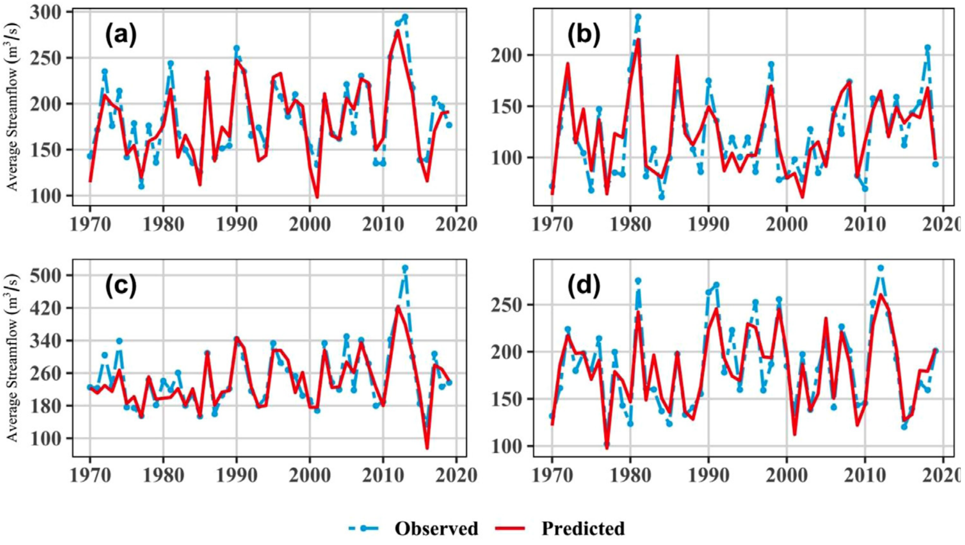

Fig. 5. Radial Basis Function (RBF) network average out-of-sample forecasts and observed streamflow for input dataset S2 (a) Seasonal, (b) May, (c) June and (d) July for the period 1970-2019.

In addition, the prediction accuracy bias was empirically estimated using the average out-of-sample cross-validation and a separate test set, and was found to be very small. For example, the Seasonal, May, June and July RBF models without PCA had an estimated prediction accuracy from the training/validation process of approximately 20.54, 22.27, 36.38 and 23.49 m^3/s, respectively. The RMSE calculated using a separate test set for the Seasonal, May, June and July models were 21.75, 21.32, 37.33 and 22.8 m^3/s, respectively. Hence, the average estimated prediction performance bias was less than 1.2 m^3/s (≅5.5, 4.4, 2.5 and 3.0% for the Seasonal, May, June and July models, respectively). Therefore, the presented workflow can give a reasonable estimate of the prediction accuracy on the test set using cross-validation average out-of-sample results. Additionally, because of the short historical record where every data point is potentially needed to determine model parameters, from this point on we relied on the repeated cross-validation process to estimate the prediction accuracy (Kuhn and Johnson, 2013), while keeping in mind that a small bias may exist in this estimate.

Table 5 lists the average out-of-sample performance measures for dataset S2 for the period of 1970-2019 for the Seasonal, May, June and July models. Fig. 4 summarizes the overall performance evaluation measures for the twenty machine learning models: the tuned MLR, ANN, SVM, ELM and RBF monthly and seasonal forecasting models including the median, interquartile ranges and outliers for the 30 (3 times repeated 10-fold cross-validation) out-of-sample forecasts each for the period 1970-2019. The nonlinear models all outperformed the seasonal and monthly MLR models, and the RBF model provided the best average prediction accuracy and performance consistency across different resamples. The SVM model had overall better forecasting accuracy and evaluation measures, smaller MAE and RMSE, higher R^2 and NSE, than both the ELM and ANN models.

