https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Evaluation of Random Forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds

Leo T. Pham[1], Lifeng Luo[2], and Andrew O. Finley[1,2]

[1]Department of Forestry, Michigan State University, East Lansing, Michigan, USA
[2]Department of Geography, Environment, and Spatial Sciences, Michigan State University, East Lansing, Michigan, USA

Correspondence: Leo Pham (phamleo@msu.edu)

Abstract. In the past decades, data-driven Machine Learning (ML) models have emerged as promising tools for short-term streamflow forecasts. Among other qualities, the popularity of ML for such applications is due to the methods' competitive performance compared with alternative approaches, ease of application, and relative lack of strict distributional assumptions. Despite the encouraging results, most applications of ML for streamflow forecast have been limited to watersheds where rainfall is the major source of runoff. In this study, we evaluate the potential of Random Forest (RF), a popular ML method, to make streamflow forecast at 1-day lead time at 86 watersheds in the Pacific Northwest. These watersheds span climatic conditions and physiographic settings and exhibit varied contributions of rainfall and snowmelt to their streamflow. Watersheds are classified into three hydrologic regimes: rainfall-dominated, transisent, and snowmelt-dominated based on the timing of center of annual flow volume. RF performance is benchmarked against Naïve and multiple linear regression (MLR) models, and evaluated using four metrics Coefficient of determination, Root mean squared error, Mean absolute error, and Kling-Gupta efficiency. Model evaluation metrics suggest RF performs better in snowmelt-driven watersheds. Largest improvement in forecasts, compared to benchmark models, are found among rainfall-driven watersheds. We obtain Kling-Gupta Efficiency (KGE) scores in the range of 0.62 - 0.99. RF performance deteriorates with increase in catchment slope and increase in soil sandiness. We note disagreement between two popular measures of RF variable importance and recommend jointly considering these measures with the physical processes under study. These and other results presented provide new insights for effective application of RF-based streamflow forecasting.

1 Introduction

Nearly all aspects of water resource management, risk assessment, and early warning water quality and flood systems rely on accurate streamflow forecast. Yet streamflow forecasting remains a challenging task due to the dynamic nature of runoff in response to spatial and temporal variability in rainfall and catchment characteristics. Therefore, development of skillful and robust streamflow models is an active area of study in hydrology and related engineering disciplines.

In the past decades, Machine Learning (ML) models have gained popularity as promising tools to predict streamflow in addition to physical and stochastic models. These data-driven tools identify patterns in input-output relationship without explicit knowledge of the physical processes or formulation of mathematical equations. To make up for their lack of ability to provide interpretation of the underlying mechanisms, these models often require fewer data, have demonstrated high accuracy in their performance, are computationally efficient, and can be used in real-time forecast (Adamowski, 2008; Mosavi et al., 2018). ML models are particularly useful when accurate prediction is the central inferential goal (Dibike and Solomatine, 2001). Artificial neural networks (ANNs), neuro-fuzzy, support vector machine (SVM), and decision trees (DT) are reported to be among the most popular and effective for both short-term and long-term flood forecast (Mosavi et al., 2018). For example, Dawson et al. (2006) provided flood risk estimation at ungauged sites using ANN at catchments across United Kingdom. Rasouli et al. (2012) predicted streamflow at lead times of 1-7 days with local observations and climate indices using three ML methods Bayesian neural network (BNN), SVM, and Gaussian process (GP). They found BNN outperformed multiple linear regression (MLR) as well as two other competing ML models. Their study also found models trained using climate indices yielded improved longer lead time forecasts (e.g., 5–7 days). Tongal and Booij (2018) forecasted daily streamflow in four rivers in the United States with SVR, ANNs, and Random Forest (RF) coupled with a baseflow separation method. Obringer and Nateghi (2018) compared eight parametric, semi-parametric, and non-parametric ML algorithms to forecast urban reservoir levels in Atlanta, Georgia. Their results showed RF yielded the most accurate forecasts.

Despite the promising results reported in existing literature, most ML streamflow forecast applications are limited to watersheds where rainfall is the major contributor. In many settings, particularly, non-arid mountainous regions, a combination of rainfall and spring snowmelt can drive streamflow (Johnstone, 2011; Knowles et al., 2007). The amount of snow accumulation and its contribution to discharge also vary among the watersheds (Knowles et al., 2006). A natural question is whether ML models can produce comparable performance in watersheds where streamflow contributions come from a mix of snowmelt and rainfall, as well as where snowmelt dominates sources. Considering the prominent role of snowpack in water management and contribution of rapid snowmelt in flood events, such question is worth exploring. To this end, we evaluate the potential of RF in making short-term streamflow forecast at 1-day lead time across 86 watersheds in the Pacific Northwest Hydrologic Unit. The United States Geological Survey (USGS) defines this region as hydrologic unit code (HUC) 17 (U.S. Geological Survey, 2020). HUC-17 consists of sub-basins and watersheds of the Columbia River that span varying hydrologic regimes. The selected watersheds have long-term record of unregulated streamflow and different streamflow contributions of rainfall and snowmelt. Other streamflow forecast studies commonly apply several ML models to a chosen watershed and evaluate the performance of the models in terms of R^2 or other goodness-of-fit measures. Drainage basin factors such as topography, vegetation, and soil can affect the response time and mechanisms of runoff (Dingman, 2015). Few studies attempted to account for and reported these effects on models' performance. Without such consideration, it is difficult to determine if a data-driven model can be generalized to watersheds not included in the given study. Therefore, our objectives are to (1) examine and compare the performance of RF in a number of watersheds across hydrologic regimes and (2) explore the role of catchment characteristics in model performance that are overlooked in previous studies.

In practice, RF can be trained to forecast streamflow at various timescales, depending on the selection of input variables. We focus on 1-day lead time because we assume only antecedent information of predictors are available at the time forecast is made. At longer lead times, changes in weather conditions would likely exert much greater control on runoff and the performance of the model.

We select RF to forecast streamflow for two reasons. First, RF has been referenced to deliver high performance in short-term streamflow forecasts (Mosavi et al., 2018; Papacharalampous and Tyralis, 2018; Li et al., 2019; Shortridge et al., 2016), making it a good candidate for our study. Second, RF allows for some level of interpretability. This is delivered through two measures of predictive contribution of variables: Mean Decrease in Accuracy (MDA) and Mean Decrease in Node Impuritiy (MDI). These two metrics have been widely used as means for variable selection in classification and regression studies in bioinformatics (Chen and Ishwaran, 2012), remote sensing classification (Pal, 2005), and flood hazard risk assessment (Wang et al., 2015). This can be considered an advantage of RF compared with the more “black-box” nature of competing ML algorithms. While the referred interpretability does not directly translate to interpretation of the physical processes, it can provide insight into relationships among predictor variables and streamflow response.

The remainder of the paper is arranged as follows. Section 2 provides a brief introduction to RF and relevant parameters, and selected evaluation indices. Section 3 describes the study area, datasets, and predictor selection. Results and discussion are in Sect. 4. Acknowledgement of limitations and recommendation for future research are also discussed. A summary of conclusions is presented in Sect. 5.

2 Methodology

2.1 Random Forest

Proposed by Breiman (2001), RF is a semi-unsupervised, non-parametric algorithm within the decision tree family that comprises an ensemble of uncorrelated trees to yield prediction for classification and regression tasks. Since a single decision tree can produce high variance and is prone to noise (James et al., 2013), RF addresses this limitation by generating multiple trees where each tree is built on a bootstrapped sample of the training data. Each time a binary split is made in a tree (also known as split node), a random subset of predictors (without replacement) from the full set of predictor variables is considered. One predictor from these candidates is used to make the split where the expected sum variances of the response variable in the two resulting nodes is minimized. The randomization process in generating the subset of the features prevents one or more particularly strong predictor from getting repeatedly chosen at each split, resulting in highly correlated trees (James et al., 2013). After all the trees are grown, each tree casts a vote on a label class for classification task or a prediction value for regression task. The output is the most popular class or the average of all regression values. Breiman (2001) provided full details on Random Forest and its merit. The `randomForest` package in R developed by Liaw et al. (2002) was used for model training and validation in our study. The step-by-step of building a regression RF follows:

https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Algorithm 1 Building a regression Random Forest

Step 1: *n* bootstrap samples are drawn from training set, each has the same size as the training sample. This is also known as `ntree` or number of trees in the forest.

Step 2: At each binary node split, a subset of `mtry` predictors, X_i, is randomly selected from *p* predictor space, Ω_p, that results in $X_i \in \Omega_p$ for {i ∈ 1,..., `mtry`}, `mtry` < *p*.

Step 3: The single best combination of predictor X among X_i predictor variables and threshold *t* is selected to split the observations, y_j, into binary regions

R_1 = { y_j | $X_i < t$ } and R_2 = { y_j | $X_i > t$ } that minimize:

$$\sum_{j:y_j \in R_1} (y_j - \hat{y}_{R_1})^2 + \sum_{j:y_j \in R_2} (y_j - \hat{y}_{R_2})^2 \tag{1}$$

where $\hat{y}_{R_1}$ is the mean of observations in R_1 and $\hat{y}_{R_2}$ is the mean of observations in R_2.

