I thank the authors for their hard work in revising the manuscript, and I find it much improved. However, I still have several major comments (listed below) and many minor comments (made inline in the manuscript; document below), so I recommend major revisions before publication. Note that I have made inline comments on some of the authors' responses, as well as the manuscript itself.

Major comments

1. Results still contain no significance tests or error bars. The authors claim that their random forest outperforms the two baseline models (persistence, which they call the "naïve" model, and linear regression), but there are no significance tests or error bars to support this claim. I made the same comment in round 1, and I find the authors' response (explaining why they chose not to include significance tests as requested) unsatisfactory. See page 6 of the document below for the authors' response (and my response to their response).

2. The authors' claim that random forests are more interpretable than other machine-learning models, is highly debatable. For details, see my first two comments on page 9 of the document below. This comment should be easy to address – I would just like to see the authors add a few sentences to the manuscript, discussing the controversy of model interpretability. *i.e.*, There are properties of random forests that make them more interpretable than other ML models, but there are also properties that make them less interpretable, so it's unfair to simply say "RF allows for some level of interpretability" (as their justification for using random forests) and then move on.

3. The authors have not clarified which data (training only or training plus validation) they use to compute standardization parameters – *i.e.*, to compute the minimum and maximum for min-max scaling. See my fourth comment on page 12 of the document below. The distinction is important. Only training data should be used to compute standardization parameters; if the validation data are also used to compute standardization parameters, this means that validation data are used to pre-process training data, which means that information from the validation data has "leaked" into the training data, which means that the two datasets are no longer independent.

4. The authors should justify why they experimented with only two hyperparameters. See my fifth comment on page 12 of the document below.

5. The authors use ROC curves to diagnose (the absence of) systematic overestimation, which is an invalid interpretation of ROC curves. See my comment on line 14 of the document below.

6. The "no-skill line" in the ROC curves (Figure 7) is misplaced. The no-skill line should be the $x = y$ line (or POD = false-alarm rate), which is the ROC curve that would be achieved by a random model such as a coin flip.

7. I don't understand why the authors include only 3 probability thresholds in their ROC curves (90%, 95%, and 99%). A typical ROC curve includes probability thresholds spanning the range from 0-100% (typically in increments of 10% at most – usually in increments of 1% or 0.1%, which allows for a smoother curve). Seeing such a small portion of the full ROC curve, it is difficult to assess the models' performance. I request that the authors plot the full ROC curve for each model, like the ones shown/explained here: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. Plotting the full ROC curves would also allow

the authors to compute area under the ROC curve (AUC), which is a scalar typically used to quantify the goodness of a ROC curve.

8. As in round 1, I request that the authors include performance diagrams along with the ROC curves (see my last comment on page 13 of the document below). Performance diagrams plot probability of detection (POD) on the *y*-axis vs. success ratio (1 minus false-alarm **ratio**, which is different than false-alarm **rate**) on the *x*-axis. Since frequency bias = POD / success ratio, performance diagrams can be used to diagnose systematic overestimation, which the authors have identified as a goal. Performance diagrams look like Figure 4 in this paper: https://journals.ametsoc.org/view/journals/wefo/34/6/waf-d-19-0094_1.xml?tab_body=fulltext-display. If it proves too complicated to plot the contours in the background (frequency bias and critical success index), I would be okay with the authors leaving the contours out.

9. The discussion on lines 384-391 is unclear throughout. See inline comments.

10. Mean-flow benchmark. On lines 331-332, the authors cite a previous paper to claim that a Kling-Gupta efficiency (KGE) $\geq$ -0.41 improves upon the “mean-flow benchmark,” which is the KGE achieved by always predicting the time-mean flow (“climatology”) at the given basin. I request that the authors compute the mean-flow benchmark for their own dataset, since it may be different than the dataset used in the paper they cite.

We would like to thank Editor Dimitri Solomatine and two reviewers for their time and valuable feedback to our manuscript. The manuscript has benefited from the informative suggestions. We appreciate the opportunity to submit a revised version for further consideration. Please find below a list of our answers to the comments and changes to the paper carried out to carefully address the remarks and the suggestions proposed by the two referees. The comments are shown below in black italic font. Our responses and revisions are presented below each comment. A marked-up version of the manuscript showing the specific changes we made is submitted along with this letter. We hope that this will allow the editor to assess the revision.

1 Referee 1

1.1 Overall comments

Pham et al. developed a Random-Forest-based (RF) algorithm for day-ahead streamflow forecasting. The algorithm employs 8 weather-snow-time features (see Table 2) and was tested across 86 watersheds in the Pacific Northwest (PNW), where it was compared with multilinear regression and previous-day streamflow as a minimum information forecasting approach (so called Naïve method). Results show that RFs provide quite robust predictions across catchments with different climatology (rainfall dominated, snowfall dominated, or transient) and generally perform better than the two benchmark approaches – especially in rainfall-dominated and transient watersheds. Drops in accuracy for RFs were correlated with watershed slope and sandiness. This is an interesting, well-written, and concise paper about an emerging machine learning technique in hydrology and its use for streamflow forecasting. While I have tangential knowledge of RF technical issues, I found the description of the algorithm clear and rigorous, which facilitates replicability and ultimately allows readers to learn more about RFs in general rather than only looking at is as a black box (note that these technical details are often bypassed or heavily summarized in other papers I have read on this matter).

Thank you for the review and supportive and constructive feedback. Yes, we too thought it would be helpful to provide this level of detail about the regression algorithms.

Also, RFs and machine-learning approaches in general are on the rise in hydrology, meaning I expect the paper to have some impact on the community.

Yes, we agree, these underused tools offer some exciting opportunities in this field. Tyralis et al. [2019] reviewed and compiled the applications of RF in water resources, where the number of papers has risen exponentially in the period between 2000 and 2019 (Figure 2). The authors of this study also acknowledged in their conclusion that, "it is quite remarkable that only a few studies recognize possible shortcomings of random forests and their variants." The analysis of the two commonly used measures of predictor importance(permutation-based Mean Decrease in Accuracy and Gini-based Mean Decrease in Node Impurity) in our manuscript suggests they might not be reliable and such analysis should be coupled with the understanding of the physical processes under study. We believe identifying and providing evidence for this shortcoming constitute a contribution to the current literature and for future research on the applicability of RF.

There are still some major and minor comments that I recommend for authors (see below), and I recommend the editor reconsider this manuscript after minor revisions

We appreciate the comments. We clarified and revised the manuscript according to your suggestions.

1.2 Major comments

1.The manuscript sometimes reads like a technical note, as it describes the algorithm and its implementation in great details but ultimately falls a little short on hydrological process interpretation. I see that the main goal of the paper is testing an algorithm, and applied research is certainly within the scope of HESS. And yet, I feel like implementing RFs across 86 watersheds with different characteristics and 10 yrs with different climatology without looking more specifically at how performance changes across the landscape and between years with different characteristics is kind of a missed opportunity.

In the manuscript, we tried to maintain a focus on (1) implementating and evaluating Random Forest for 1-day streamflow forecast and (2) exploring how the variations in hydrological processes across watersheds (e.g., contributions of rainfall and snowfall to streamflow, physical factors) might affect the model performance. In the Introduction, we provided an synopsis on the climatology of the Pacific Northwest and the roles of snowmelt and rainfall in driving streamflow dynamics. In Section 3.1 Study

Area and Data, we described the regional hydrographic systems of the Columbia River Basin. We also included physical characteristics of the watersheds under study, which were compiled and documented in the GAGESII Dataset (`https://water.usgs.gov/lookup/getspatial?gagesII_Sept2011`) and available in Supplementary material. We further systematically classified the watersheds into three hydrologic regimes: snowmelt-driven, rainfall-driven, and transient. As seen in Figure 2, the majority of the rainfall-dominated and transient watersheds are located on the west side of the Cascades while snowmelt-dominated watersheds are east of the Cascades and at higher elevation (Table 1). We reported the performance of RF for each watershed and explored how factors such as drainage area and slope could affect model performance in Section 4.6. Wenger et al. (2009) simulated and evaluated runoff at 55 watersheds in the Pacific Northwest using variable infiltration capacity (VIC) model and followed a similar classification scheme. Summary statistics of individual watersheds were reported similarly in Table A1 (`doi:10.1029/2009WR008839`). While we also thought about exploring the inter-annual and seasonal variability in model performance, reporting multiple sets of evaluation metrics (e.g., KGE, R^2, RMSE, MAE) for 86 watersheds would become confusing and dilute the scope of the paper.

I was a little surprised by the choice of benchmark models and particularly by the fact that authors did not consider a full hydrologic model. I understand that authors would probably like to stay within the realm of data-driven models, but a Naïve approach looks very simplistic, especially at a daily time scale and in basins where rainfall and snowfall coexist.