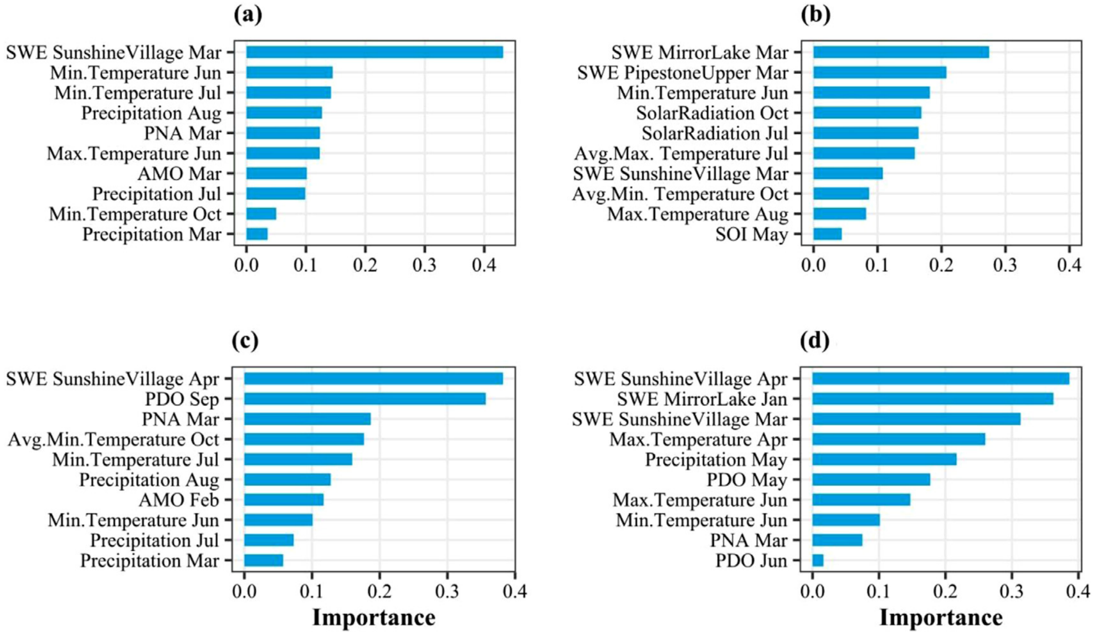

Fig. 6. The ten most important variables for (a) Seasonal, (b) May, (c) June and (d) July streamflow forecasting models and their relative importance, measured as the accumulated percentage decrease in the estimated error with the addition of each predictor to the model.

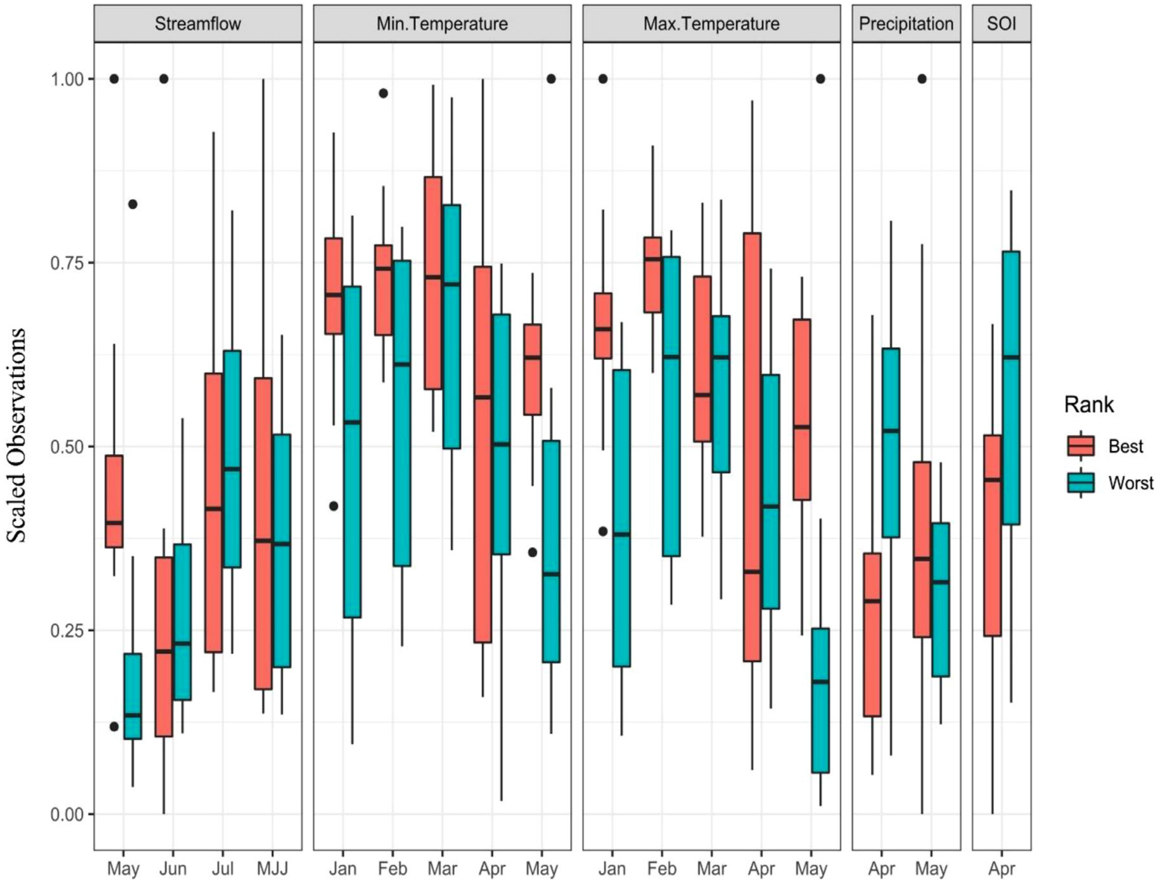

Fig. 7. Boxplots for the May models of the streamflow and other input variables (temperature, precipitation, SOI) with clear difference between the best and worst predicted years. The best years are the set of ten years with the lowest prediction residuals, and the worst years are the set of ten years with the highest prediction residuals.

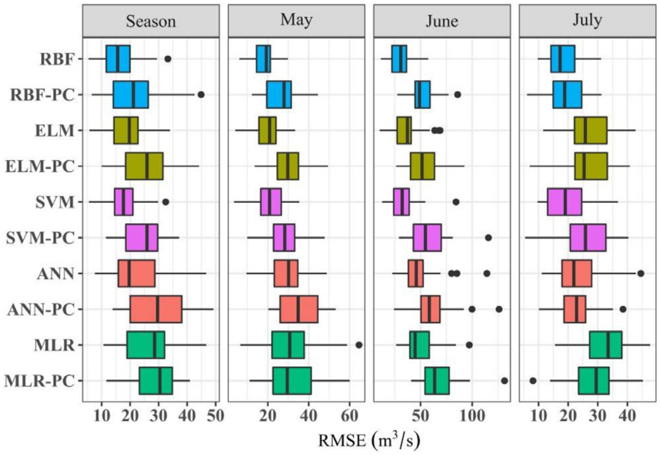

Fig. 8. Prediction RMSE (m^3/s) change for season and monthly models developed with and without using PCA.