Step 4: Repeat step 2-3 until all terminal region contains less than `nodesize` observations.

Due to sampling with replacement, some observations may not be selected during the bootstrap. These are referred as out-of-bag or OOB, and used to estimate the error of the tree on unseen data. It has been estimated that approximately 37% of samples constitute OOB data (Huang and Boutros, 2016). An average OOB error is calculated for each subsequently added tree to provide an estimate of the performance gain. The OOB error can be particularly sensitive to the number of random predictors used at each split `mtry` and number of trees `ntree` (Huang and Boutros, 2016). Generally, predictive performance improves (or reduction in OOB error) as `ntree` increases. However, recent research has shown that depending on the dataset, there is a limit for number of trees where additional growing does not improve performance (Oshiro et al., 2012). It has been advised that `mtry` is set to no larger than 1/3 of total number of predictors for optimal regression prediction (Liaw et al., 2002), which is also the default value in `randomForest` function in R and widely adopted in literature. Nevertheless, Huang and Boutros (2016) found that this value is dataset-dependent and could be tuned to improve the performance of Random Forest. Bernard et al. (2009) argued the number of relevant predictors highly influences optimal `mtry` value. In this study, we select the optimal `mtry` using an exhaustive search strategy, in which all possible values of `mtry` are considered, using R package `Caret` (Kuhn et al., 2008). Figure 1 illustrates the step-by-step operating principle of growing a Random forest and its relevant parameters.

2.2 Variable importance in Random Forest

In addition to assessing a model's overall predictive ability, there is also interest in understanding the contribution of each predictor variable to model performance. There are two built-in metrics for assessing variable importance in RF: MDA and MDI. After all trees are grown, OOB data during training is used to compute the first measure. At each tree, the Mean squared error (MSE) between predicted and observed is calculated. Then the values of each of the *p* predictors are randomly permuted with other predictor variables held constant. The difference between the previous and new MSE is averaged over all trees. This is considered the predictor variable's MDA (Liaw et al., 2002) and values are reported in percent difference in MSE. The procedure is repeated for each predictor variable. Given that there is a strong association between a predictor and response variable, breaking such bond would result in large error in the prediction (ie., large MDA). It is noted MDA value can be negative where a predictor has no predictive power and adds noise to the model.

The second method, MDI, measures the average gain in residual error reduction each time a predictor is selected to make a split during training. It is based on the principle that a binary split only occurs when residual errors (or impurity) of two descendent nodes are less than that of their parent node. The MDI of a predictor is the sum of all gains across all trees divided by the number of trees. Because the scale of MDI depends on values of response variable, raw MDI provides little interpretation. Following Wang et al. (2015), we computed relative MDI for each variable, which in our case is calculated by dividing each predictor variable's MDI by the sum of MDI from all predictors at each watershed. When scaled by 100, this relative MDI is a percentage and can be interpreted as the relative contribution of each predictor to the total reduction in node impurities. In case the predictor makes no contribution during the splitting, the relative MDI would be effectively zero. In both metrics, the larger the value, the more important the predictor.

2.3 Benchmark models

We benchmark the performance of RF during the validation period against multiple linear regression (MLR) and simple Naïve models using the calculated Pearson correlation coefficient (r) between forecasted and observed values for each model. In Naïve model, we assume “minimal-information" scenario and the best estimate of the streamflow from the next day is the observed value from current day (Gupta et al., 1999). Its r, in this case, is the 1-day autocorrelation coefficient in the time series and measures of the strength of persistence. We train and verify MLR model using same data sets and predictors supplied to RF model.

2.4 Evaluation metrics

There exists different model performance metrics and each provides unique insights on the correspondence between forecasted and observed streamflow values. While Pearson correlation coefficient (r) and its square, namely Coefficient of determination (R^2), are often used, Legates and McCabe Jr (1999) discussed the limitation of these two measures. The authors recommended that absolute error measures (i.e., Root mean squared error or Mean absolute error) and goodness-of-fit measure, such as the Nash-Sutcliffe efficiency (NSE), could provide more reliable and conservative assessment of the models. Kling-Gupta efficiency (KGE) is a relatively new metric that was developed based on a decomposition of NSE (Gupta et al., 2009). This goodness-of-fit measure is gaining popularity as a benchmark metric for hydrologic models by addressing several shortcomings diagnosed with NSE. For these reasons, we selected the following four metrics to evaluate RF performance: R^2, RMSE, MAE, and KGE. These metrics cover various aspects of model's performance and are also provide intuitive interpretation.

R^2 can be interpreted as the proportion of the variance in the observed values that can be explained by the model. Values are in the range between 0 and 1 where 1 indicates the model is able to explain all variation in the observed dataset.

$$R^2 = \left(\frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (\hat{y}_i - \overline{\hat{y}})^2} \sqrt{\sum_{i=1}^{N} (y_i - \overline{y})^2}} \right)^2 \tag{2}$$

where $\hat{y}_i$ and y_i are the forecasted and observed values respectively with

$$\overline{y_i} = \frac{1}{N} \sum_{i=1}^{N} y_i \quad \text{and} \quad \overline{\hat{y}_i} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i \tag{3}$$

MAE provides an average magnitude of the errors in the model's predictions without considering the direction (underestimation or overestimation).

$$MAE = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{N} \tag{4}$$

RMSE is the standard deviation of the residuals between the predictions and observers. It is more sensitive to larger error due to the squared operation. Both MAE and RMSE values are scale-dependent as they depend on the magnitude of values. The standardization in streamflow measurements (described in Sect. 3) allows comparison of MAE and RMSE across gauges.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}} \tag{5}$$

KGE metric ranges between -inf and 1. While there currently is not a definitive KGE scale, negative KGE values are considered “not satisfactory” or “undesirable” (Schönfelder et al., 2017; Siqueira et al., 2018) and model performance is considered as “poor" with $0 < \text{KGE} < 0.5$ (Rogelis et al., 2016). KGE is calculated as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{6}$$

where *r* is the correlation coefficient, α is a measure of relative variability in the forecasted and observed values, and β represents the bias:

$$\alpha = \frac{\sigma_{\hat{y}}}{\sigma_y} \quad \text{and} \quad \beta = \frac{\mu_{\hat{y}}}{\mu_y} \tag{7}$$

where $\sigma_{\hat{y}}$ is the standard deviation in observations, σ_y is the standard deviation in forecasted values, $\mu_{\hat{y}}$ is the forecasted mean, and μ_y is observation mean.

In hydrological forecast, one might be interested in the ability of the model to capture more extreme events rather than the overall performance. The definition of "extreme" depends on the objective of the studies. Here, we adopt the peak-over-threshold method of classifying points extreme daily discharge at 90th, 95th, and 99th percentile thresholds during the validation period. We measure the ability of RF to capture these events using two additional metrics: Probability of Detection (POD) and False Alarm Rate (FAR). The calculation followed as in (Karran et al., 2013).

$$POD = \frac{P(\hat{y}_i > \omega | y_i > \omega)}{P(y_i > \omega)} \tag{8}$$

$$FA = \frac{P(\hat{y}_i > \omega | y_i < \omega)}{P(y_i < \omega)} \tag{9}$$

where ω is a specified threshold.

3 Study Area and data

3.1 Watersheds in Pacific Northwest Hydrological Unit

In this study, we focus on watersheds in the Pacific Northwest or USGS designated HUC 17. This region covers an area of 836,517 km^2 and encompasses all of Washington, six other states, and British Columbia, Canada. For the purpose of maintaining consistency in monitoring protocol and data, we only consider watersheds on the US territory. Columbia River and its tributaries make up the majority of the drainage area, traveling more than 1,240 miles with an extensive network of more than 100 hydroelectric dams and reservoirs have been built along these river channels. Hydropower in the Columbia River Basin supplies approximately 70 percent of Pacific Northwest energy (Payne et al., 2004). Flood control is also an important aspect of reservoir operation in this region.

The north-south running Cascade Mountain Range divides the region into eastern and western parts and strongly influence the regional climate. The windward side of the mountain receives an ample amount of winter precipitation compared to the leeward side. When temperature falls near freezing point, precipitation comes in the form of snow and provides water storage for dry summer months. East of the Cascades, summer rainfall result from rapidly built thunderstorm and convective events that can produce flash floods (Mass, 2015). Proximity to the ocean provides buffering effect, resulting in more mild temperature in the winter.

3.2 Data

3.2.1 Streamflow

Our analysis uses streamflow data available through the USGS National Water Information System (NWIS) (https://waterdata.usgs.gov/nwis/sw). From NWIS, we selected daily streamflow time series for gauges using the following criteria: 1) continuous operation during the 10-year period between 2009 and 2018; 2) have less than 10 percent of missing data; 3) positioned in watersheds with "natural" flow that is minimally interrupted by anthropogenic intervention such reservoirs. The third criterion was met using the GAGES-II: Geospatial Attributes of gauges for Evaluating Streamflow dataset (Falcone, 2011) classification to identify watersheds with least-disturbed hydrologic condition and represented natural flow. Additional screening was performed to remove gauges that were inconsistent with others based on correlation coefficient comparison between the respective gauge and mean basin streamflow. We also excluded small creeks with drainage area less than 50 km^2. In total, 86 watersheds were selected (complete watershed physical characteristics are provided in Supplementary materials).