We did not consider a full hydrologic model and instead selected Naïve model as a benchmarking model for two reasons. First, as we briefly mentioned in the Introduction, both data-driven and hydrologic models have their advantages and disadvantages, and yet our objective was not to show one is superior to the other. Safeeq et al. [2014] provides an assessment of Variable Infiltration Capacity (VIC) model for predicting hydrologic regimes of 217 watersheds in the Pacific Northwest at 1/16 degree (6km×6km) grid-scale resolution. The author discussed the large underestimation of simulated SWE from meteorological data forcing compared to that at SNOTEL sites, which could be contributing to model bias. As we supplied RF with SNOTEL data, a comparison on runoff forecast between two models would be inappropriate. Second, we showed that, to our surprise, RF was not able to beat a simple Naïve model in its forecast among snowmelt-driven watersheds with strong persistence (Figure 5). We observed in some other papers where multiple ML models were tested and compared against one another without a baseline model. We wondered how much of the prediction were actually due to persistence in streamflow. As Pappenberger et al. [2015] pointed out, "benchmarking with simpler models can be viewed as a gain-based approach." For this reason, we believe the use of the Naïve model is justified.

How can this approach predict, e.g., intense rain-on-snow events that are ubiquitous in the PNW?

While this is a relevant question given the climatology of the PNW and the potential risk of rain-on-snow (ROS) floods, ROS events themselves are quite complex hydrometeorological phenomenon. Since we supplied RF with 1-day antecedent streamflow, SWE conditions, and meteorological data (precipitation and temperature), we would not expect the model to predict these ROS events. This would require additional future weather forecast information and fall under real-time forecast, which is out of the scope of our approach. We discussed the potential of including $t+1$ precipitation forecast on line 377-383 under Section 4.7 Limitations and future research.

Most flood-forecasting tools I have been exposed to use full hydrologic models, and I encourage authors to at least discuss this matter in their manuscript.

Good suggestion. We have now added the following discussion on the application of hydrologic models in this region under Introduction section. *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 in Western USA, 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]. Both watershed-scale hydrologic models and statistical models have been used to access the current and future stream hydrology and associated flood risks [Salathé Jr et al., 2014, Wenger et al., 2010, Tohver et al., 2014, Pagano et al., 2009]. Safeeq et al. [2014] simulated streamflows using the Variable Infiltration Capacity (VIC) model at* $1/16°$ *and* $1/20°$ *spatial resolutions and evaluated against observed values from 217 watersheds at at annual and season time scales. The study found the model was able to capture the hydrologic behavior of the study watersheds with reasonable accuracy. Yet authors recommended careful site-specific model calibration using not only streamflow but also SWE data would be expected to improve model performance and reduce model* *bias. Pagano et al. [2009] applied Z-Score Regression to daily SWE from SNOTEL stations and year-to-date precipitation data to predict seasonal streamflow volume in unregulated streams in Western US. Authors reported the skill of these forecasts was comparable to the official published outlooks. A natural question is whether ML models can produce comparable performance in these watersheds where streamflow contributions come from a mix of snowmelt and rainfall, as well as where snowmelt dominates sources.*

3. Relatedly, I was also a little surprised that authors did not consider rainfall and snowfall as separate features in their model (see their Table 2). I am aware that PRISM only provides total precip, but it also provides temperature and relative humidity that could be employed to separate snowfall from rainfall. Perhaps considering SWE already makes up for this, but I encourage authors consider this at least for future work. This was again particularly puzzling to me given the well-known role of rain on snow in this region.

We did consider the need to differentiating falling precipitation as rainfall from snowfall using surface air temperature data from PRISM as suggested. However, there are certain drawbacks of this approach. We expect the shifts in temperature can be dramatic at daily time scale, especially in the mountainous region with complex terrain, and may not be well captured by PRISM data. Moreover, watersheds in this study can span a wide range of elevation, making the determination of threshold temperature difficult. We chose not to explicitly differentiate the precipitation types and alternatively included maximum temperature (Tmax) and minimum temperature (Tmin) recorded at SNOTEL stations as predictors in the model. While it is uncertain to say whether the model was able to pick such signal due to the "black-box" nature of RF, we can see in Figure 8 that Tmin variable is ranked high among Snowmelt-dominant watersheds in variable importance. But this could also an indication of better prediction due to snowpack's sensitivity in temperature shifts rather than rain-snow differentiation.

1.3 Specific comments

We here address the specific comments. Simple typo fixes and clarifications were made directly in the revised manuscript.

- Title: I have usually seen "rainfall-dominated" and "snowfall-dominated" being used, rather than rainfall and snowmelt driven. Consider revising.

We used "snowmelt-dominated" and "snowmelt-driven" interexhangably throughout the manuscript. We included here two publications that employed similar usage, *"Soil moisture states, lateral flow, and streamflow generation in a semi-arid, snowmelt-driven catchment"* and *"Climate change impacts on the hydrology of a snowmelt driven basin in semiarid Chile"*.

- Line 24: I believe even ML algorithms need the formulation of some mathematical equations, although maybe not in a predictive role.

We agree and deleted "formulation of mathematical equations" from the sentence.

I think "ease of application" might be relative, especially in ungauged areas or for users with limited computational capabilities. Consider revising or expanding

The sentence has been revised to, "Among other qualities, the popularity of ML for such applications is due to the methods' competitive performance compared with alternative approaches, relative ease in implementation, and less strict distributional assumption."

Line 38: maybe also mention glaciers here, although they might not be an important driver for hydrology in your study region.

It's a good point and by saying, "most ML streamflow forecast applications are limited to watersheds where rainfall is the *major* contributor," we acknowledge there are other sources that can contribute to streamflow.

- Line 56: indeed, statistical forecasting models are widely used across the western US to predict summer flow (e.g., April to July total runoff). I understand this is out of the scope of your paper, but maybe mention this application to provide broader framing to your work

We appreciate the recommendation and mentioned a statistical forecasting model along with application of physical models in the Introduction.

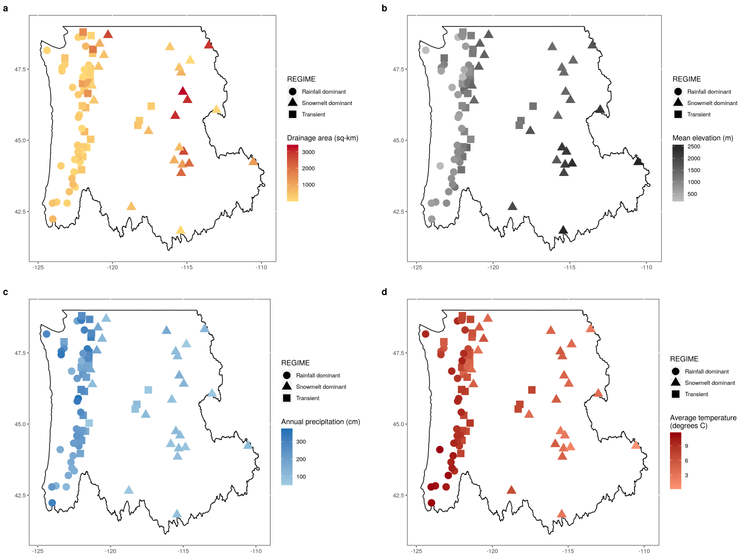
Line 149: Knoben at al. (https://hess.copernicus.org/articles/23/4323/2019/) have recently pointed out that KGE = 0 has a different implication from NSE = 0, and so KGE = 0 should be used with caution. Please revise as relevant.

Thank you for poiting this out. The following is our revision, "KGE metric ranges between -inf and 1. While there currently is not a definitive KGE scale, Knoben et al. (2019) showed KGE values in the range between -0.41 and 1 indicate the model a model improves upon the mean flow benchmark , which assumes the predicted streamflow values equal to the mean of all observations. Generally, KGE value of 1 suggests the model can perfectly reproduce observations."

Line 176: maybe some more quantitative climatology would be more appropriate here. For instance, replace "ample amount of winter precipitation" with statistics of winter precip for your watersheds. Same for "mild temperature". It would also be interesting to provide some statistics of mean-max SWE across the basins.

We updated Table 1 to include summary statistics for mean annual temperature and mean annual precipitation across three hydrologic regimes. We also added the following plot to the manuscript. SWE from SNOTEL stations was calculated at HUC-6 level and min-max statistics would be available in Supplementary.

Figure 1: Gauge locations with color gradient indicating variations in (a) watershed drainage area, (b) mean watershed elevation, (c) mean watershed annual precipitation, and (d) mean watershed annual temperature.