Fig. 5 compares the RBF summer season and monthly forecasting models against the observed streamflow for the period from 1970 to 2019. For the seasonal, May, June and July streamflows, the average RBF out-of-sample forecasts over the period from 1970 to 2019 had coefficients of determination of 0.9, 0.8, 0.82, and 0.88 and NSE coefficients of 0.82, 0.72, 0.75, and 0.80.

In addition to model performance evaluation, we investigated all model predictors and extracted their relative importance, measured as the accumulated percentage decrease in the estimated error for each predictor added to the model. Values are shown in Fig. 6. While different models selected different predictors for the four monthly and seasonal streamflows, they shared many common attributes. First, the dominant variables selected in the four cases were the snow water equivalent (SWE) from individual snow courses. The SWE at Sunshine Village in March and/or April was the top predictor for the Seasonal, June and July models and the 6th-most important predictor in May, perhaps because of the central location of the Sunshine Village snow course in the mountainous portion of the watershed. Among the global climate indices, the PDO in September was the 2nd most important predictor for June forecasts, and the PDO values in May and June of the previous year were among the top ten predictors in the July forecasting model. Inputs from other global climate indices, such as the AMO and PNA, were also important in some forecasts. Further, the importance of global climate indices increased with the length of the forecast lead time, as in the cases of the June (one-month lead time) and July (two-months lead time) models. The Arctic Oscillation (AO), the North Atlantic Oscillation (NAO), the North Pacific Index (NP) were unimportant for all four models. In addition, SOI and other El Niño indices were not good predictors despite their general connection to snow cover variability and temperature in the study area; this result may be attributed to the fact that responses to ENSO events vary substantially (Bonsal and Shabbar, 2008).

Although May streamflows did not exhibit as much variability as June streamflows (Fig. 2), they had the least accurate predictions. Further, the interannual variability in the May streamflows explained by the May prediction models was the lowest across all model types, as shown by the R^2 value in Table 5. We therefore investigated the factors related to best versus worst performance for the May models. Focusing on two sets of ten years – one set with the highest residuals (worst performance) and the other with the lowest residuals (best performance), as shown in Fig. 7 – revealed that the most difficult years to predict were those with a significantly lower May streamflow, which has a median of 0.13 compared to a median of 0.4 for the best years, while there was no clear difference based on the full summer-season streamflow (whether the year in question was “dry” or “wet”). Given no clear difference in the total seasonal streamflow, the worst years were likely those in which the snowpack melted later in June, as indicated by the significantly lower minimum and maximum temperatures in May, which have medians below 0.3 and 0.2 for the worst years compared to medians greater than 0.5 for best years, and higher June streamflow values, which nullified the predictive power of the snowpack – the most important predictor of May streamflow, as shown in Fig. 6. The same conclusion results from a recognition that the worst years were generally colder than the best years, particularly in terms of the May maximum temperature, as shown in Fig. 7. In contrast, there was no significant difference in May precipitation for the best and worst years, which suggests that the snowpack melt for the worst years resulted from an increase in maximum temperature rather than rain-on-snow events. Finally, note the difference in the SOI values between the best and worst years, which indicates that May streamflows in La Niña years were harder to predict.

Although a number of studies have suggested the use of PCA (Dehghani et al., 2014; Hao et al., 2018; Mortensen et al., 2018; Noori et al., 2011; Salas et al., 2011), its value for predictor preprocessing to improve model accuracy is not clearly established. Therefore, we investigated predictor preprocessing with PCA using dataset S2 for both seasonal and monthly models, with the effects on RMSE shown in Fig. 8. Clearly, PCA decreased forecasting accuracy, with the average prediction error increasing in almost all cases except for the MLR in July. This result shows that for carefully selected and uncorrelated input predictors and nonlinear models, using PCA may still eliminate important information. This behavior can be attributed to one or more of the following reasons: 1) PCA does not consider the target variable when summarizing variability, 2) PCA is a linear transformation for the input variables, which require different treatment by nonlinear models,

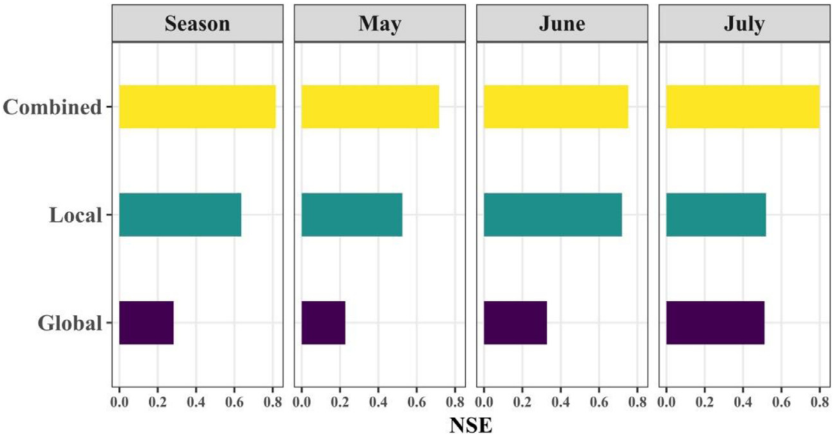

Fig. 9. Change in forecasting accuracy using local hydrometeorological variables only (local), global climate-indices only (global), and using the two sets of input variables in combination (combined).