Following methodology proposed in Wenger et al. (2010), the watersheds were further grouped into three classes of hydrologic regimes based on the timing of center of annual flow, which is defined as the date at which half of the total annual flow volume is exceeded. The annual flow calculations follow a water-year calendar that begins October 1st and ends September 30th. These three hydrologic regimes include: "early" streams with flow time $<$ 150 (27 February), "late" streams with flow time $>$ 200 (18 April), and "intermediate" streams with flow time between 150 and 200. These hydrologic regimes correspond to rainfall-dominated, snowmelt-dominated, and transient or transitional (mixture of rain and snowmelt) hydrographs, respectively. While this particular classification and its variants have been used in various studies related to water resources in this region (Mantua et al., 2009; Elsner et al., 2010; Vano et al., 2015), we adopted this partition in our study for two reasons. First, as Regonda et al. (2005) pointed out, the classification provides a summary of information about type and timing of precipitation, timing of snowmelt, and the contribution of these hydro-climatic variables to streamflow. This helps us assess model performance in consideration of sources of runoff. Second, the classification provides a basis to generalize the results to other watersheds that are not part of the study.

On average, records at these watersheds have less than 3 percent missing data during the 2009–2018 period. The drainage area of the watersheds range between 51 km^2 and 3355 km^2, and the mean elevation range from 239 m and 2509 m, estimated from 30-m resolution digital elevation model (Table 1). Spatial distribution of watersheds is shown in Fig. 2.

3.2.2 Precipitation

Daily precipitation observations were obtained from the AN81d PRISM dataset (Di Luzio et al., 2008). This gridded dataset has a resolution of 4km, covers the entire continental US from January 1981 to present, and is continuously updated every 6 months. Best estimate gridded value is derived by using all the available data from numbers of station networks ingested by the PRISM Climate Group. A combination of Climatologically aided interpolation (CAI) and RADAR interpolation were used in developing PRISM dataset. In our study, watershed daily precipitation (measured in mm) time series were constructed by computing the arithmetic mean for precipitation values of all grid points that fall within the given watershed.

3.2.3 Snow water equivalent and temperatures

Snow water equivalent (SWE) is defined as the depth of water that would be obtained if a column of snow were completely melted (Pan et al., 2003). Daily SWE data were retrieved from 201 Snow Telemetry (SNOTEL) Stations in the PNW. These stations are part of the network of over 800 sites located in remote, high-elevation mountain watersheds in the western U.S. The elevation of these stations are in the range of 128 m and 3142 m. At SNOTEL sites, SWE is measured by a snow pillow—a pressure sensitive pad that weighs the snowpack and records the reading via a pressure transducer. As the temperature shift is the primary trigger for snowmelt, daily maximum temperature (TMAX) and minimum temperature (TMIN) from SNOTEL sensors were also retrieved and included as predictors. The obtained data reflected the last measurement recorded for the respective day at each site. The dataset is mostly complete, with 99.6%/99.6%/99.9% of the observations are available for three variables TMAX, TMIN, and SWE respectively. Because of the sparse coverage of SNOTEL sites, daily average values were calculated at USGS basin level (or 6-digit Hydrological Unit) and subsequently applied to the watersheds located in that basin. There is a total of 15 basins, each contains a number of SNOTEL stations in the range between 6 and 30. It is noted the *in situ* data from these of stations cannot capture the spatial variability of snow accumulation and computing an area-averaged snowpack value from observations remains a challenging task (Mote et al., 2018). The SNOTEL averages, therefore, represent first-order estimates of snow coverage and temperature conditions.

3.2.4 Predictor selection

Future daily mean streamflow (Q_{t+1}) is the response variable in our study. We attempt to explain the variability in Q_{t+1} using eight relevant predictors from the three datasets (Table 2). Selection of predictors is based on thorough review of the literature from previous studies and our understanding of the hydrology of this region. Specifically, precipitation (P_t) is intuitively a driver of streamflow. SWE_t metric provides storage information on the amount of accumulated snow available for runoff and is influenced by changes in temperature ($TMAX_t$ and $TMIN_t$). Previous day streamflow (Q_t) is particularly important due to high degree of persistence that exist in the time series. The Pentad Index (PEN_t) is introduced to account for highly seasonal characteristics of the streamflow in this region (Zheng et al., 2018). A hydrological year consists of 73 pentads where each comprises of five consecutive days. Data preprocessing showed moderate to strong non-linear correlation between daily streamflow and Pentad Index across gauges. We also derived two variables: sum of 3-day precipitation ($P3_t$) and snowmelt (SD_t) from available data. Inclusion of 3-day precipitation was to account for large winter storms that can last for several days, which often result in surges in streamflow. SD_t was calculated as the difference between SWE at day t and $t-1$. It is noted that we use the term "snowmelt" to facilitate discussion in the context of runoff generated mechanism. A positive value of SD_t indicates snow accumulation and negative value indicates melt.

Soil moisture is also a relevant variable in streamflow modeling as it controls the partition between infiltration and runoff of precipitation (Aubert et al., 2003). However, soil moisture data is often limited and incomplete, especially at daily interval and therefore not included in this study. The data were divided into two sets: training consisting of seven years 2009–2015 and a validation set of three years 2016–2018. All data was standardized using Min-Max Scaling to facilitate comparison across gauges. A flowchart representing the input-output model based on RF is shown in Fig. 3.

4 Results and discussion

4.1 Parameter tuning

As we mentioned in Sect. 2, error rate in RF can be sensitive to two parameters: the number of trees `ntree` and number of randomly selected predictors available for splitting at each node `mtry`. We trained RF on a sample of training data sets and observed that the reduction in error is negligible after 2000 trees. Therefore, `ntree`=2000 was set across watersheds. `mtry`, on the other hand, was tuned empirically using a combination of exhaustive search approach and cross-validation.

The goal of tuning is to select the `mtry` parameter value that would optimize the performance of the model. The candidates were evaluated based on their out-of-bag Mean absolute error (MAE). At each watershed, eight possible candidate values of `mtry` (1-8) were analyzed by 3 repetitions of 10-fold cross validation from the train data set. Averaging the MAE of repetitions of the cross-validation procedure can provide more reliable results as the variance of the estimation is reduced (Seibold et al., 2018). To illustrate, in Fig. 4, lowest cross-validation MAE is obtained at `mtry` = 3 at Carbon River Watershed (USGS Site 12094000). The results of tuning for all gauges (Table 3) show that the optimal `mtry` values are {3,4,5} with median MAE of 0.0127, 0.0116, and 0.0079 respectively. These values are close to the suggested default `mtry` for regression (i.e., round-up of the square root of total number of predictors or 3 in our study). The optimal `mtry` at each gauge was then used in both training and validating the model. Because the number of predictors in our study is relatively small, computation burden of the exhaustive search was manageable. As the number of candidate grows, a random search strategy (Probst et al., 2019) in which values are drawn randomly from a specified space can be more computationally efficient.

4.2 Benchmark RF against Multiple Linear Regression and Naïve models

Figure 5 shows the pair-wise comparisons of r values for RF, MLR, and Naïve models. In Fig. 5a, we observe RF mostly outperforms Naïve Model in rainfall-driven and transient watersheds. We also discern large improvement, defined as the positive difference in r values between RF and Naïve Model, tends to occur where persistence is relatively lower. This suggests that application of RF would be most benefiting at watersheds where next-day streamflow is less dependent on the condition of the current day. Among snowmelt-driven watersheds, three models show marginal difference in r values. As Mittermaier (2008) pointed out, the choice of reference can affect the perceived performance of the forecast system. Our pair-wise comparisons highlight the fact that evaluating data-driven models should be performed in consideration of the autocorrelation structure in the data (Hwang et al., 2012). Without accounting for persistence model, it would inadequate to conclude that RF gives better performance in snowmelt-driven watersheds. Nevertheless, we observe RF outperformed MLR in all watersheds in rainfall-dominated and transitional watersheds and 19 out of 25 snowmelt-dominated watersheds. The median r values for RF in the

https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

three groups are (0.88, 0.89, 0.98) compared to (0.85, 0.87, 0.98) for MLR. This may reflect RF's better ability to capture non-linear relationship between streamflow and other variables.

4.3 Evaluation of RF overall performance

We next evaluated the overall performance of RF across three flow regimes using four metrics R^2, KGE, MAE, and RMSE (Fig. 6). We observe similar trend reported in Fig. 5 where RF performs better in snowmelt-dominated than rainfall-dominated (higher R^2, lower MAE). Snowmelt-dominated watersheds have the smallest range of R^2 values across the three groups. This may suggest that there is less variability in flow behaviors at individual gauges in this group. Not surprisingly, transitional group has the largest spread in R^2 values as watersheds in this group share characteristics from the other two groups.