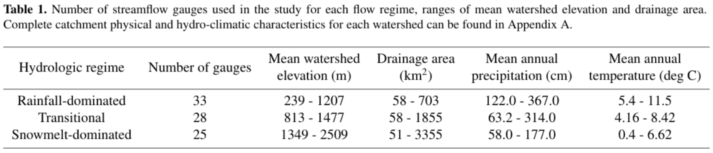


Table 1. Number of streamflow gauges used in the study for each flow regime, ranges of mean watershed elevation and drainage area. Complete catchment physical and hydro-climatic characteristics for each watershed can be found in Appendix A.

Hydrologic regime	Number of gauges	Mean watershed elevation (m)	Drainage area (km^2)	Mean annual precipitation (cm)	Mean annual temperature (deg C)
Rainfall-dominated	33	239 - 1207	58 - 703	122.0 - 367.0	5.4 - 11.5
Transitional	28	813 - 1477	58 - 1855	63.2 - 314.0	4.16 - 8.42
Snowmelt-dominated	25	1349 - 2509	51 - 3355	58.0 - 177.0	0.4 - 6.62

Line 186: have you tried to impute missing values? What's the impact of gaps in your framework?

In the initial screen, we selected gauges that "have less than 10 percent of missing data". The majority of 86 gauges that were eventually included in the study had continuous record or less than 3 percent missing data. Therefore, we do not think imputation was necessary.

Line 196: how "place-based" is this classification based on the day of the water year? I would have expected one to classify basins based on proportion of rainfall over total precipitation, which look more general to me.

This classification scheme were applied to streams in the Pacific Northwest from previous studies and based on the shape of the mean annual hydrograph. We provided examples of three types of hydrographs in Figure 2 (b-d). Classifying the watersheds based on proportion of rainfall over total precipitation is another approach but this also requires differentiating rainfall from snowfall at daily timescale. We addressed a shortcoming with this approach above.

- Line 246: how were the validation and calibration period chosen? May this choice have played a role in your results? What were the climatological characteristics of these two periods? Please expand and support your choice here.

Seven years and three years of data were used for training and validating the model, respectively. While there is no clear requirement, this selection is consistent with convention in which training dataset should be larger than validation data set to ensure that the model is exposed to a wide range of hydrological conditions. We tested the another partition scheme where we reserved 8 years (2009-2016) for training and 2 years (2017-2018) for validating. The KGE scores across the watersheds typically improved within 0.05 margin. However, we felt 2 years of observations (approximately 700 data points) would not be sufficient to evaluate the model and decided to go with the former scheme, which might be more conservative. Moreover, our forecast is 1-day ahead so we don't expect the wet year vs dry-year characteristics, which is defined and takes place at annual time scale, would play a major role on the model's performance.

- Line 269: I might have missed this, but do you show any statistics of persistence for your catchments to support this statement? Again, I may be missing something here.

We discussed this at lines 279-285.

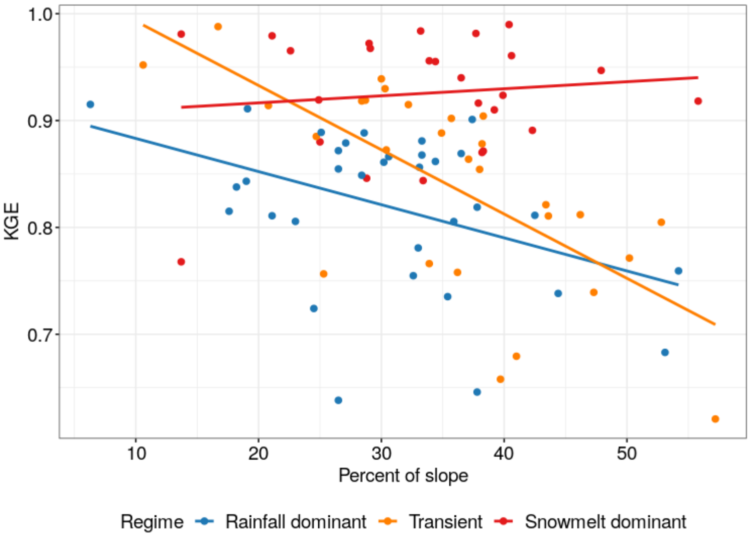
- Line 307: may these be due to rain on snow?

This is a possibility.

Figure 9: consider adding the scatter plot for slope

Thanks for the suggestion. We will add the following scatter plot slope vs. KGE in the revised manuscript.

Figure 2: KGE scores plotted against average percent of slope at each watershed. Best-fit lines were determined using simple linear regression.



- Table 1: is there any reason why snowmelt-driven catchments have a larger range of drainage areas? Just out of my curiosity.

We believe this is because the Columbia River and its major tributaries including the largest branch, the Snake River, flow through east of the Cascades and on the Oregon-Idaho border where most of the the snowmelt-dominated watersheds in our study are located.

- Table 5: what is the data source for these characteristics? Especially sandiness and forested area

The data from this table was part of the USGS GAGEII dataset, which was cited on line 200 and above. Percentage of sand in soil was estimated from the Department of Agriculture Digital General Soil Map of the United States or STATSGO2 Database. The data can be found in the submitted Supplementary document. Percentage of forested area was estimated using the National Land Cover Database 2006 classification. We added the following clarification to the manuscript, *"These watershed characteristics were compiled as part of GAGESII dataset using national data sources including US National Land Cover Database (NLCD) 2006 version, 100m-resolution National Elevation Dataset (NED), and Digital General Soil Map of the United States (STATSGO2)."*

2 Referee 2

2.1 Overall comments

This paper describes the use of random forests (ensembles of decision trees) to predict streamflow in various basins in the Pacific Northwest at one-day lead time. The authors use two methods to understand the most important predictor variables, and they also investigate the effect of other basin characteristics (not included as predictor variables) on the performance of the random forest. With some improvements this work could be a valuable contribution to the literature, especially given the analyses of predictor importance and confounding variables (basin characteristics not included as predictors). However, at this time I have chosen to reject, due to several major issues with the paper. Major comments are summarized below, and inline comments are attached in a PDF.

We appreciate the reviewer's kind words on the merit of the manuscript and its contribution to the literature. While we believe there are instances in the manuscript that can be improved, most of them can be addressed with better clarification. Please see the Inline comments section below.

2.2 Major comments

1. The paper has serious grammatical issues, which make it difficult to follow. I have pointed all grammatical errors in the abstract (see `pham2020_annotations_abstract.pdf`*). After the abstract, I have mostly abstained from pointing out every grammatical error. However, the frequency of grammatical errors is approximately the same throughout the paper. I would like to make it clear that I am not rejecting on the basis of grammar alone, but it does make the paper difficult to follow (there are many sentences that I simply do not understand).*

Thanks for carefully reviewing our manuscript. We have revised these errors based on your suggestions and believe the overall language is now clearer and improved.

2. The explanation of machine-learning methods (the random forest and associated predictor-importance methods) is unclear and contains several false statements. See inline comments for more detail. The explanation is probably not clear enough for readers unfamiliar with ML to follow, and it contains enough false statements that readers familiar with ML will probably be left scratching their heads.

We believe our description and discussion of random forests algorithm is sufficient given the scope of the paper. We also clarified statements the reviewer considered "false" and supported with cited literature.

3. No significance-testing. The authors claim that their random forest outperforms the two baseline models (persistence, which they call the "naïve" model, and linear regression), but no significance-testing is conducted to support this claim. Especially for the comparison of the random forest with linear regression, the numbers are close enough (see line 277) that I doubt the differences are statistically significant.

Our comparisons for the three models using correlation coefficient statistics are consistent with published literature. Specifically, Yaseen et al. [2015] conducted a review of ML models for streamflow forecasting in the period 2000–2015 and noted, "It was also found, the majority of the reviewed articles evaluated based on the correlation coefficient (R) and Root Mean Square Error (RMSE) in addition to other evaluation criteria (e.g., Relative error; RE, Mean Absolute Error; MAE, Mean Square Error; MSE, Nash–Sutcliffe Coefficient; NS, Sum of Square Error; SSE), clearly illustrated in (Fig. 1). The modeling that yields a maximum value of correlation coefficient and minimum values of RMSE and MAE presented a good evaluation of model performance." Also, result of a significance test for the differences between the correlation coefficients does not provide useful information. R, in this case, is an evaluation score for the models similar to KGE, RMSE, and MAE. Because we used persistence model as baseline, the positive difference in R values between the two respective models can be interpreted as a gain in forecast skillfulness. It would not make sense to perform a statistical significance test for the difference between two evaluation scores of two models otherwise.

Anonymous: This response justifies the evaluation scores used, which is a straw-man argument. I have no issue with the evaluation scores; my issue is with the lack of significance-testing. It is currently impossible to tell if any of the inter-model differences are statistically significant or due to noise.

Anonymous: ? I don't understand what "otherwise" means in this sentence, and throughout this response I see no justification for the lack of significance tests.