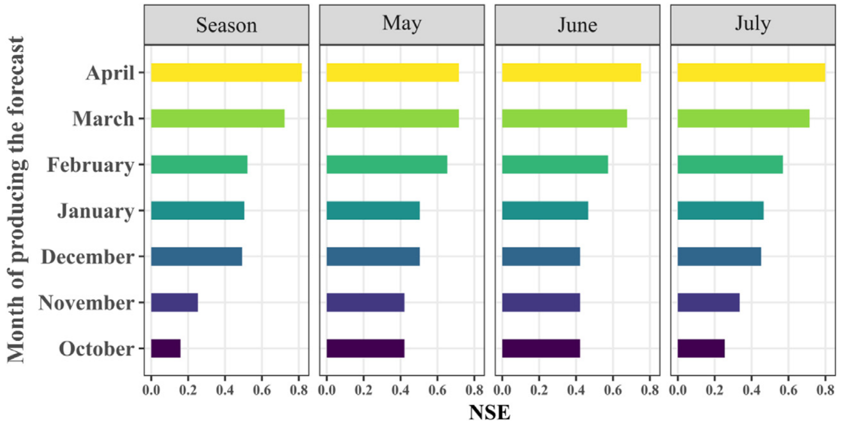

Fig. 10. Change in forecasting accuracy at longer lead times (1-6 months in advance).

and 3) the loss of important information resulted from retaining only the principal components comprising 95% of the variance (Kuhn and Johnson, 2013).

Results from three sets of predictors were compared in order to clarify the effect of global climate indices on projection accuracy: 1) hydrometeorological variables combined with global climate indices, 2) local hydrometeorological variables only and 3) global climate-indices only. Input dataset S2 provided the combination of hydrometeorological and climate-index values. Local hydrometeorological predictors alone were obtained from S2 by omitting global climate-indices from the dataset. Global climate index predictors were obtained from dataset S3. Fig. 9 shows the effect of the three input datasets on the forecasting performance. Incorporating the global climate indices in the predictor set clearly improved the model forecasting skill significantly, while using local climate indices only caused a 10-25% reduction in forecasting skill. Adding global climate index predictors clearly improved streamflow prediction accuracy, particularly at longer lead times.

Finally, Fig. 10 shows the forecasting performance for longer lead times of one to six months. For the Seasonal, June and July forecasts, the forecasting skill in March, at a one-month lead time, was lower than in April but the loss in skill was small in comparison with cases of more than two-month lead time (October-February). Of practical significance for government agencies and water managers, although forecasting skill decreased with increasing the lead time, satisfactory forecasts (NSE>0.5) could also be obtained two months in advance of the spring melt, in February.

6. Summary and conclusions

Data-driven modeling with ML algorithms is one of the most rapidly developing topics in the field of hydrologic modeling and forecasting, especially for its ability to exploit increasing volumes of available data. However, the complexity of the models and their ability to extract useful information from input data requires following a careful model building process. This study identified and addressed pitfalls and challenges facing data-driven ML modeling in hydrology with a novel workflow (Fig. 3) that avoids both by representing the uncertainty involved in variable selection. Further, it leads to better generalization performance on unseen data with unbiased estimation of the prediction accuracy.

The workflow addresses selection bias by considering variable selection as a part of the model building process and using two resampling or cross validation loops to reduce the chance of overfitting to the predictors. It also uses repeated k-fold cross validation to avoid the resubstitution validation problem. The workflow has a consistent, generic model building process applicable to any ML algorithm that allows an equal effect of the performance estimation bias on all models, which is important for cases where relatively short historical records will not permit the use of a test set. It also limits the role of the test set to estimate the prediction accuracy rather than model selection. Finally, it ensures that input variable transformation and preprocessing occur inside cross validation to avoid leaking information about the mean and variance of the validation data to the training data. The flexibility of the workflow for application to any ML model, variable selection algorithm or resampling method allows it to serve as a guide for future machine learning studies in hydrology.

To demonstrate the performance of the new workflow, we developed 76 monthly (May, June and July) and Seasonal streamflow forecasting models for a study location with a relatively short historical record and a large number of predictors compared to samples: the cordilleran, snowmelt-dominated Bow River Basin in Alberta, Canada. We used four different ML algorithms, ANN, SVM, ELM and RBF, and compared them against a more traditional multiple linear regression approach.

In applying the new workflow, we first empirically estimated the bias in the prediction accuracy using the cross-validation results and a separate test set and found the bias to be very small, less than 1 m^3/s (≅5%). This result indicated that following a robust and appropriate resampling and validation process presented in the workflow improved our estimates of the generalization performance, which guides the selection of more accurate models.

The performance of each ML model was evaluated using standard statistical metrics: mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2) and Nash-Sutcliffe model efficiency coefficient (NSE). The RBF performed best of the five model types in terms of average prediction accuracy and consistency in performance across different resamples. Specifically, for the seasonal and monthly predictions, RBF out-of-sample forecasts over the period from 1970 to 2019 had coefficients of determination of 0.90, 0.80, 0.82 and 0.88 and Nash-Sutcliffe Efficiency coefficients of 0.82, 0.72, 0.75 and 0.80. Preprocessing input variables with PCA was found to be detrimental to prediction accuracy.