Because RMSE gives more weight to larger errors compared to MAE, the difference between the two metrics represents the extent in which outliers are present in error values (Legates and McCabe Jr, 1999). In rainfall-driven and transient groups, the shape of the boxplot distributions remain fairly consistent between the two error metrics, suggesting that distribution of large errors is similar to that of mean errors in these watersheds. In snowmelt-driven watersheds, we observe a noticeably wider interquartile range (difference between first quartile and third quartile) in RMSE plot compared to MAE plot. This indicates that RF can still be susceptible to underestimation or overestimation in watersheds where the mean error is relatively low.

In Table 4, KGE scores are reported in a range of 0.64–0.99 for all watersheds. The median values for each flow regime are 0.84, 0.87, and 0.94. Based on assessment proposed by Rogelis et al. (2016) where model performance is considered "poor" for $0.5 > \text{KGE} > 0$, RF can be seen to give satisfactory performance at all watersheds. Our results are comparable to findings in Tongal and Booij (2018) where authors compare the performance of RF, SVM, and ANNs to simulate daily discharge with baseflow separation at four rivers in California and Washington. Although authors did not classified these basins, it can be inferred three of the rivers were rainfall-driven and one was snowmelt-driven. RF model in their study produced KGE scores of 0.41, 0.81, and 0.92 for the rainfall-driven water basins (without baseflow separation). However, our KGE scores for snowmelt-fed watersheds (with a median of 0.94) are higher compared to the reported 0.55 in their study.

4.4 RF performance on extreme streamflows

We also examine the model's capacity to forecast extreme events because of their potential high impact and associated flood risks in this region. As seen in Fig. 7a, RF becomes expectedly less skilful in its forecasts with increase in magnitude of the events. At 90th percentile threshold, we observe the same pattern as seen in the R^2 and KGE boxplots. The model tends to perform better among snowmelt-dominated watersheds compared to those in transient and rainfall-driven groups. At 95th threshold, RF can forecast correctly at least 50 percent of the times at most watersheds. At 99th threshold, there are large spreads in POD and the difference in RF's ability to forecast extreme streamflow among the three flow regimes becomes less obvious. In snowmelt-driven watersheds, 8 out of 25 have POD > 0.5, 9 have POD between 0.01 and 0.5, and 8 have a POD of 0. While few studies have examined complex diel hydrologic responses in high-elevation catchments (Graham et al., 2013), our particular result suggests large surges in streamflow sustained by spring and early summer snowmelt can be difficult to predict, even at 1-day lead time, and is an ongoing research subject (Ralph et al., 2014). We see in Fig. 7b, False alarm rate

(FAR) is in agreement with POD and suggests that RF is consistent in its forecasts of rare events. That is, a high POD value is not a result of systematic overestimation. In such cases, we would observe both high POD and FAR among snowmelt-driven watersheds.

4.5 Analysis of variable importance

Variable importance is a useful feature in both understanding the underlying process of current model and generating insights for selection of variable in future studies (Louppe et al., 2013). RF quantifies variable importance through two metrics: DMA and MDI (Fig. 8). In both metrics, the higher value indicates variable contributes more to the model accuracy. Intuitively, streamflow from previous day is shown to be the most importance variable due to persistence. This is reflected across three flow regimes and two metrics. We also observe the sum of 3-day precipitation tends to more predictive power than than 1-day precipitation. Maximum temperature and minimum temperature share similar contribution where minimum temperatures tend to receive slightly higher scores. Among snowmelt-dominated watersheds (Fig. 8c and 8f), we anticipate snow indices (SD_t and SWE_t) contribute more in the prediction than precipitation and this is also reflected. Surprisingly, Pentad Index comes third in both metrics. This supports the long-term snowpack memory of daily streamflow (Zheng et al., 2018) and can be useful in real prediction. Precipitation does not seem to have significant contribution to the model's accuracy in this group. Although PRISM precipitation data includes both rainfall and snowfall, it is likely that the majority of fallen precipitation in these high-altitude watersheds is stored as snow on the surface and does not immediately contribute to runoff. Li et al. (2017) estimated that 37% of the precipitation falls as snow in western US, yet snowmelt is responsible for 70% of the total runoff in mountainous areas. It is still very surprising to observe such low contribution of precipitation variable to RF model accuracy. Nevertheless, we observe general agreement between the two metrics in ranking of the variables in snowmelt-driven group.

In transient and rainfall-dominated groups, there are noticeable disagreement between the two metrics. Precipitation (P_t) and 3-day precipitation ($P3_t$) tend to rank lower in MDA measure (Fig. 8a and 8b) compared to MDI (Fig. 8d and 8e). Specifically, in rainfall-dominated group, 3-day precipitation and precipitation are placed 2nd and 3rd based on median MDI compared to 4th and 7th in MDA. Maximum and minimum temperatures, on the other hand, tend to be more important in MDA calculation compared to in MDI. In Shortridge et al. (2016), RF model was used to predict streamflow at five rain-fed rivers in Ethiopia. Similarly calculated MDA in this study suggested precipitation were less important (7.71%) than temperature (12.74%). Linear model in the same study, however, considered the coefficient for precipitation to be significant ($p << 0.01$) while temperature coefficient was not ($p = 0.08$). In Obringer and Nateghi (2018), authors predicted daily reservoir levels in three reservoirs in Indiana, Texas, and Atlanta using RF and other ML techniques. Precipitation was reported as the least important variable and ranked behind dew point temperature and humidity. Inspecting the density distribution of our predictors, we suspect that for variables that are heavily skewed and zero-inflated (e.g., precipitation), permutation-based MDA may underestimate their importance compared to those that are more normally distributed such as maximum and minimum temperatures. Strobl et al. (2007) showed RF variable importance measures can be unreliable in situations where potential predictor variables vary in their scale of measurement or their number of categories. There is also an ongoing discussion regarding the stability of both metrics across different datasets (Nicodemus, 2011; Calle and Urrea, 2010). Although results from MDI make more sense in our case, we suggest RF users to exert caution when interpreting outputs from these two metrics.

4.6 Effects of watershed characteristics on model performance

To explore the role of catchment characteristics such as geology, topography, and land cover on the performance of RF model, we perform Pearson correlation test between the KGE scores and selected basin physical characteristics for each flow regime. The results are shown in Table 5. There is a strong negative correlation ($p < 0.05$) between KGE scores and watershed slopes among rainfall-dominated and transient watersheds. As steeper hillslope often associates with faster surface and subsurface water movement during event-flow runoff, this can result in shorter response time. We observe a similar trend between KGE scores and percent of sand in the soil (Fig. 9) where the RF performs worse in watersheds with higher hydraulic conductivities (i.e., higher sand content). This could be a result of rapid subsurface flow from soil profile enabled by soil macropores in mountainous forested area (Srivastava et al., 2017), where subsurface flow is the predominant mechanism. Without a quantification of the partition of discharge into surface flow and subsurface flow at individual watersheds, it is difficult to determine the relative importance of subsurface runoff mechanisms in regulating streafmlow and how that may have affected the RF performance. The findings, however, suggest RF performance can deteriorate at watersheds with quick-response runoff when supplied with 1-day delayed observation data.

It appears that stream density and the amount of vegetation cover may also affect the performance of RF, but the relationships are not statistically significant at $\alpha = 0.05$. Aspect eastness, darainage area, and basin compactness are not determining factors to variability in the KGE scores. We also explored the impact of land-use/land-cover, which can be represented by the extent of impervious cover in each watershed. However, because we only selected unregulated watersheds that experienced minimal human disruption during the initial screening, most watersheds have very little impervious cover (less than 5%). It is noted that these selected characteristics are not meant to be exhaustive, but rather representative of various types of factors that could help explain the variability in model performance. Furthermore, an alternative approach to Pearson's correlation is to use ANOVA to test for marginal significance of each catchment variable to KGE while accounting for their interaction. Because our objective is not to make inference on KGE based on these variables and ANOVA analysis can be complicated to interpret, we choose to compute correlation coefficient r.

4.7 Limitations and future research

There area some notable limitations in our study as well as RF in general. The classification of watersheds into three flow regimes was based on the timing of the climatological mean of the annual flow volume, which can fluctuate from year to year. This is particularly true for watersheds in transient group where streamflow is contributed by a mix of runoff from winter rainfall and springtime snowmelt, where inter-annual variability is tremendous in both magnitude and timing (Lundquist et al., 2009). Therefore, the membership of the classified watersheds from this group can vary. In fact, Mantua et al. (2009) discussed the future shift of transient runoff watersheds towards rainfall-dominated in Washington. Because we trained RF using the same input variables for all watersheds regardless of flow regimes and calculated performance metrics separately, the classification does not alter the results at individual watershed.

In the study, we used estimated precipitation from PRISM, which is an interpolation product and combines data from various rain gauges from multiple networks. Despite of possible introduced errors and uncertainty, we believe the use of spatially distributed product better represents the areal estimation of precipitation over the basin than a single rain gauge measurement. In real-time forecast, this would be not be feasible due to the added time to compile and process such data. As our results indicate that RF can produce reasonable forecasts, potential future research could explore the sensitivity of the model using a station data or even include $t+1$ precipitation forecast as a predictor.

An inherent limitation of RF is the lack of direct uncertainty quantification in prediction. In our case, forecasted streamflow using RF does not yield a standard error comparable to that provided traditional linear model, and hence no way to provide probabalistic confidence intervals on predictions. Estimation confidence interval methods have been proposed by Wager et al. (2014); Mentch and Hooker (2016); Coulston et al. (2016), but they are not widely applied. For future work, computation of confidence interval in RF prediction will be useful in addressing and understanding uncertainty.