4. Interpretation of model performance is lacking in detail and contains several confusing statements. See inline comments on lines 274, 283, 287, 293, 295, 298, 302, 304, 308, 309, and 311.

We addressed these comments and suggestions below.

2.3 Inline comments

Simple typos and punctuation errors are corrected directly to the revised manuscript.

2.3.1 Abstract

Line 1: "Data-driven machine learning" is redundant, since all machine learning is data-driven. Also, "machine learning" should not be capitalized.

In this sentence, we included "data-driven" to contrast ML approach with the traditional hydrologic models, which are physically based and we later discussed this under the Introduction section.

Line 3: What do you mean by "performance"? Forecast quality, or computational efficiency?

We agree this was vague. Following is the revised sentence, "Among other qualities, the popularity of ML models for such applications is due to their relative ease in implementation, less strict distributional assumption, and competitive computational and predictive performance."

Line 6: This phrase seems to be missing a word. Perhaps you mean many/diverse climatic conditions and physiographic settings?

We revised it as following, "These watersheds cover diverse climatic conditions and physiographic settings."

Line 8: 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.)

We believe the current description, "timing of center-of-annual-flow volume" is adequate for the purpose of the Abstract.

Line 12-13: 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).

Although 0.62 - 0.99 KGE score range would be considered "good", we believe these values should be evaluated comparatively not absolutely. We also did not compute "percent improvement" in our analysis. In the interest of conciseness, we also only reported raw KGE score, which has increasingly been used as a metric to access model performance in hydrology and is more objective in our opinion, in the Abstract.

Line 15-16: 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.

We were referring to these insights: (1) "RF performance deteriorates with increase in catchment slope and increase in soil sandiness" and (2) "We note disagreement between two popular measures of RFvariable importance and recommend jointly considering these measures with the physical processes under study."

2.3.2 Manuscript body

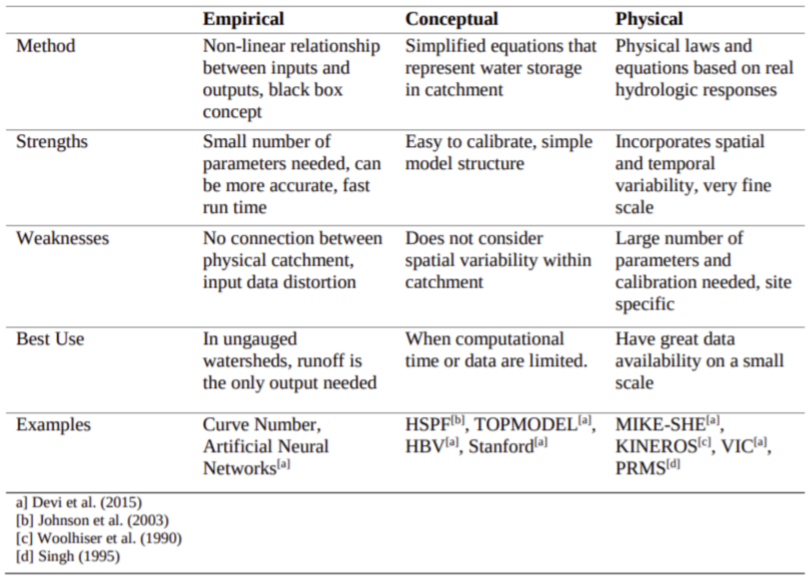
Line 23-24: This is very close to the definition of ML, although you do not explicitly say so.

You are right; we were giving the definition of ML models.

Line 24-25: This is a controversial statement. Many methods have been developed to interpret ML models and use ML to understand the underlying physical processes. Entire books have been written on the subject (`https://christophm.github.io/interpretable-ml-book/`).

We believe ML models do not provide the same level of interpretation as physical models. We included below a comparison table from the EPA that provides an overview on different rainfall-runoff model types [Sitterson et al., 2018]. ML models would fall under "Empirical" category. Also, Breiman [2001] himself wrote, "A forest of trees is impenetrable as far as simple interpretations of its mechanism go. In some applications, analysis of medical experiments for example, it is critical to understand the interaction of variables that is providing the predictive accuracy." He later discussed the application of permutation-based variable importance as a proxy to understand the input-output relationship. We discussed the limitation of this method in our manuscript under Section 4.5.

Table 1. Comparison of the basic structure for rainfall-runoff models



	Empirical	Conceptual	Physical
Method	Non-linear relationship between inputs and outputs, black box concept	Simplified equations that represent water storage in catchment	Physical laws and equations based on real hydrologic responses
Strengths	Small number of parameters needed, can be more accurate, fast run time	Easy to calibrate, simple model structure	Incorporates spatial and temporal variability, very fine scale
Weaknesses	No connection between physical catchment, input data distortion	Does not consider spatial variability within catchment	Large number of parameters and calibration needed, site specific
Best Use	In ungauged watersheds, runoff is the only output needed	When computational time or data are limited.	Have great data availability on a small scale
Examples	Curve Number, Artificial Neural Networks[a]	HSPF[b], TOPMODEL[a], HBV[a], Stanford[a]	MIKE-SHE[a], KINEROS[c], VIC[a], PRMS[d]

a] Devi et al. (2015)
[b] Johnson et al. (2003)
[c] Woolhiser et al. (1990)
[d] Singh (1995)

Line 27: What are other possible goals for ML? This will not be obvious to readers unfamiliar with ML.

We revised the sentence to, "ML models are particularly useful when accurate prediction is the central inferential goal (Dibike and Solomatine, 2001), whereas conceptual rainfall-runoff model can help gain a better understanding of hydrologic phenomena and catchment yields and responses (Sitterson et al., 2018).

Line 28: Neuro-fuzzy what? This is an adjective in a list of nouns.

Neuro-fuzzy refers to combinations of artificial neural networks and fuzzy logic and is commonly used as a noun in the literature. For example, Mosavi et al. [2018] discussed flood forecasting using ML models and wrote, "Many ML algorithms, e.g., artificial neural networks (ANNs), neuro-fuzzy, support vector machine (SVM), and support vector regression (SVR) were reported as effective for both short-term and long-term flood forecast."

Anonymous: Okay, but it still won't make sense to most readers to use "neuro-fuzzy" as a noun. Please replace in the text with "neuro-fuzzy (a combination of ANNs and fuzzy logic)".

Line 33: Which other models? Be specific. Your literature review should motivate the methods that you end up using.

We were referring to the two models, SVM and GP, from the previous sentence. In this paragraph, we would like to review the applications of various ML models in streamflow forecasting. We specifically discussed the reasons we used RF for our study on line 74-81.

Anonymous: This is still unclear in the text. Please replace "two other competing ML models" with "the other two ML models" or something similar. Currently it's not clear that "two other competing ML models" are the ones you mentioned in the previous sentence.

Line 35: Define.

We feel like this is not relevant because we did not use "baseflow separation" in our study. Readers who are interested in this method can look into it the referenced paper.

Anonymous: If you don't think "baseflow separation" is worth defining, please don't mention it. If it's worth mentioning, I think it's worth defining.

Line 46: If this the region shown in Figure 2? If yes, you should reference Figure 2 here. If no, you should include a map of the region in a different figure.

Yes, it is the same region in Figure 2 and we referenced it here in the revised manuscript.

Line 48: Does "unregulated" mean "not human-modified"?

Yes.

Line 56: "RF can be trained to forecast streamflow at various timescales, depending on the selection of input variables." I don't know what you mean by this. Random forests (and the individual trees therein) perform variable selection automatically, so random forests should not "depend on the selection of input variables".

The performance of the model to forecast "at various timescales" does depend on selection of predictor variables. For example, a study focuses on seasonal streamflow forecasting would consider including climate indices such as Southern Oscillation Index as one of the predictor variables.

Line 58: I don't understand why this prevents you from forecasting at longer lead times. All forecasts are made with antecedent information, but some phenomena can still be forecast skillfully at lead times much longer than 1 day.

In [Rasouli et al., 2012], the authors forecasted streamflow at 1-7 day lead times using three models: Bayesian neural network, support vector regression, and Gaussian process, and data from

Anonymous: Please include this explanation in the text. The following sentence is a non-sequitur, because like I said in the first round of reviews, using only antecedent information as predictors should not limit your lead time.

"We focus on 1-day lead time because we assume only antecedent information for predictors are available at the time forecast is made."

combinations of climate indices and local meteo-hydrologic observations. They concluded local observations as predictors were generally best at shorter lead times while local observations plus climate indices were best at longer lead times of 5–7 days. Also, the skillfulness of all three models decreased with increasing lead times. We cited this study on line 35. In our study, we focused on 1-day lead time forecasting and therefore did not include long-term climate information.