The relative importance of each predictor on model accuracy was investigated, in terms of the reduction in model error with its inclusion. SWE was the most important predictor for seasonal and monthly models; further, using SWE values from individual snow course stations with an individual assignment of model weight for each station produced better results than an average SWE over the basin. For the seasonal models, SWE measures in March and the previous year's summer temperature and precipitation were the most important predictors. For the May models, prediction accuracies were significantly affected by a late snowpack melt in June. An aggregated seasonal forecast abated such temporal effects. Further, including global climate indices, especially, the Pacific decadal oscillation index, Atlantic multidecadal oscillation index, and Pacific-North American pattern index improved Nash–Sutcliffe coefficient of efficiency by 6% to 50%. Of practical significance for government agencies and water managers, we found that although forecasting skill decreased with increasing forecast lead time, satisfactory forecasts (NSE>0.5) could be obtained two months in advance of the spring melt, at the end of February. We also found that a combination of global climate indices and local hydrometeorological variables led to improved models and forecasting performance. Therefore, reliable seasonal and monthly streamflow forecasts could be obtained for the Bow River Basin using RBF and a combination of local hydrometeorological conditions and large-scale climate indices. RBF forecasts can be applied to support integrated water resources management in Alberta.

Extensions of this study should focus on improving understanding of the tendency of different variable selection algorithms to pick irrelevant or redundant predictors with changes in training data. Better understanding of variable selection algorithms can also improve a major drawback of the proposed workflow – high computational complexity – by finding simpler variable selection algorithms that can lead to similar streamflow forecasting model accuracy. Additionally, exploring an emerging paradigm, theory-guided data science, may improve the interpretability of ML models by limiting the search space through imposing background knowledge of the modeled process (Babovic, 2009; Chadalawada et al., 2020; Herath et al., 2020; Karpatne et al., 2017). The application of the presented workflow could be extended to modeling at finer temporal time resolutions, such as weekly or even daily. The Supplementary Materials demonstrate the impact of increasing the number of observations or of modeling at finer temporal resolutions on the workflow runtime for different ML algorithms. They show that a relative increase in the number of observations by 80 times can lead to a relative increase in the measured runtimes from a minimum of 2.5 (ELM) to a maximum of 40 times (SVM), with an average of 14 times greater. Finally, we encourage other researchers to experiment further and evaluate the application of the presented workflow to forecast different processes, including rainfall, evaporation and urban water demand.

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.

CRediT authorship contribution statement

Amr Gharib: Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **Evan G.R. Davies:** Supervision, Conceptualization, Formal analysis, Writing – review & editing.

Acknowledgements

We thank Dr. Nesa Ilich for his valuable contributions and revisions, and for his provision of the naturalized streamflow data. We also thank Dr. John Quilty of the University of Waterloo and two anonymous reviewers for their helpful comments on, and revisions to, the manuscript.

This work was supported financially by the University of Alberta, Mitacs Accelerate Fellowship, and Optimal Solutions Ltd. It also benefited indirectly from funding to project AI-2336, “Seasonal operational model for water management within irrigation districts”, of Alberta Innovates, with matching support from the City of Calgary, and the Bow River (BRID), Eastern (EID), Raymond (RID), St. Mary River (SMRID), Taber (TID), and Western (WID) Irrigation Districts in Alberta, Canada.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2021.103920.

References

Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.-T., 2012. Learning from data: a short course. [United States] : AMLBook.com.

Alberta Agriculture and Forestry, 2020. Alberta Irrigation Information 2019.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018. Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting. Atmos. Res. 213, 450–464. https://doi.org/10.1016/j.atmosres.2018.07.005.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99, 6562–6566. https://doi.org/10.1073/pnas.102102699.

Babovic, V., 2009. Introducing knowledge into learning based on genetic programming. J. Hydroinform. 11, 181–193. https://doi.org/10.2166/hydro.2009.041.

Behzad, M., Asghari, K., Eazi, M., Palhang, M., 2009. Generalization performance of support vector machines and neural networks in runoff modeling. Expert Syst. Appl. 36, 7624–7629. https://doi.org/10.1016/j.eswa.2008.09.053.

Bergmeir, C., Benítez, J.M., 2012. Neural networks in R using the Stuttgart neural network simulator: RSNNS. J. Stat. Softw. 46, 1–26. https://doi.org/10.18637/jss.v046.i07.

Bishop, C.M., 1996. Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA.

Bonsal, B., Shabbar, A., 2011. Large-Scale Climate Oscillations Influencing Canada, 1900-2008. Canadian Biodiversity: Ecosystem Status and Trends 2010 , Technical Thematic Report No. 4. 15.

Bonsal, B., Shabbar, A., 2008. Impacts of large-scale circulation variability on low streamflows over Canada: a review. Can. Water Resour. J. /Rev. Can. des ressources hydriques 33, 137–154. https://doi.org/10.4296/cwrj3302137.

Bonsal, B.R., Shabbar, A., Higuchi, K., 2001. Impacts of low frequency variability modes on Canadian winter temperature. Int. J. Climatol. 21, 95–108. https://doi.org/10.1002/joc.590.

Boulesteix, A.L., Strobl, C., 2009. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med. Res. Methodol. 9, 1–14. https://doi.org/10.1186/1471-2288-9-85.