5 Conclusions

Accurate streamflow forecast has extensive applications across disciplines from water resources and planning to engineering design. In this study, we assessed the ability of RF to make daily streamflow forecast at 86 watersheds in the Pacific Northwest. Key results are summarized below:

- Based on KGE scores (ranging from 0.62 to 0.99), we show RF is able to produce useful forecasts across all watersheds.
- RF performs better in snowmelt-dominated watersheds, which can be attributed to stronger persistence in the streamflow time series. Largest improvements in forecast compared to Naïve model are found among rainfall-dominated watersheds.
- The two built-in approaches for measuring predictor importance yield noticeably different results. We recommend interpretation of the these two metrics should be coupled with understanding of the physical processes and how these processes are connected.
- Steepness of slope and amount of sand content are found to deteriorate RF performance in two flow regime groups. This demonstrates catchment characteristics can cause variability in performance of the model and should be considered in both predictor selection and evaluation of the model.

Considering the current and future vulnerabilities of the Pacific Northwest to flooding caused by extreme precipitation and significant snowmelt events (Ralph et al., 2014), skillful streamflow forecasts can have important implications. Due to its practical applications, RF and RF-based algorithms continue to gain popularity in hydrological studies (Tyralis et al., 2019). Given the promising results from our study, RF can be used as part of an ensemble of models to achieve better generalization ability and accuracy not only in streamflow forecast but also in other water-related applications in this region.

Code and data availability. Example code for building Random Forest model in R and data are available at https://github.com/leopham95/RandomForestStreamflowForecast

Author contributions. **Leo Pham**: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft. **Lifeng Luo**: Conceptualization, Investigation, Methodology, Funding acquisition, Supervision, Project administration, Resources, Writing - original draft. **Andrew Finley**: Resources, Supervision, Funding acquisition, Writing - original draft.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. Leo Pham was supported by the Algorithms and Software for SUpercomputers with emerging aRchitEctures Fellowship funded by National Science Foundation Grant No.1827093. We wish to express deep gratitude to the researchers at the National Supercomputing Center in Wuxi, China and Tyler Willson at Michigan State University for the inital brainstorming and project development.Luo's effort was partially supported by the national science foundation (NSF-1615612).

References

Adamowski, J. F.: Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis, Journal of Hydrology, 353, 247–266, 2008.

Aubert, D., Loumagne, C., and Oudin, L.: Sequential assimilation of soil moisture and streamflow data in a conceptual rainfall–runoff model, Journal of Hydrology, 280, 145–161, 2003.

Bernard, S., Heutte, L., and Adam, S.: Influence of hyperparameters on random forest accuracy, in: International Workshop on Multiple Classifier Systems, pp. 171–180, Springer, 2009.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Calle, M. L. and Urrea, V.: Letter to the editor: stability of random forest importance measures, Briefings in bioinformatics, 12, 86–89, 2010.

Chen, X. and Ishwaran, H.: Random forests for genomic data analysis, Genomics, 99, 323–329, 2012.

Coulston, J. W., Blinn, C. E., Thomas, V. A., and Wynne, R. H.: Approximating prediction uncertainty for random forest regression models, Photogrammetric Engineering & Remote Sensing, 82, 189–197, 2016.

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y., and Wilby, R. L.: Flood estimation at ungauged sites using artificial neural networks, Journal of hydrology, 319, 391–409, 2006.

Di Luzio, M., Johnson, G. L., Daly, C., Eischeid, J. K., and Arnold, J. G.: Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States, Journal of Applied Meteorology and Climatology, 47, 475–497, 2008.

Dibike, Y. B. and Solomatine, D. P.: River flow forecasting using artificial neural networks, Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere, 26, 1–7, 2001.

Dingman, S. L.: Physical hydrology, Waveland press, 2015.

Elsner, M. M., Cuo, L., Voisin, N., Deems, J. S., Hamlet, A. F., Vano, J. A., Mickelson, K. E., Lee, S.-Y., and Lettenmaier, D. P.: Implications of 21st century climate change for the hydrology of Washington State, Climatic Change, 102, 225–260, 2010.

Falcone, J. A.: GAGES-II: Geospatial attributes of gages for evaluating streamflow, Tech. rep., US Geological Survey, 2011.

Graham, C. B., Barnard, H. R., Kavanagh, K. L., and McNamara, J. P.: Catchment scale controls the temporal connection of transpiration and diel fluctuations in streamflow, Hydrological Processes, 27, 2541–2556, 2013.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, Journal of Hydrologic Engineering, 4, 135–143, 1999.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of hydrology, 377, 80–91, 2009.

Huang, B. F. and Boutros, P. C.: The parameter sensitivity of random forests, BMC bioinformatics, 17, 331, 2016.

Hwang, S. H., Ham, D. H., and Kim, J. H.: A new measure for assessing the efficiency of hydrological data-driven forecasting models, Hydrological sciences journal, 57, 1257–1274, 2012.

James, G., Witten, D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning, volume 103 XIV of, 2013.

Johnstone, J. A.: A quasi-biennial signal in western US hydroclimate and its global teleconnections, Climate dynamics, 36, 663–680, 2011.

Karran, D. J., Morin, E., and Adamowski, J.: Multi-step streamflow forecasting using data-driven non-linear methods in contrasting climate regimes, Journal of Hydroinformatics, 16, 671–689, 2013.

Knowles, N., Dettinger, M. D., and Cayan, D. R.: Trends in snowfall versus rainfall in the western United States, Journal of Climate, 19, 4545–4559, 2006.

Knowles, N., Dettinger, M., and Cayan, D.: Trends in snowfall versus rainfall for the western united states, 1949-2001. prepared for California energy commission public interest energy research program, Trends in Snowfall Versus Rainfall for the Western United States, 1949-2001. Prepared for California Energy Commission Public Interest Energy Research Program, 2007.

Kuhn, M. et al.: Building predictive models in R using the caret package, Journal of statistical software, 28, 1–26, 2008.

Legates, D. R. and McCabe Jr, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water resources research, 35, 233–241, 1999.

Li, D., Wrzesien, M. L., Durand, M., Adam, J., and Lettenmaier, D. P.: How much runoff originates as snow in the western United States, and how will that change in the future?, Geophysical Research Letters, 44, 6163–6172, 2017.

Li, X., Sha, J., and Wang, Z.-L.: Comparison of daily streamflow forecasts using extreme learning machines and the random forest method, Hydrological Sciences Journal, 64, 1857–1866, 2019.

Liaw, A., Wiener, M., et al.: Classification and regression by randomForest, R news, 2, 18–22, 2002.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P.: Understanding variable importances in forests of randomized trees, in: Advances in neural information processing systems, pp. 431–439, 2013.

Lundquist, J. D., Dettinger, M. D., Stewart, I. T., and Cayan, D. R.: Variability and trends in spring runoff in the western United States, Climate warming in western North America: evidence and environmental effects. University of Utah Press, Salt Lake City, Utah, USA, pp. 63–76, 2009.

Mantua, N., Tohver, I., and Hamlet, A.: Impacts of climate change on key aspects of freshwater salmon habitat in Washington State, 2009.

Mass, C.: The weather of the Pacific Northwest, University of Washington Press, 2015.

Mentch, L. and Hooker, G.: Quantifying uncertainty in random forests via confidence intervals and hypothesis tests, The Journal of Machine Learning Research, 17, 841–881, 2016.

Mittermaier, M. P.: The potential impact of using persistence as a reference forecast on perceived forecast skill, Weather and forecasting, 23, 1022–1031, 2008.

Mosavi, A., Ozturk, P., and Chau, K.-w.: Flood prediction using machine learning models: Literature review, Water, 10, 1536, 2018.

Mote, P. W., Li, S., Lettenmaier, D. P., Xiao, M., and Engel, R.: Dramatic declines in snowpack in the western US, Npj Climate and Atmospheric Science, 1, 1–6, 2018.

Nicodemus, K. K.: Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures, Briefings in bioinformatics, 12, 369–373, 2011.

Obringer, R. and Nateghi, R.: Predicting urban reservoir levels using statistical learning techniques, Scientific reports, 8, 5164, 2018.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A.: How many trees in a random forest?, in: International workshop on machine learning and data mining in pattern recognition, pp. 154–168, Springer, 2012.

Pal, M.: Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26, 217–222, 2005.

Pan, M., Sheffield, J., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., Robock, A., Lohmann, D., Cosgrove, B., Duan, Q., et al.: Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent, Journal of Geophysical Research: Atmospheres, 108, 2003.

Papacharalampous, G. A. and Tyralis, H.: Evaluation of random forests and Prophet for daily streamflow forecasting, Advances in Geosciences, 45, 201–208, 2018.

Payne, J. T., Wood, A. W., Hamlet, A. F., Palmer, R. N., and Lettenmaier, D. P.: Mitigating the effects of climate change on the water resources of the Columbia River basin, Climatic change, 62, 233–256, 2004.