Line 63: This is a controversial point. Interpretation methods have been developed for many ML models. Also, many interpretation methods are model-agnostic and therefore can be applied to any ML model. In fact, one could argue that neural networks are more interpretable than random forests, since many interpretation methods rely on gradients (of the prediction with respect to model weights or input variables) and therefore cannot be applied to random forests, which are gradient-free models.

We understand you think the statement is controversial. However, we provided the reason why we believe RF "allows for some level of interpretability." This is delivered through its permutation-based and Gini-based variable importance measures, which have been used across disciplines (citied on lines 76-78). The permutation-based feature importance in particular was developed in the original paper and also included as one of the model-agnostic methods from the book you cited above. The author also discussed the advantages and disadvantages of each method. While you are suggesting, "one could argue that neural networks are more interpretable than random forests, since many interpretation methods rely on gradients," it doesn't seem fair as this implies gradient-based methods are better than permutation-based or Gini-based methods.

Anonymous: You cannot use the existence of the permutation test as an argument that random forests are more interpretable than other ML models, because like you say in this paragraph, the permutation test is model-agnostic. So your only valid argument here is that the Gini-based importance test can be applied to random forests and not other models. However, there are many interpretation methods that can be applied to other models (e.g., neural networks) and not random forests.

Anonymous: Please include some of this discussion in the main text. ML interpretability is a very controversial topic, and I think it's unfair to simply say "RF allows for some level of interpretability" and then move on.

Line 66: What do you mean by this? Internal parameters (those adjusted by training), or hyperparameters?

While we are aware of the term "hyperparameter" that is used in the ML literature, we chose to adhere to the original usage of "parameter" in [Breiman, 2001] and randomForest R package description [Liaw et al., 2002]. These two parameters are discussed under Section 2 Methodology.

Anonymous: Please clarify this in the main text. ML specialists might be confused when they see the word "parameter" being used to mean hyperparameter.

Line 75: This is false. Random forests (at least the ones you trained) are supervised learning, because the correct answer is supplied for each training example.

As you said, we trained RF to perform supervised learning in our study. However, random forests can be used to perform supervised and unsupervised learning [Liaw et al., 2002, Criminisi et al., 2012].

Anonymous: "Semi-supervised training" means training with a small amount of labeled data and a large amount of unlabeled data, which you are not doing. In any case, I'm happy with the change made to the main text.

Line 75: What do you mean by this? Random forests have two parameters for each split node (the predictor variable and threshold), so a forest with 2000 trees has (10^4) *parameters at least.*

The term "non-parametric" does not suggest the model doesn't contain parameters. Rather, "Non-parametric methods do not assume any particular family for the distribution of the data and so do not estimate any parameters for such a distribution" [Altman and Bland, 1999].

Anonymous: Please clarify this in the main text.

*Line 76: This is false. The trees are always somewhat correlated, because there is overlap among training sets for the different trees (since training sets are resampled *with replacement* from the full training set).*

A fundamental element of random forests is the randomization of predictor selection at each split, thus minimizing the correlation among the trees. Breiman [2001] himself wrote, "The randomness used in tree construction has to aim for low correlation ρ while maintaining reasonable strength." Bharathidason and Venkataeswaran [2014] discussed the idea of only including uncorrelated trees in the forest, which was shown to improve the performance of the model. As you pointed out, "trees are always somehwhat correlated," we believe this is true and changed "uncorrelated" to "decorrelated". This is consistent with the description of RF in the Elements of Statistical Learning text [Friedman et al., 2001].

Line 83: How does this happen? How does each tree make a prediction for a new example? Please clarify.

We added the following sentence to the manuscript, "After all the trees are grown, the forests make prediction on a new data point by having all trees run through the predictors. In the end, the trees cast a majority vote on a label class for classification task or produce a value for regression task by averaging all predictions."

Algorithm 1: This algorithm will probably not be intuitive for readers unfamiliar with ML. I think a plain-language explanation, along with a figure, would be much better.

In the submitted manuscript, we included both plain-language explanation (lines 75:85) and a figure (Figure 1).

Line 88: This is not an estimate. It is called the "0.632" rule and has been mathematically proven: https://www.jstor.org/stable/2965703?seq=1

In the cited paper, the author discussed that the value "0.632+" was "estimated" using bootstrap. It is also shown in [Albert et al., 2008] that the probability of not selecting an event in the bootstrap procedure becomes $e^{-1} \approx 0.369$ or approximately 37%.

Line 97: Unnecessary detail, since the hyperparameter experiment you described could be implemented with a simple for-loop (does not require a special library).

We believe citation of packages used in the study is important for reproducibility.

Anonymous: Okay, but what did you do with the Caret package? Did you use the Caret package to train the random forests, or just to loop through hyperparameters?

Line 104: So you compute a different MSE for each tree, rather than computing one MSE for the whole random forest?

MSE is calculated at tree level because the OOB sample used to compute MSE is different for each tree.

Line 105: Is this method any different than the permutation test created by Breiman (2001)?

It is the same method.

Line 108: Not necessarily. If there are two highly correlated predictor variables (x_1 and x_2), permuting one of the two may not decrease the model's performance. For example, if you permute only x_1, even if x_1 is highly important, the model may still perform well by relying on x_2, since x_1 and x_2 contain a lot of redundant information.

We did not consider this and thank you for pointing it out. Boulesteix et al. [2012] discussed the challenge of accurately measuring variable importance in computational biology and bioinformatics studies when highly correlated predictor variables are involved. It is relevant in our study as we supplied the model with maximum temperature and minimum temperature, which are correlated. We will add a discussion of this issue under Section 4.5 Variable importance analysis.

Line 111: ? I generally don't know what you mean in this sentence.

We provided details in the next sentence. In regression decision tree, split only occurs when the residual sum of squares (from Step 3 in Algorithm 1) of two descendent nodes is less than that of their parent node. In other words, there is a reduction in residual errors and the MDI measures this reduction.

Line 114: This shouldn't matter if you have only one response variable, right? It should matter only if you have multiple response variables with different scales (e.g., one response variable that ranges from 0...1 and another that ranges from 500...5000).

We standardized all variables in the training and validation sets. Because of this, raw MDI does not have an associated unit and provides little interpretation. Scaled MDI, on the other hand, can be interpreted as the relative contribution, in percentage, of each predictor to the total reduction in node impurities.

Line 125: I suggest calling this the "persistence baseline," rather than the naïve model. The word "naïve" evokes naïve Bayes for many people.

This is valid but we believe naïve model is commonly used in the context of hydrologic forecasting. We also defined this terminology.

Line 125: What are these limitations? Please discuss. Model evaluation is very important, and the methods you use should be explicitly justified.

We added the following clarification,"Among the limitations, these measures were reported to be especially oversensitive to extreme values (outliers)."

Line 136: How?

We explained this for each evaluation metric in the following paragraphs in the manuscript.

Anonymous: In this case, please add "as explained in the remainder of this section" to the end of the sentence.

Line 139: Please define all variables in this equation, including the N and i.

We added the following to the sentence, " $\hat{y}_i$ and y_i are the forecasted and observed values at day i respectively, and N is total number of the observations during the validation period."

Line 145: What do you mean by this? More sensitive to outliers?

By "error", we were referring to the difference between the predicted and observed values ($|\hat{y}_i - y_i|$). Due to the squared operation, RMSE is therefore more sensitive to large errors.

Line 147: Please state the ranges and optimal values for MAE and RMSE, like you did for R^2. (I know that it's probably obvious to most readers, but it's a small amount of additional text and worth specifying.)

Actually, both MAE and RMSE depend on the raw value of response variable, y, which will vary from one study to another. They are better interpreted comparatively.

Anonymous: For both MAE and RMSE, the range is always [0, inf) and the optimal value is 0. (Of course most models are good enough that they do not produce errors of infinity, but they could.)

Line 160: I don't know what this means.

For the validation period, we calculated the 90th, 95th, and 99th percentile streamflow values at each watershed. These are considered thresholds. If an observed daily streamflow exceeded this threshold, it would be considered an extreme event.

Line 180: A buffering effect to what? What does the ocean "buffer".

We acknowledge the term “buffering effect” might be vague and revised the sentence to, “Proximity to the ocean creates a more moderate climate with a narrower temperature range, particularly in the winter.”

Line 221- 223: Predictors of what? Streamflow, or SWE?

They were included as predictors for the RF model. All eight predictors are listed in Table 2.

Anonymous: Please replace "as predictors" in this sentence with "as predictors in the random forest" or "as predictors for streamflow" or something similar.

Line 223: Why only the last measurement of each day?

We only supplied the last measurement from SNOTEL stations because not all predictors have sub-daily values.

Anonymous: Please put this in the main text.

Line 225: How big are these basins? Please show a map.