Broomhead, D.S., Lowe, D., 1988. Multivariable functional interpolation and adaptive networks. Complex Syst. 2.

Brown, R.D., Goodison, B.E., 1996. Interannual variability in reconstructed Canadian snow cover, 1915–1992. J. Clim. 9, 1299–1318. https://doi.org/10.1175/1520-0442(1996)009<1299:IVIRCS>2.0.CO;2.

Burn, D.H., 2008. Climatic influences on streamflow timing in the headwaters of the Mackenzie River Basin. J. Hydrol. 352, 225–238. https://doi.org/10.1016/j.jhydrol.2008.01.019.

Carrier, C., Kalra, A., Ahmad, S., 2013. Using Paleo reconstructions to improve streamflow forecast lead time in the western United States. J. Am. Water Resour. Assoc. 49, 1351–1366. https://doi.org/10.1111/jawr.12088.

Chadalawada, J., Herath, H.M.V.V., Babovic, V., 2020. Hydrologically informed machine learning for rainfall-runoff modeling: a genetic programming-based toolkit for automatic model induction. Water Resour. Res. 56, 1–23. https://doi.org/10.1029/2019WR026933.

Çimen, M., 2008. Estimation of daily suspended sediments using support vector machines. Hydrol. Sci. J. 53, 656–666. https://doi.org/10.1623/hysj.53.3.656.

Danandeh Mehr, A., Kahya, E., Bagheri, F., Deliktas, E., 2014. Successive-station monthly streamflow prediction using neuro-wavelet technique. Earth Sci. Inform. 7, 217–229. https://doi.org/10.1007/s12145-013-0141-3.

Dehghani, M., Saghafian, B., Nasiri Saleh, F., Farokhnia, A., Noori, R., 2014. Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. Int. J. Climatol. 34. https://doi.org/10.1002/joc.3754.

Deo, R.C., Şahin, M., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monit. Assess. 188, 90. https://doi.org/10.1007/s10661-016-5094-9.

Deo, R.C., Şahin, M., 2015. Application of the Artificial Neural Network model for prediction of monthly Standardized Precipitation and Evapotranspiration Index using hydrometeorological parameters and climate indices in eastern Australia. Atmos. Res. 161–162, 65–81. https://doi.org/10.1016/j.atmosres.2015.03.018.

El-Shafie, A., Abdin, A.E., Noureldin, A., Taha, M.R., 2009. Enhancing inflow forecasting model at aswan high dam utilizing radial basis neural network and upstream monitoring stations measurements. Water Resour. Manag. 23, 2289–2315. https://doi.org/10.1007/s11269-008-9382-1.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Softw. 62, 33–51. https://doi.org/10.1016/j.envsoft.2014.08.015.

Ghorbani, M.A., Khatibi, R., Goel, A., FazeliFard, M.H., Azani, A., 2016. Modeling river discharge time series using support vector machine and artificial neural networks. Environ. Earth Sci. 75, 1–13. https://doi.org/10.1007/s12665-016-5435-6.

Gobena, A.K., Gan, T.Y., 2006. Low-frequency variability in Southwestern Canadian stream flow: links with large-scale climate anomalies. Int. J. Climatol. 26, 1843–1869. https://doi.org/10.1002/joc.1336.

Gobena, A.K., Gan, T.Y., 2009. The role of Pacific climate on low-frequency hydroclimatic variability and predictability in Southern Alberta. Canada. J. Hydrometeorol. 10, 1465–1478. https://doi.org/10.1175/2009JHM1119.1.

Hao, Z., Singh, V.P., Xia, Y., 2018. seasonal drought prediction: advances, challenges, and future prospects. Rev. Geophys. 56, 108–141. https://doi.org/10.1002/2016RG000549.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7.

He, Z., Wen, X., Liu, H., Du, J., 2014. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. J. Hydrol. 509, 379–386. https://doi.org/10.1016/j.jhydrol.2013.11.054.

Herath, H.M.V.V., Chadalawada, J., Babovic, V., 2020. Hydrologically informed machine learning for rainfall-runoff modelling: towards distributed modelling. Hydrol. Earth Syst. Sci. Discuss. 1–42. https://doi.org/10.5194/hess-2020-487

Hsieh, W.W., Tang, B., 2001. Interannual variability of accumulated snow in the Columbia Basin, British Columbia. Water Resour. Res. 37, 1753–1759. https://doi.org/10.1029/2000WR900410.

Huang, G.Bin, Babri, H.A., 1998. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. IEEE Trans. Neural Networks 9, 224–229. https://doi.org/10.1109/72.655045.

Huang, G.Bin, Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man, Cybern. Part B Cybern. 42, 513–529. https://doi.org/10.1109/TSMCB.2011.2168604.

Huang, G.-B., Siew, C.-K., 2004. Extreme learning machine: RBF network case. In: ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004. IEEE, pp. 1029–1036. https://doi.org/10.1109/ICARCV.2004.1468985.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541). IEEE, pp. 985–990. https://doi.org/10.1109/IJCNN.2004.1380068.

Humphrey, G.B., Gibbs, M.S., Dandy, G.C., Maier, H.R., 2016. A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. J. Hydrol. 540, 623–640. https://doi.org/10.1016/j.jhydrol.2016.06.026.