Probst, P., Wright, M. N., and Boulesteix, A.-L.: Hyperparameters and tuning strategies for random forest, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, e1301, 2019.

Ralph, F., Dettinger, M., White, A., Reynolds, D., Cayan, D., Schneider, T., Cifelli, R., Redmond, K., Anderson, M., Gherke, F., et al.: A vision for future observations for western US extreme precipitation and flooding, Journal of Contemporary Water Research & Education, 153, 16–32, 2014.

Rasouli, K., Hsieh, W. W., and Cannon, A. J.: Daily streamflow forecasting by machine learning methods with weather and climate inputs, Journal of Hydrology, 414, 284–293, 2012.

Regonda, S. K., Rajagopalan, B., Clark, M., and Pitlick, J.: Seasonal cycle shifts in hydroclimatology over the western United States, Journal of climate, 18, 372–384, 2005.

Rogelis, M. C., Werner, M., Obregón, N., and Wright, N.: Hydrological model assessment for flood early warning in a tropical high mountain basin, Hydrology and Earth System Sciences Discussions, 2016.

Schönfelder, L. H., Bakken, T. H., Alfredsen, K., and Adera, A. G.: Application of HYPE in Norway, SINTEF Energi. Rapport, 2017.

Seibold, H., Bernau, C., Boulesteix, A.-L., and De Bin, R.: On the choice and influence of the number of boosting steps for high-dimensional linear Cox-models, Computational Statistics, 33, 1195–1215, 2018.

Shortridge, J. E., Guikema, S. D., and Zaitchik, B. F.: Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds, Hydrology and Earth System Sciences, 20, 2611–2628, 2016.

Siqueira, V. A., Paiva, R. C. D. d., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., Paris, A., Calmant, S., and Collischonn, W.: Toward continental hydrologic–hydrodynamic modeling in South America, Hydrology and Earth System Sciences. Göttingen: Copernicus. Vol. 22, n. 9 (set. 2018), p. 4815-4842, 2018.

Srivastava, A., Wu, J. Q., Elliot, W. J., Brooks, E. S., and Flanagan, D. C.: Modeling streamflow in a snow-dominated forest watershed using the Water Erosion Prediction Project (WEPP) model, Transactions of the ASABE. 60 (4): 1171-1187., 60, 1171–1187, 2017.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC bioinformatics, 8, 25, 2007.

Tongal, H. and Booij, M. J.: Simulation and forecasting of streamflows using machine learning models coupled with base flow separation, Journal of hydrology, 564, 266–282, 2018.

Tyralis, H., Papacharalampous, G., and Langousis, A.: A brief review of random forests for water scientists and practitioners and their recent history in water resources, Water, 11, 910, 2019.

U.S. Geological Survey: U.S. Geological Survey, 2019, National Hydrography Dataset (ver. USGS National Hydrography Dataset Best Resolution (NHD) for Hydrologic Unit (HU) 4 - 2001), https://www.usgs.gov/core-science-systems/ngp/national-hydrography/access-national-hydrography-products, 2020.

Vano, J. A., Nijssen, B., and Lettenmaier, D. P.: Seasonal hydrologic responses to climate change in the P acific N orthwest, Water Resources Research, 51, 1959–1976, 2015.

Wager, S., Hastie, T., and Efron, B.: Confidence intervals for random forests: The jackknife and the infinitesimal jackknife, The Journal of Machine Learning Research, 15, 1625–1651, 2014.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X.: Flood hazard risk assessment model based on random forest, Journal of Hydrology, 527, 1130–1141, 2015.

Wenger, S. J., Luce, C. H., Hamlet, A. F., Isaak, D. J., and Neville, H. M.: Macroscale hydrologic modeling of ecologically relevant flow metrics, Water Resources Research, 46, 2010.

Zheng, X., Wang, Q., Zhou, L., Sun, Q., and Li, Q.: Predictive Contributions of Snowmelt and Rainfall to Streamflow Variations in the Western United States, Advances in Meteorology, 2018, 2018.

Figure 1. Structure of a RF and relevant parameters

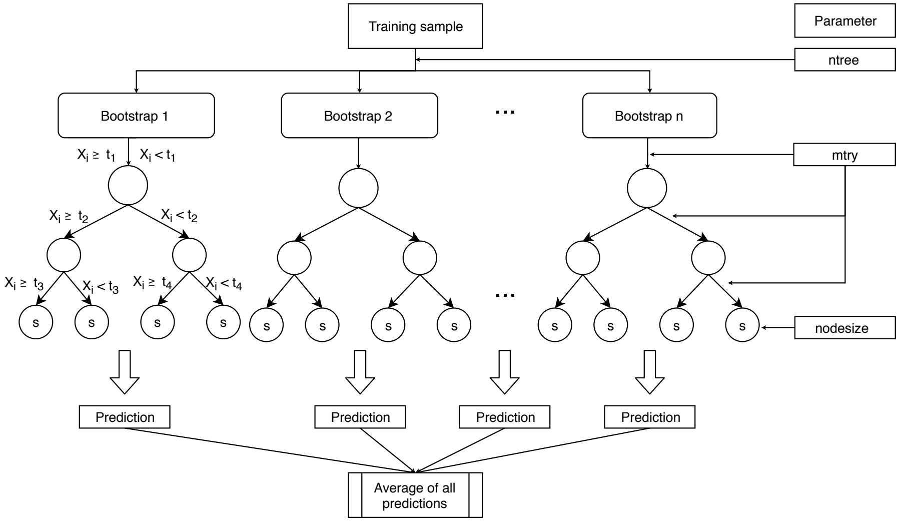
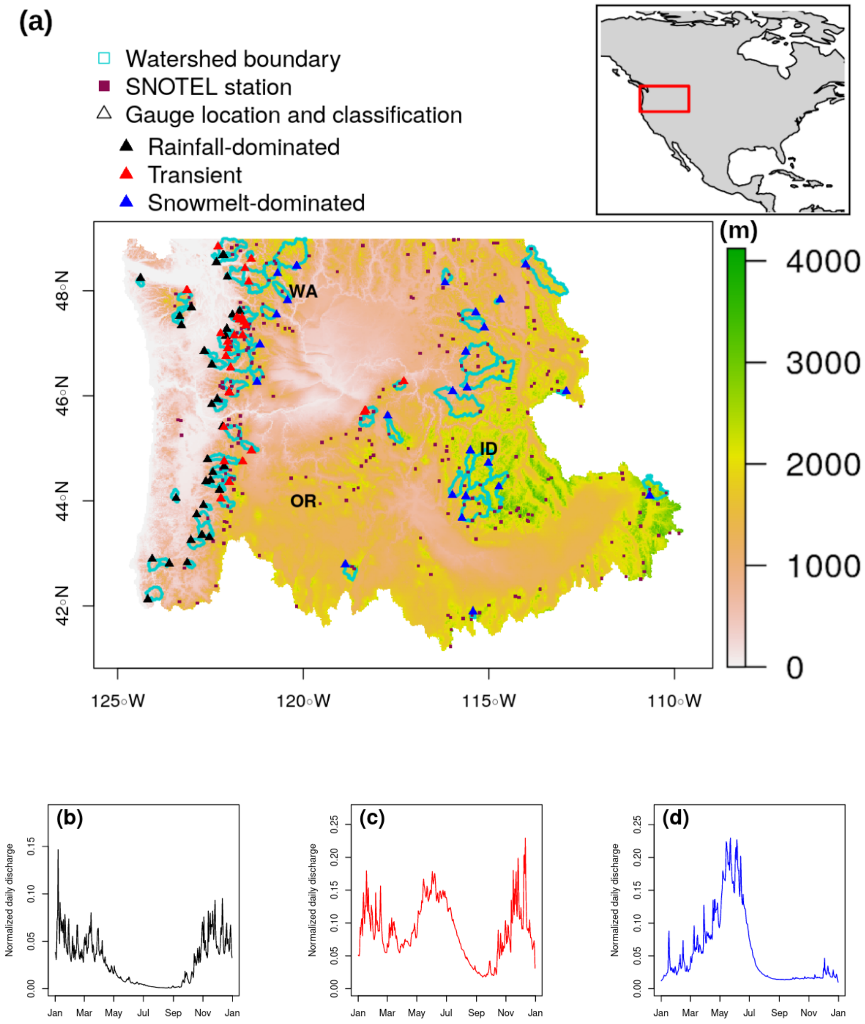
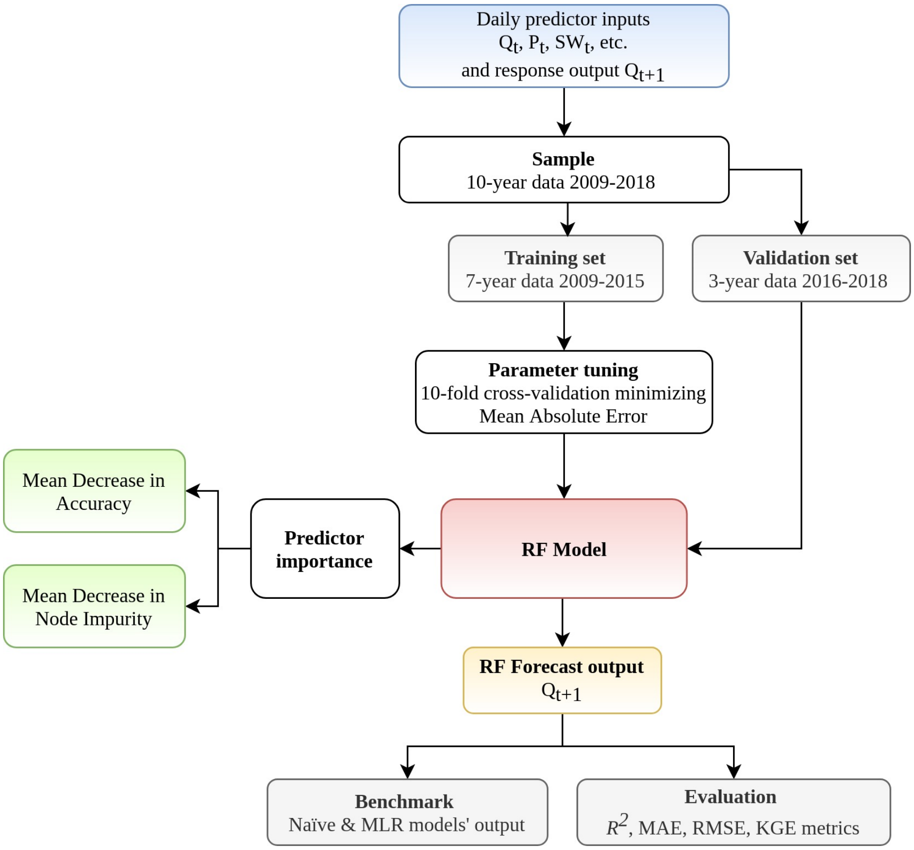