We added the following map and table to Supplementary material.

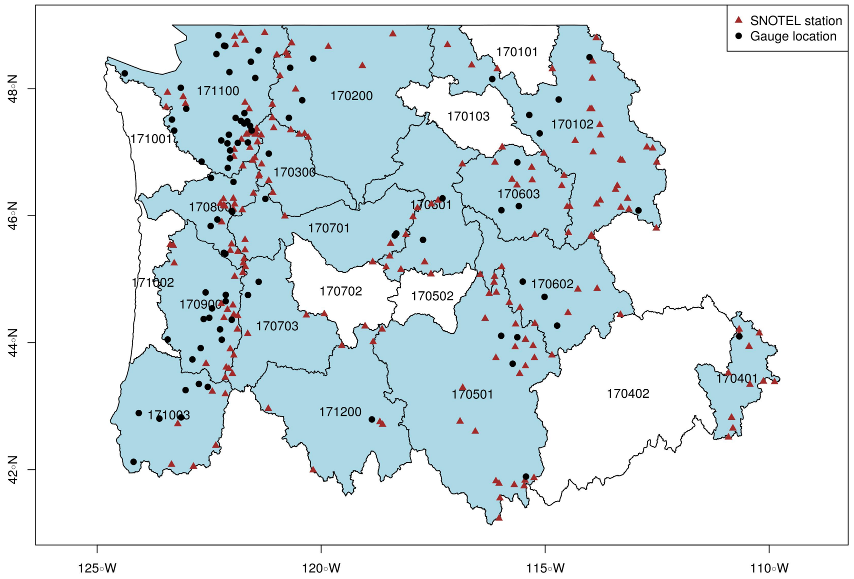


Figure 3: Map of basins within the Pacific Northwest Hydrological Unit. Blue basins contained at least one watershed and were included in the the study.

HUC-6	Name	Area (km^2)	Number of SNOTEL
170102	Pend Oreille	67598.70	30
170200	Upper Columbia	119755.57	10
170300	Yakima	15928.20	9
170401	Snake Headwaters	14812.20	11
170501	Middle Snake-Boise	85150.16	27
170601	Lower Snake	30198.02	7
170602	Salmon	36248.15	11
170603	Clearwater	24318.13	9
170701	Middle Columbia	29124.57	11
170703	Deschutes	27789.56	8
170800	Lower Columbia	16120.04	15
170900	Willamette	29697.66	15
171003	Southern Oregon Coastal	34510.01	6
171100	Puget Sound	52958.23	26
171200	Oregon Closed Basins	45143.34	6

Line 227: Can you discuss how much this affects the accuracy of your model? It seems like a major caveat.

This represents a shortcoming of the study due to limited spatial coverage of SNOTEL stations and the introduced uncertainty likely affects the accuracy of the model. We acknowledged this by stating, "The SNOTEL averages, therefore, represent first-order estimates of snow coverage and temperature conditions." We also considered an alternative approach by drawing information only from the SNOTEL station closest to gauge location but decided the basin-average better represented SWE conditions. Using basin-average SNOTEL SWE is consistent with previous studies in that focused on streamflow forecast [Abudu et al., 2011] as well as contribution of snowmelt to streamflow [Zheng et al., 2018] in western USA. Nevertheless, we believe supplying the RF with a more spatially consistent SWE data would improve model accuracy and is certainly worthy of future research. The reported RF performance in our study might be an underestimation.

Anonymous: Please state explicitly that this is a limitation of your work.

Anonymous: Please put this sentence in the conclusion, along with other future work.

Line 232: Why not use all predictors available? Random forests are not computationally expensive and perform predictor selection automatically.

We believe the current selection is appropriate. We also had to consider the practical purpose of the model. While there are other variables we could include such as soil moisture content, many would not be available for 1-day ahead forecasting in real time.

Line 235: Why use T_{min} and T_{max} as predictors if their only influence is on SWE (another predictor)? In that case you should just use SWE.

It's worth mentioning that these predictors are at 1-day lag. SWE only reflects the current state of snow condition and melting of snow is often triggered by changes in temperature. Given that there is high temporal correlation in daily temperatures, T_{min} and T_{max} data can provide useful signal to our streamflow forecast.

Anonymous: Please put this in the main text.

Line 237: This only tells me what a pentad is, not what the "pentad index" is. Please define the "pentad index".

In this case, the "pentad index" refers to the numerical sequence of pentads in a calendar year (1 to 73).

Line 239: What do you mean by "across gauges"? Is this correlation a spatial correlation, or is it computed at each gauge (in which case it's temporal but not spatial)?

Temporal correlation between daily streamflow and Pentad Index was computed at each gauge here.

Line 247: How? There are many ways to do min-max scaling. For example, what are the min and max values after standardization? Also, which dataset do you use to compute the min and max values for scaling? Just the training set, or both training and validation?

Anonymous: You still have not answered this question.

We added the following clarification to the manuscript, "We standardized training and validation data at each gauge using min-max scaling. The new data for all variables have values between zero and one."

Line 251: Random forests have many other hyperparameters: minimum sample size per split node, minimum sample size per leaf node, maximum depth, cost function, etc.).

We were aware of this but preferred to focus on the two parameters discussed in [Breiman, 2001].

Anonymous: Please clarify in the main text that there are other hyperparameters. Also, please justify why you experimented with only two hyperparameters.

What is a "sample of training data sets"? I thought you had only one training set (2009-15). Thanks for allowing us to clarify this. We tested RF on training data sets of 30 randomly chosen watersheds and observed that the reduction in error is negligible after 2000 trees. Then we set the number of trees to 2000 for all watersheds.

Anonymous: What do you mean by "error"? Out-of-bag MAE? Whatever the answer is, please state it in the main text.

Line 256: Why optimize for MAE, instead of one of the other scores you looked at (RMSE, R^2, or KGE)?

We actually optimized using both MAE and RMSE. The results were similar except for a few watersheds. The reason we moved forward based on MAE results was because RMSE penalizes larger errors and it was our interest to minimize the average errors, not large errors. KGE and R^2, on the other hand, do not directly capture the magnitude of errors but rather the overall performance of the model. They are also not commonly used in parameter tuning.

Anonymous: Please put this justification in the main text.

Line 261: On line 95 you said that the default is M/3, where M = number of predictors.

Yes, we have 8 predictors so round-up of 8/3 is 3.

Anonymous: Yes, in this case M/3 and ceil(sqrt(M)) are the same, but in general they are not. So my comment remains.

Anonymous: Is this shown in the figures?

Yes, this is shown in figure 5a with correlation coefficient of RF (y-axis) plotted against correlation coefficient of naïve model (x-axis). For rainfall-driven and transient watersheds, most points lie on the left of the 1-to-1 line, suggesting RF outperforms naïve model. For snowmelt-driven watersheds, the points lie on the 1-to-1 line, which indicates there is marginal difference in the models' performance.

Anonymous: Please put this explanation in the text.

Line 271: Is this shown in the figures?

Yes.

Line 274: ? I don't understand this sentence.

Sorry for the typo. The sentence should read, "Without accounting for persistence, it would be inadequate to conclude that RF delivered better performance compared to the other two groups."

Line 275: How many of these differences are statistically significant? In general, all comparisons between two models should be accompanied by a significance test.

We addressed this comment in the Major comments section.

Line 283: Could you verify this hypothesis by analyzing the data?

Yes, the hydrographs of snowmelt-driven watersheds tend to be less flashy compared to rainfall-driven watersheds.

Anonymous: Please put this in the main text, to clarify that the statement "there is less variability in flow behaviors at individual gauges in this group" is backed up by data, rather than being pure conjecture.

Line 287: How can large errors and mean errors have the same distribution? By definition, large errors are greater than mean errors.

By "distribution", we were referring to the dispersions of the RMSE and MAE score values for 3 groups in Figure 6b. For example, the MAE scores are heavily skewed towards 0 while RMSE scores are more evenly spread among snowmelt-driven watersheds.

Anonymous: Please clarify this in the main text. Without the clarification, I imagine a lot of readers will have the same question I did.

Line 293: The opposite of "poor" is not "satisfactory". Does the Rogelis paper define other ranges of KGE as "fair," "good," "excellent," etc.? Or does it just say that the 0-0.5 range is "poor"?

As we explained above, these scores should be evaluated comparatively rather than absolutely.

Line 295: ? Define.

We addressed this comment above.

Line 299: Is this a fair comparison? The sets of watersheds is your paper vs. Tongal and Booij are completely different, no?

You are right. For this very reason, we simply reported the KGE scores in our study and theirs without making a comparison of the two models.

