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

Anonymous: This should not be capitalized.

Anonymous: This whole phrase should be hyphenated ("rainfall- and snowmelt-driven").

Anonymous: "Data-driven machine learning" is redundant, since all machine learning is data-driven. Also, "machine learning" should not be capitalized.

Anonymous: What do you mean by "performance"? Forecast quality, or computational efficiency?

Anonymous: streamflow-forecasting

Anonymous: random forests

Anonymous: forecasts

Anonymous: This phrase seems to be missing a word. Perhaps you mean many/diverse climatic conditions and physiographic settings?

Anonymous: comma

Anonymous: What does this mean? (I know you define the term later, but it should be defined at first mention. Alternatively, if you do not want to define jargon in the abstract, you could use a different term.)

Anonymous: Do not capitalize.

Anonymous: Delete comma.

Anonymous: Insert colon.

Anonymous: None of these terms should be capitalized.

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.

Anonymous: "Metric" has a specific definition in forecast evaluation, and not all these numbers are metrics. Replace with "Evaluation scores".

Anonymous: that the RF

Anonymous: The largest

Anonymous: is

Anonymous: Do not capitalize.

Anonymous: This acronym should be defined at the first mention of Kling-Gupta efficiency.

Anonymous: What does this mean? Is 0.62-0.99 good, or bad? Since most readers are probably unfamiliar with KGE, I suggest reporting the other evaluation scores here as well. Alternatively, you could just report percent improvement over the baseline models (since this is what really tells you how good the random forest is).

Anonymous: Delete (unnecessary).

Anonymous: Delete (unnecessary).

Anonymous: What are the "new insights"? This statement is very generic and could be found in almost any paper abstract, so you should make it a bit more specific.