Figure 2. (a) Elevation (m) shading map showing the Pacific Northwest Hydrological Unit, 86 selected stream gauges (triangles), and their drainage area (cyan delination lines), and SNOTEL stations (brown squares). Examples of annual hydrographs of (a) rainfall-dominated, (b) transient regime, and (c) snowmelt-dominated watersheds. Figures (a-c) are based on 2009-2018 daily flow data at three sites 12043300 (124.4W, 48.2N), 12048000 (123.1W, 48N), and 10396000 (118.9W, 42.7N), respectively.



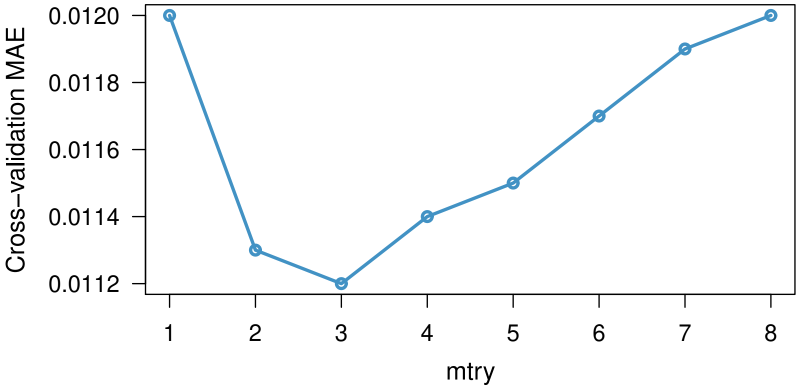
https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Figure 3. Flowchart showing the input-output model based on Random Forest



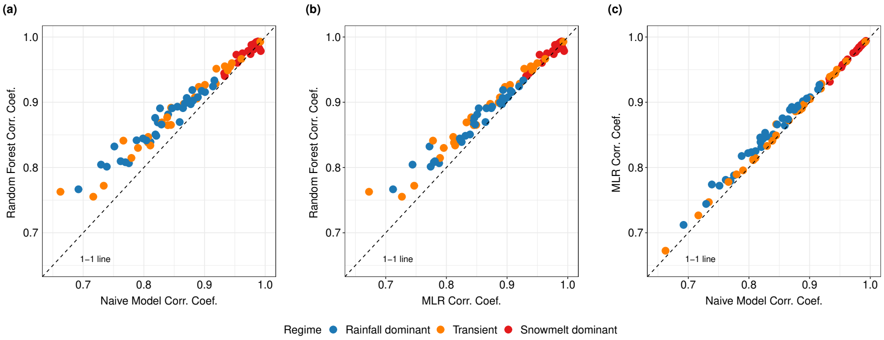
https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Figure 4. Out-of-bag Mean Absolute Error plotted against `mtry` during optimal parameter search at site USGS 12094000.



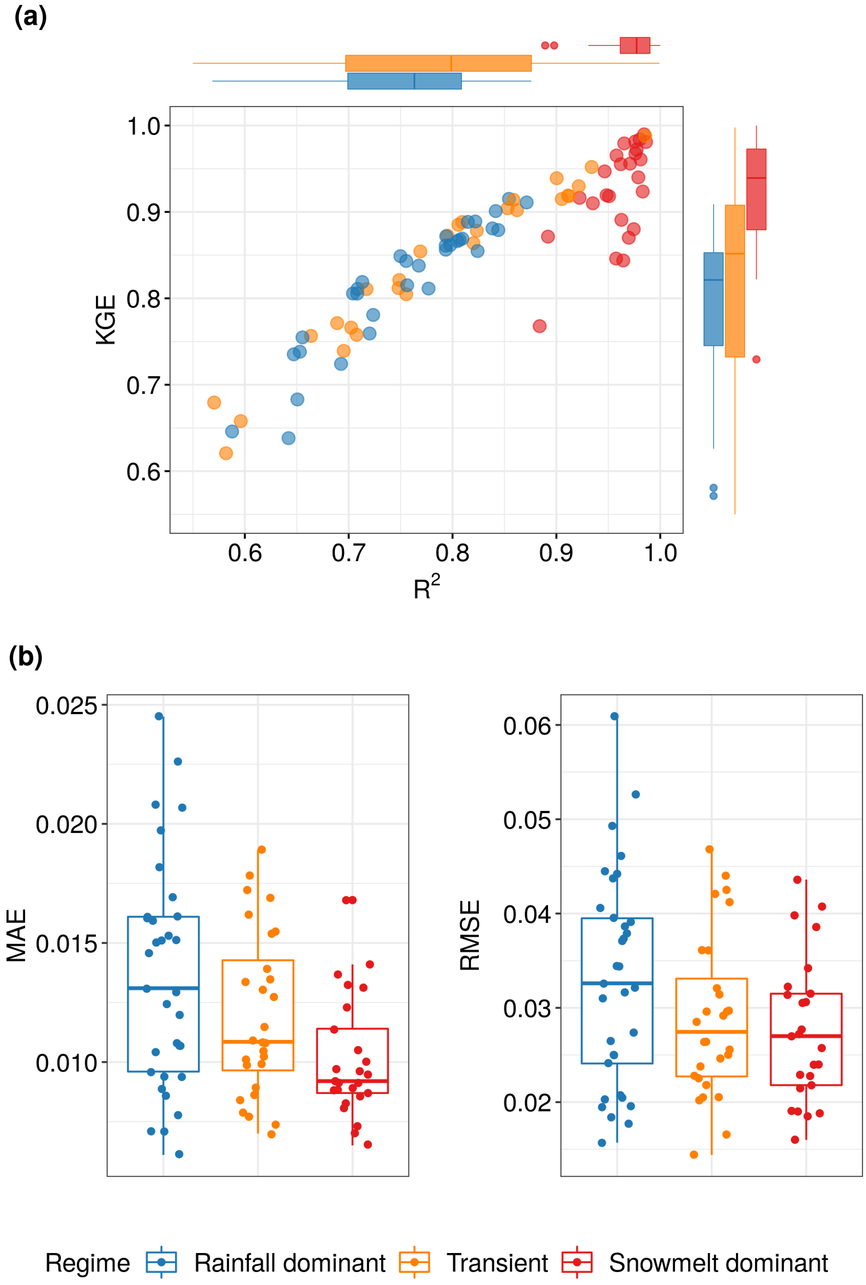
https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Figure 5. Pairwise scatter plots of Pearson correlation coefficient between forecasted and observed values for (a) Random Forest vs. Naïve Model, (b) Random Forest vs Multiple Linear Regression, and (c) Multiple Linear Regression vs. Naïve Model. Each dot represents one watershed (n=86).



https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Figure 6. Streamflow daily forecast scores computed over the validation period (2016-2018) for RF model across four metrics (a) R^2 and KGE (b) MAE and RMSE.



https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Figure 7. (a) Number of times RF *correctly* forecasted events that exceeded 90th, 95th, and 99th thresholds divided by the total number of exceedance. (b) Number of times RF *incorrectly* forecasted events that exceeded 90th, 95th, and 99th thresholds divided by the total number of non-exceedance.