Line 301: Figure 7 should be plotted on a performance diagram. This would allow you to show POD, FAR, frequency bias, and CSI all in the same figure. For example, see Figure 12 in this paper: https://journals.ametsoc.org/waf/article/35/4/1523/347594

We think this is a good suggestion. We included here the relative operating characteristic (ROC) plot, which measures the ability of forecast model to discriminate between events and no-events across thresholds. This is similar to the performance diagram in the paper you referenced. We will modify the discussion on POD and FAR based on this new plot.

Anonymous: The ROC curve and performance diagram are very different things. The ROC curve plots POD vs. false-alarm rate (sometimes called probability of false detection or POFD), which is b / (b + d) in the contingency table. The performance diagram plots POD vs. false-alarm *ratio*, which is b / (a + b) in the contingency table. False-alarm *rate* and *ratio* tend to be very different for rare events, because b + d (the number of actual non-events) tends to be much greater than a + b (the number of forecast events). Thus, models with very good ROC curves often have poor performance diagrams. This is why it is crucial to show both.

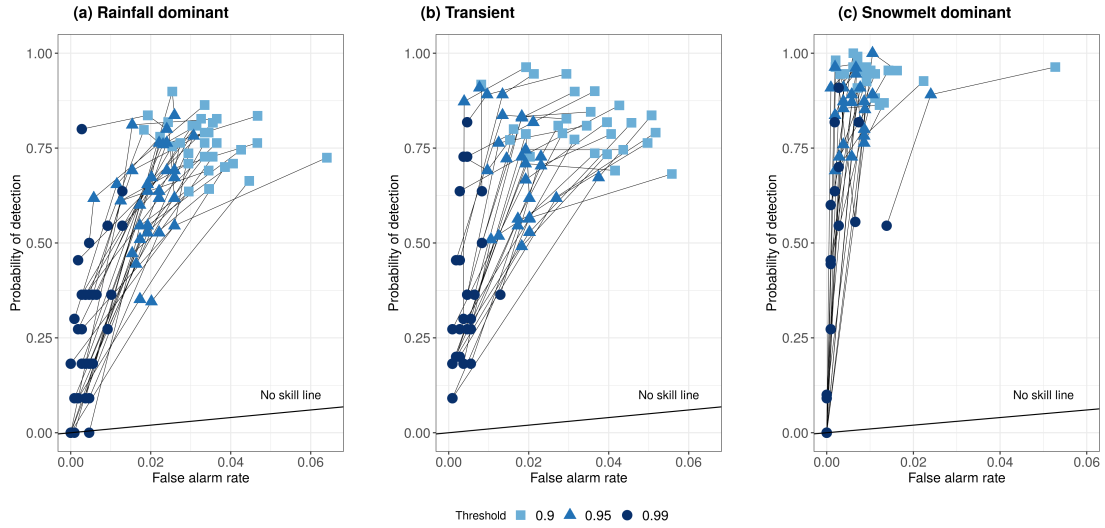


Figure 4: Probability of detection plotted is plotted against false alarm rate for three extreme thresholds: 90th, 95th, and 99th percentiles.

Line 302: What is the actual value corresponding to each percentile?

The actual values corresponding to each of the three thresholds varied across watersheds. Because of this, we did not record the actual values them and focused on the FAR and POD values.

Line 308: This hypothesis seems like just a guess. Can you verify it by looking at the data (i.e., explicitly looking at predictions for cases with large surges vs. cases without large surges)?

This is not a guess but suggested by our understanding of the hydrology of the region, examination of hydrographs, and the POD rate of RF among snowmelt-driven watersheds. The large surges of runoff from these watersheds likely occur during spring and early summer (March-June in Figure 2d).

Line 309: What does this mean? FAR and POD measure very different things, so what does it mean for them to be "in agreement"?

Although POD and FAR provide different measurements, high POD and low FAR suggest skillful forecast. This is shown on slide 25 here (`https://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/events_announce/STEWksp_Training_Hydro_Verification_30Nov06.pdf`).

Line 311: You don't know this until you have calculated frequency bias (which is shown in performance diagrams).

We're not sure how "frequency bias" is calculated as the paper cited above did not mention it. However, based on our ROC plot, if there were systematic overestimation, FAR would be exceed POD and the ROC curve would fall below the no-skill line (slide 38 in the document from NOAA we referenced above).

Anonymous: You cannot detect systematic overestimation from a ROC curve. Systematic overestimation occurs when frequency bias > 1, and frequency bias = (a + b) / (a + c), where a = number of true positives; b = number of false positives; and c = number of false negatives. In other words, frequency bias is (number of forecast events) / (number of actual events).

Frequency bias can be contoured in the background of a performance diagram, which is another reason that I've requested you include performance diagrams along with ROC curves.

References

S. Abudu, J. P. King, and A. S. Bawazir. Forecasting monthly streamflow of spring-summer runoff season in rio grande headwaters basin using stochastic hybrid modeling approach. *Journal of Hydrologic Engineering*, 16(4):384–390, 2011.

J. Albert, E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio, C. Baixeras, J. Barrio, H. Bartko, D. Bastieri, et al. Implementation of the random forest method for the imaging atmospheric cherenkov telescope magic. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 588(3):424–432, 2008.

D. G. Altman and J. M. Bland. Statistics notes variables and parameters. *Bmj*, 318(7199):1667, 1999.

S. Bharathidason and C. J. Venkataeswaran. Improving classification accuracy based on random forest model with uncorrelated high performing trees. *Int. J. Comput. Appl*, 101(13):26–30, 2014.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

J. A. Johnstone. A quasi-biennial signal in western us hydroclimate and its global teleconnections. *Climate dynamics*, 36(3-4):663–680, 2011.

N. Knowles, M. D. Dettinger, and D. R. Cayan. Trends in snowfall versus rainfall in the western united states. *Journal of Climate*, 19(18):4545–4559, 2006.

N. Knowles, M. Dettinger, and D. Cayan. 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.

A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

A. Mosavi, P. Ozturk, and K.-w. Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

T. C. Pagano, D. C. Garen, T. R. Perkins, and P. A. Pasteris. Daily updating of operational statistical seasonal water supply forecasts for the western us 1. *JAWRA Journal of the American Water Resources Association*, 45(3):767–778, 2009.

F. Pappenberger, M.-H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon. How do i know if my forecasts are better? using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, 522:697–713, 2015.

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

M. Safeeq, G. S. Mauger, G. E. Grant, I. Arismendi, A. F. Hamlet, and S.-Y. Lee. Comparing large-scale hydrological model predictions with observed streamflow in the pacific northwest: Effects of climate and groundwater. *Journal of Hydrometeorology*, 15(6):2501–2521, 2014.

E. P. Salathé Jr, A. F. Hamlet, C. F. Mass, S.-Y. Lee, M. Stumbaugh, and R. Steed. Estimates of twenty-first-century flood risk in the pacific northwest based on regional climate model simulations. *Journal of Hydrometeorology*, 15(5):1881–1899, 2014.

J. Sitterson, C. Knightes, R. Parmar, K. Wolfe, B. Avant, and M. Muche. An overview of rainfall-runoff model types. 2018.

I. M. Tohver, A. F. Hamlet, and S.-Y. Lee. Impacts of 21st-century climate change on hydrologic extremes in the pacific northwest region of north america. *JAWRA Journal of the American Water Resources Association*, 50(6):1461–1476, 2014.

H. Tyralis, G. Papacharalampous, and A. Langousis. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11(5):910, 2019.

S. J. Wenger, C. H. Luce, A. F. Hamlet, D. J. Isaak, and H. M. Neville. Macroscale hydrologic modeling of ecologically relevant flow metrics. *Water Resources Research*, 46(9), 2010.

Z. M. Yaseen, A. El-Shafie, O. Jaafar, H. A. Afan, and K. N. Sayl. Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology*, 530:829–844, 2015.

X. Zheng, Q. Wang, L. Zhou, Q. Sun, and Q. Li. Predictive contributions of snowmelt and rainfall to streamflow variations in the western united states. *Advances in Meteorology*, 2018, 2018.

Evaluation of ~~Random Forest~~ random forests for short-term daily streamflow forecasting 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~~ forecasting. Among other qualities, the popularity of ML models for such applications is ~~due to the method's competitive performance compared with alternative approaches, ease of application, and relative lack of strict distributional assumptions~~ due to their relative ease in implementation, less strict distributional assumption, and competitive computational and predictive performance. Despite the encouraging results, most applications of ML for streamflow forecasting have been limited to watersheds where rainfall is the major source of runoff. In this study, we evaluate the potential of ~~Random Forest~~ random forests (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~~ cover diverse 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, transient, and snowmelt-dominated, based on the timing of center-of-annual flow volume. RF performance is benchmarked against ~~N~~naïve and multiple linear regression (MLR) models and evaluated using four ~~metrics~~criteria: coefficient of determination, root mean squared error, mean absolute error, and Kling-Gupta efficiency (KGE). Model evaluation ~~metric~~scores suggest that RF performs better in snowmelt-driven watersheds compared to rainfall-driven watersheds. The largest improvement in forecasts, compared to benchmark models, are found among rainfall-driven watersheds. ~~We obtain 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.