Ilich, N., Gharib, A., Davies, E.G.R., 2018. Kernel distributed residual function in a revised multiple order autoregressive model and its applications in hydrology. Hydrol. Sci. J. 63, 1745–1758. https://doi.org/10.1080/02626667.2018.1541090.

Jain, A., Kumar, A.M., 2007. Hybrid neural network models for hydrologic time series forecasting. Appl. Soft Comput. J. https://doi.org/10.1016/j.asoc.2006.03.002.

Jean, M.È., Davies, E.G.R., 2016. Towards best water management policies: how current irrigation reservoir operation practices compare with theory in Alberta. Water Int. 41, 948–965. https://doi.org/10.1080/02508060.2016.1210562.

Kalra, A., Ahmad, S., 2009. Using oceanic-atmospheric oscillations for long lead time streamflow forecasting. Water Resour. Res. 45, 1–18. https://doi.org/10.1029/2008WR006855.

Kalra, A., Ahmad, S., Nayak, A., 2013a. Increasing streamflow forecast lead time for snowmelt-driven catchment based on large-scale climate patterns. Adv. Water Resour. 53, 150–162. https://doi.org/10.1016/j.advwatres.2012.11.003.

Kalra, A., Miller, W.P., Lamb, K.W., Ahmad, S., Piechota, T., 2013b. Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins. Hydrol. Process. 27, 1543–1559. https://doi.org/10.1002/hyp.9236.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., 2004. kernlab - an S4 package for kernel methods in R. J. Stat. Softw. 11, 1–20. https://doi.org/10.18637/jss.v011.i09.

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168.

Kim, T., Shin, J.Y., Kim, H., Heo, J.H., 2020. Ensemble-based neural network modeling for hydrologic forecasts: addressing uncertainty in the model structure and input variable selection. Water Resour. Res. 56, 1–19. https://doi.org/10.1029/2019WR026262.

Kisi, O., Cimen, M., 2011. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. J. Hydrol. 399, 132–140. https://doi.org/10.1016/j.jhydrol.2010.12.041.

Kisi, O., Kerem Cigizoglu, H., 2007. Comparison of different ANN techniques in river flow prediction. Civ. Eng. Environ. Syst. 24, 211–231. https://doi.org/10.1080/10286600600888565.

Kuhn, M., 2020.

Kuhn, M., Johnson, K., 2019. Feature Engineering and Selection : a Practical Approach for Predictive Models. Chapman & Hall/CRC Press, Boca Raton, FL.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, Applied Predictive Modeling. Springer, New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3.

Ladd, M.J., Gajewski, K., 2009. The North American summer Arctic front during 1948-2007. Int. J. Climatol. 123, n/a. https://doi.org/10.1002/joc.1940.

Legates, D.R., McCabe Jr., G.J., 2005. Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. https://doi.org/10.1029/1998WR900018.

Li, P.H., Kwon, H.H., Sun, L., Lall, U., Kao, J.J., 2010. A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir, Taiwan. Int. J. Climatol. 30, 1256–1268. https://doi.org/10.1002/joc.1954.

Lima, A.R., Cannon, A.J., Hsieh, W.W., 2015. Nonlinear regression in environmental sciences using extreme learning machines: A comparative evaluation. Environ. Model. Softw. 73, 175–188. https://doi.org/10.1016/j.envsoft.2015.08.002.

Lima, A.R., Cannon, A.J., Hsieh, W.W., 2016. Forecasting daily streamflow using online sequential extreme learning machines. J. Hydrol. 537, 431–443. https://doi.org/10.1016/j.jhydrol.2016.03.017.

Liu, H., Motoda, H., 2007. Computational Methods of Feature Selection. Chapman & Hall/CRC.

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. https://doi.org/10.1016/j.jhydrol.2014.06.050.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Softw. 15, 101–124. https://doi.org/10.1016/S1364-8152(99)00007-9.

Makkeasorn, A., Chang, N.B., Zhou, X., 2008. Short-term streamflow forecasting with global climate change implications – a comparative study between genetic programming and neural network models. J. Hydrol. 352, 336–354. https://doi.org/10.1016/j.jhydrol.2008.01.023.

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M., Francis, R.C., 1997. A Pacific interdecadal climate oscillation with impacts on salmon production. Bull. Am. Meteorol. Soc. 78, 1069–1079. https://doi.org/10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2.

Miller, J.D., Forte, R.M., 2017. Mastering Predictive Analytics with R: Machine Learning Techniques for Advanced Models, Second. ed. Packt Publishing.

Moradi, A.M., Dariane, A.B., Yang, G., Block, P., 2020. Long-range reservoir inflow forecasts using large-scale climate predictors. Int. J. Climatol. 6526. https://doi.org/10.1002/joc.6526.

Mortensen, E., Wu, S., Notaro, M., Vavrus, S., Montgomery, R., De Piérola, J., Sánchez, C., Block, P., 2018. Regression-based season-ahead drought prediction for southern Peru conditioned on large-scale climate variables. Hydrol. Earth Syst. Sci. 22, 287–303. https://doi.org/10.5194/hess-22-287-2018.

National Academies of Sciences Engineering and Medicine, 2016. Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts. National Academies Press, Washington, D.C. https://doi.org/10.17226/21873.

Noori, R., Karbassi, A.R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M.H., Farokhnia, A., Gousheh, M.G., 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. J. Hydrol. 401, 177–189. https://doi.org/10.1016/j.jhydrol.2011.02.021.