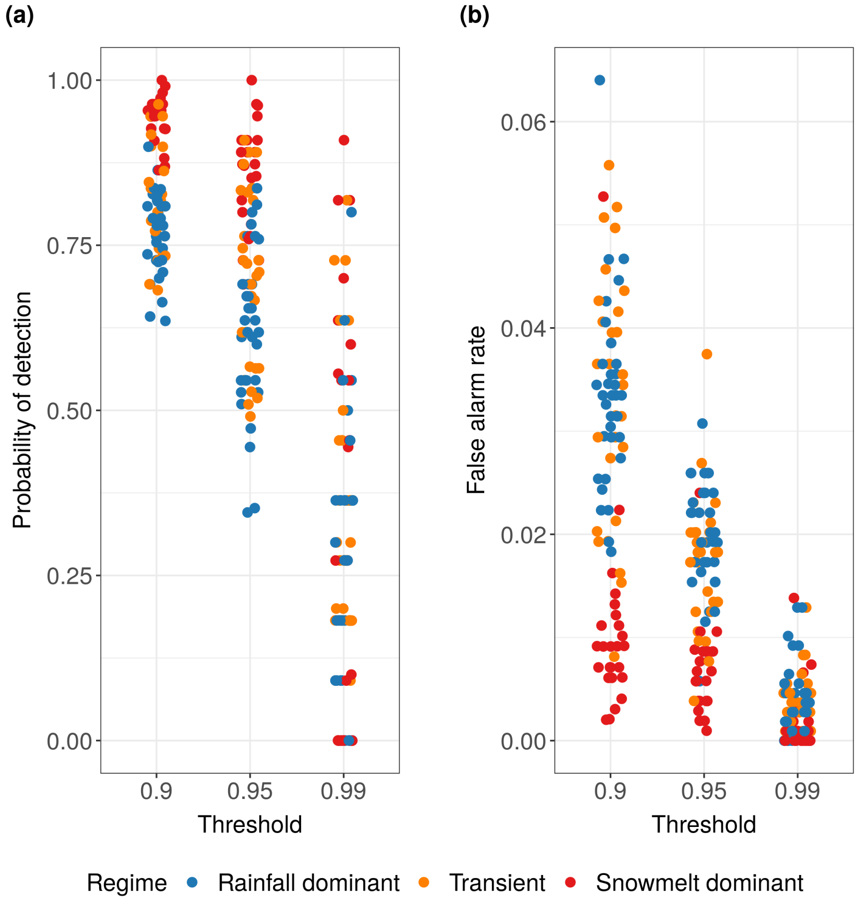
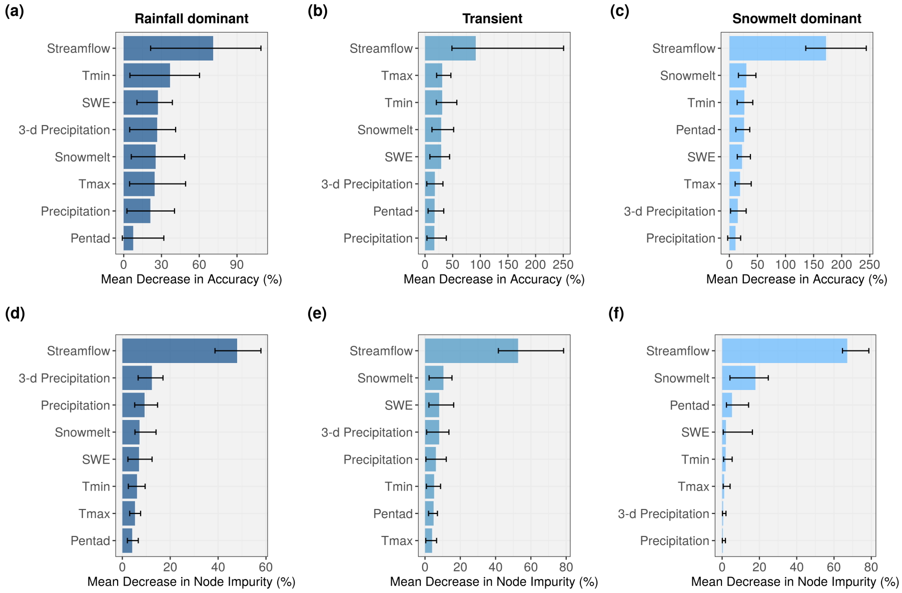


Figure 8. Barplots show importance of predictor variables using (a-c) DMA and (d-f) DMI metrics. Length of the blue bars indicates the median value across the watersheds for each flow regime and the thin black bar represents the range of the values.



https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Figure 9. KGE scores plotted against average percentage of sand in soil at each watershed. Best-fit lines were determined using simple linear regression.

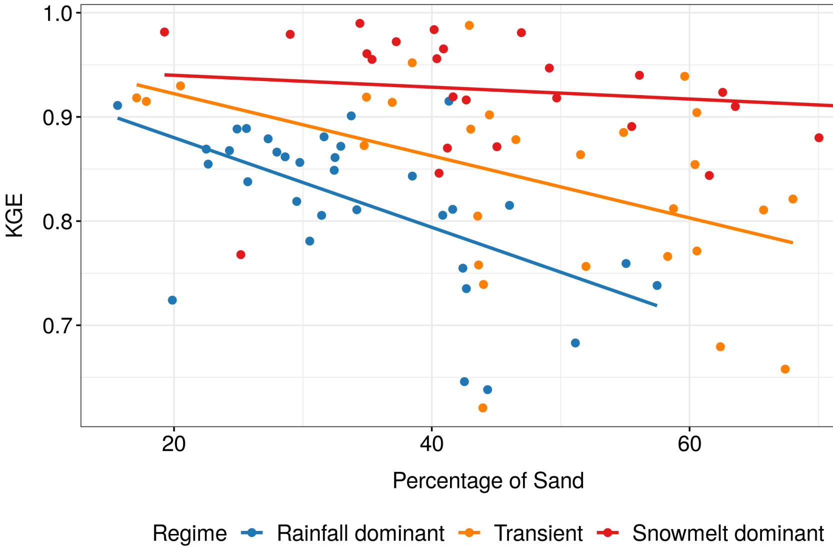


Table 1. Number of streamflow gauges used in the study for each flow regime, ranges of mean watershed elevation and drainage area. Complete catchment physical and hydro-climatic characteristics at individual site (retrieved from Falcone (2011).

Hydrologic regime	Number of gauges	Mean watershed elevation (m)	Drainage area (km^2)
Rainfall-dominated	33	239 - 1207	58 - 703
Transitional	28	813 - 1477	58 - 1855
Snowmelt-dominated	25	1349 - 2509	51 - 3355

https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Table 2. List of potential predictors.

No.	Predictors	Index	Unit	Source
1	Streamflow at day *t*	Q_t	ft^3/s	USGS
2	Precipitation	P_t	mm	PRISM
3	Sum of 3-day precipitation (P_t + P_{t-1}+ P_{t-2})	$P3_t$	mm	Derived from PRISM
4	Snow water equivalent	SWE_t	in	SNOTEL
5	Maximum temperature	$TMAX_t$	degree F	SNOTEL
6	Minimum temperature	$TMIN_t$	degree F	SNOTEL
7	Snowmelt (SW_t - SW_{t-1})	SD_t	in	Derived from SNOTEL
8	Pentad index	PEN_t	-	-

Table 3. The achieved parameter `mtry` using exhaustive-search strategy (`mtry` = {1,2,6,7,8} were considered but not found as the optimal value at any gauge).

`mtry`	Number of gauges	Median MAE
3	29	0.0127
4	44	0.0116
5	13	0.0079

https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Table 4. Descriptive statistics of the four metrics used to evaluate the overall performance of Random Forest: R^2, KGE, MAE, and RMSE.

Metric	Flow regime	Min	Q1	Median	Q3	Max
R^2	Rainfall dominant	0.59	0.71	0.77	0.81	0.87
	Transient	0.57	0.71	0.80	0.87	0.99
	Snowmelt dominant	0.88	0.95	0.97	0.98	0.99
KGE	Rainfall dominant	0.64	0.78	0.84	0.87	0.92
	Transient	0.62	0.77	0.86	0.91	0.99
	Snowmelt dominant	0.77	0.89	0.94	0.97	0.99
MAE	Rainfall dominant	0.0061	0.0096	0.0131	0.0161	0.0245
	Transient	0.0070	0.0097	0.0109	0.0143	0.0189
	Snowmelt dominant	0.0065	0.0087	0.0092	0.0114	0.0168
RMSE	Rainfall dominant	0.0157	0.0241	0.0326	0.0395	0.0609
	Transient	0.0144	0.0227	0.0275	0.0331	0.0468
	Snowmelt dominant	0.0160	0.0218	0.0270	0.0315	0.0436

https://doi.org/10.5194/hess-2020-305
Preprint. Discussion started: 23 June 2020

Table 5. Pearson correlation coefficient between KGE scores and selected basin variables. Highlighted red values indicate the relationship is significant at 5 percent or 1 percent level.

Watershed characteristics	Hydrologic regime		
	Rainfall dominant	Transient	Snowmelt dominant
Slope	-0.42	-0.68	0.12
Aspect eastness	-0.02	0.12	-0.12
Drainage area	0.14	-0.12	0.11
Basin compactness	0.09	-0.12	-0.16
Stream density	-0.10	0.29	-0.27
Percent of sand	-0.59	-0.46	-0.14
Percent of forested area	-0.11	0.32	0.32