Nourani, V., Kisi, Ö., Komasi, M., 2011. Two hybrid Artificial Intelligence approaches for modeling rainfall-runoff process. J. Hydrol. 402, 41–59. https://doi.org/10.1016/j.jhydrol.2011.03.002.

Quilty, J., Adamowski, J., 2020. A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. Environ. Model. Softw. 130, 104718. https://doi.org/10.1016/j.envsoft.2020.104718.

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, 175–202. https://doi.org/10.1029/2018WR023205.

Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap rank-ordered conditional mutual information (broCMI): a nonlinear input variable selection method for water resources modeling. Water Resour. Res. 52, 2299–2326. https://doi.org/10.1002/2015WR016959.

R Core Team, 2020. R: A language and environment for statistical computing.

Raschka, S., 2020. Model evaluation, model selection, and algorithm selection in machine learning.

Rasouli, K., Hsieh, W.W., Cannon, A.J., 2012. Daily streamflow forecasting by machine learning methods with weather and climate inputs. J. Hydrol. 414–415, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039.

Rieker, J.D., Labadie, J.W., 2012. An intelligent agent for optimal river-reservoir system management. Water Resour. Res. 48, 1–16. https://doi.org/10.1029/2012WR011958.

Salas, J.D., Fu, C., Rajagopalan, B., 2011. Long-range forecasting of colorado streamflows based on hydrologic, atmospheric, and oceanic data. J. Hydrol. Eng. 16, 508–520. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000343.

Schmidt, L., Heße, F., Attinger, S., Kumar, R., 2020. Challenges in applying machine learning models for hydrological inference: a case study for flooding events across Germany. Water Resour. Res. 56, 1–23. https://doi.org/10.1029/2019WR025924.

Shabbar, A., Skinner, W., 2004. Summer drought patterns in Canada and the relationship to global sea surface temperatures. J. Clim. 17, 2866–2880. https://doi.org/10.1175/1520-0442(2004)017<2866:SDPICA>2.0.CO;2.

Shabri, A., Suhartono, 2012. Streamflow forecasting using least-squares support vector machines. Hydrol. Sci. J. 57, 1275–1293. https://doi.org/10.1080/02626667.2012.714468.

Singh, S.K., 2016. Long-term streamflow forecasting based on ensemble streamflow prediction technique: a case study in New Zealand. Water Resour. Manag. https://doi.org/10.1007/s11269-016-1289-7.

Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling: Some past experiences and new approaches. J. Hydroinform. 10, 3–22. https://doi.org/10.2166/hydro.2008.015.

Tan, Q.F., Lei, X.H., Wang, X., Wang, H., Wen, X., Ji, Y., Kang, A.Q., 2018. An adaptive middle and long-term runoff forecast model using EEMD-ANN hybrid approach. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2018.01.015.

Tyralis, H., Papacharalampous, G., 2017. Variable selection in time series forecasting using random forests. Algorithms 10, 114. https://doi.org/10.3390/a10040114.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer, New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1.

Wickham, H., Francois, R., Henry, L., Muller, K., 2020. dplyr: A Grammar of Data Manipulation.

Wilks, D.S., 2019. Statistical Methods in the Atmospheric Sciences, 4th ed. Elsevier, Amsterdam, Netherlands. https://doi.org/10.1016/C2017-0-03921-6.

Wu, W., May, R.J., Maier, H.R., Dandy, G.C., 2013. A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. Water Resour. Res. 49, 7598–7614. https://doi.org/10.1002/2012WR012713.

Yadav, B., Ch, S., Mathur, S., Adamowski, J., 2016. Discharge forecasting using an Online Sequential Extreme Learning Machine (OS-ELM) model: a case study in Neckar River. Germany. Meas. J. Int. Meas. Confed. 92, 433–445. https://doi.org/10.1016/j.measurement.2016.06.042.

Yaseen, Z.M., Allawi, M.F., Yousif, A.A., Jaafar, O., Hamzah, F.M., El-Shafie, A., 2016a. Non-tuned machine learning approach for hydrological time series forecasting. Neural Comput. Appl. 1–13. https://doi.org/10.1007/s00521-016-2763-0.

Yaseen, Z.M., El-Shafie, A., Afan, H.A., Hameed, M., Mohtar, W.H.M.W., Hussain, A., 2016b. RBFNN versus FFNN for daily river flow forecasting at Johor River, Malaysia. Neural Comput. Appl. 27, 1533–1542. https://doi.org/10.1007/s00521-015-1952-6.

Yaseen, Z.M., El-shafie, A., Jaafar, O., Afan, H.A., Sayl, K.N., 2015. Artificial intelligence based models for stream-flow forecasting: 2000-2015. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2015.10.038.

Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J., El-Shafie, A., 2016c. Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J. Hydrol. 542, 603–614. https://doi.org/10.1016/j.jhydrol.2016.09.035.

Zhang, X., Peng, Y., Zhang, C., Wang, B., 2015. Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. J. Hydrol. 530, 137–152. https://doi.org/10.1016/j.jhydrol.2015.09.047.

Zheng, F., Maier, H.R., Wu, W., Dandy, G.C., Gupta, H.V., Zhang, T., 2018. On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: demonstration for data-driven models. Water Resour. Res. 54, 1013–1030. https://doi.org/10.1002/2017WR021470.