Anonymous: The abstract is much-improved.

1 Introduction

Nearly all aspects of water resource management, risk assessment, and early-warning ~~water quality and~~ systems for floods 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.~~While physical models remain a common and powerful tool for predicting streamflow, ML models are gaining popularity due to some of their unique qualities and potential advantages. ~~These data-driven tools identify patterns in input-output relationship without explicit knowledge of the physical processes or formulation of mathematical equations.~~ Compared with the often labor-intensive and computationally expensive task of parameterizing in physical model (Tolson and Shoemaker, 2007; Boyle et al., 2000), ML models are data-driven and can identify patterns in the input-output relationship without explicit knowledge of the physical processes and onerous computational demand. To make up for their ~~lack~~ limited ~~of~~ ability to provide interpretation of the underlying mechanisms, ~~these models~~ ML models often require fewer calibration data than physical models, have demonstrated high accuracy in their predictive performance, are computationally efficient, and can be used in real-time forecasting (Adamowski, 2008; Mosavi et al., 2018). ML models are particularly useful when accurate prediction is the central inferential goal (Dibike and Solomatine, 2001), whereas conceptual rainfall-runoff model can provide a better understanding of hydrologic phenomena and catchment yields and responses (Sitterson et al., 2018). Artificial neural networks (ANN~~s~~), 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, ANN~~s~~, and 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 in Western USA, 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). Both watershed-scale hydrologic and statistical models have been used to assess the current and future stream hydrology and associated flood risks (Salathé Jr et al., 2014; Wenger et al., 2010; Tohver et al., 2014; Pagano et al., 2009). Safeeq et al. (2014) simulated streamflows in 217 watersheds at annual and seasonal time scales using the Variable Infiltration Capacity (VIC) model at 1/16° and 1/20° spatial resolutions. The study found that the model was able to capture the hydrologic behavior of the studied watersheds with a reasonable accuracy. Yet the authors recommend careful site-specific model calibration using not only streamflow but also snow water equivalent (SWE) data would be expected to improve model performance and reduce model bias. Pagano et al. (2009) applied Z-score regression to daily SWE from Snow Telemetry (SNOTEL) stations and year-to-date precipitation data to predict seasonal streamflow volume in unregulated streams in Western US. The authors reported the skill of these forecasts is comparable to the official published outlooks. A natural question is whether ML models can produce comparable

Anonymous: less

Anonymous: Insert colon here.

Anonymous: Do not capitalize.

Anonymous: The fractions are awkward. Please replace with "0.0625" and "0.05".

Anonymous: comma

Anonymous: comma

performance in these 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 Region (Fig. 2). The U.S. Geological Survey (2020) defines this region as hydrologic ~~unit code~~region 17 or HUC 17. 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 or report 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) to examine and compare the performance of RF in a number of watersheds across hydrologic regimes and (2) to 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 for 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: Please replace with "depending on the input variables provided". The current phrasing makes it unclear whether "the selection of input variables" is done by the user or by the random forest.

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 compared with other ML models. This is delivered through two measures of predictive contribution of variables: ~~M~~mean ~~D~~decrease in ~~A~~accuracy (MDA) and ~~M~~mean ~~D~~decrease in ~~n~~node ~~I~~impurity (MDI). These two ~~metrics~~measures 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 predictors ~~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~~ criteria. Section 3 describes the study area, datasets, and predictor selection. Results and discussion are given in Sect. 4 along with limitations and recommendation for future research. A summary and indication of future work are is provided in Sect. 5.

2 Methodology

2.1 Random ~~Forest~~forests

Proposed by Breiman (2001), RF is a ~~semi-unsupervised~~ supervised, non-parametric algorithm within the decision tree family that comprises an ensemble of ~~uncorrelated~~ decorrelated 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 forests make prediction on a new data point by having all trees run through the predictors. In the end, the forests cast a majority vote on a label class for classification task or produce a value for regression task by averaging all predictions.~~The output is the most popular class or the average of all regression values.~~ Breiman (2001) provided full details on ~~Random Forest~~RF 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:

Anonymous: Figure 1 should be referenced throughout Section 2.1, not just at the end. I think referencing Figure 1 throughout could make the explanation much more clear.

(I also made this comment in round 1.)

Anonymous: Please explain this more clearly.

(I also made this comment in round 1.)

Algorithm 1 Building a regression ~~Random Forest~~RF

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_i among X predictor variables and threshold t is selected to split the observations, y_j, into binary regions $R_1 = \{ y_j | X_i < t \}$ and $R_2 = \{ y_j | X_i \geq t \}$ 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 to 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, the predictive performance improves (or ~~reduction in OOB error~~ OOB error decreases) 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~~RF. Bernard et al. (2009) argued that 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 RF and ~~its~~ the relevant parameters.

2.2 Variable importance in random forests

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~~measures for assessing variable importance in RF: ~~M~~mean ~~D~~decrease in ~~A~~accuracy (MDA) and ~~M~~mean ~~D~~decrease in ~~n~~node ~~I~~impurity (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 potentially result in large error in the prediction (i.e., large MDA).~~It is noted~~ MDA value can be negative where a predictor has no predictive power and adds noise to the model. Strobl et al. (2007), however, expressed caution that permutation-based measures such as MDA could show a bias towards correlated predictor variables by overestimating their importance, particularly in high-dimensional data sets.

Anonymous: Please write that this method was developed by Breiman (2001).

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 the case where a predictor makes no contribution during the splitting, the relative MDI would be effectively zero. For both ~~metrics~~measures, the larger the value, the more important the predictor.

Anonymous: I don't understand the need for such a convoluted phrase. Could you replace with just "average error reduction"?

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 Performance evaluation ~~metrics~~ criteria

There exist~~s~~ different model performance ~~metrics~~ criteria and each provides unique insights on the correspondence between forecasted and observed streamflow values. While ~~Pearson correlation coefficient~~ r and its square, namely ~~C~~coefficient of determination (R^2), are often used, Legates and McCabe Jr (1999) discussed the limitation of these two measures where they were reported to be especially oversensitive to extreme values or outliers. The authors recommended that absolute error measures (i.e., ~~R~~root mean squared error or ~~M~~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~~criteria to evaluate RF performance: R^2, RMSE, MAE, and KGE. These ~~metrics~~criteria cover various aspects of model's performance and are also provide intuitive interpretation.

Anonymous: Delete.

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 N is total number of the observations during the validation period, $\hat{y_i}$ and y_i are the forecasted and observed values at day i 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 observations. It is more sensitive to larger error due to the squared operation. Both MAE and RMSE ~~values~~scores 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~~ $-\infty$ and 1. While there currently is not a definitive KGE scale, ~~negative KGE values are considered “not satisfactory" or “undesirable", and model performance is considered as “poor" with 0 < KGE < 0.5~~ Knoben et al. (2019) showed KGE values in the range between -0.41 and 1 indicate the model improves upon the mean flow benchmark, which assumes the predicted streamflow values equal to the mean of all observations. KGE value of 1 suggests the model can perfectly reproduce observations. 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:

Anonymous: Please insert "Pearson" here, since you have been calling it the "Pearson correlation coefficient" elsewhere.

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

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~~study. Here, we adopt the peak-over-threshold method ~~classifying points extreme daily discharge at 90th, 95th, and 99th percentile thresholds during the validation~~. For the validation period, we calculated the 90th, 95th, and 99th percentile streamflow values at each watershed. These are considered thresholds. If an observed daily streamflow exceeded this threshold, it would be considered an extreme event. We measure the ability of RF to capture these events using two additional ~~metrics~~criteria: probability of detection (POD) and false alarm rate (FAR). The calculation followed as in (Karran et al., 2013).

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

and

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

where ω is a specified threshold.

3 Study Area and data

3.1 Watersheds in the Pacific Northwest Hydrologic ~~Unit~~Region

In this study, we focus on watersheds in the Pacific Northwest Hydrologic Region~~ 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.The Columbia River and its tributaries make up the majority of the drainage area, traveling more than ~~1,240 miles~~ 2000 km 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.

Anonymous: Figure 2 should be referenced much earlier (at the beginning of this section and also in the introduction, where you first mention the HUC 17 region). It is difficult to understand all this description without a map.

(I also made this comment in round 1.)

The north-south running Cascade Mountain Range divides the region into eastern and western parts and strongly influence the regional climate. The western 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. Summers tend to be cool and comparatively dry. 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.~~ Proximity to the ocean creates a more moderate climate with a narrower seasonal temperature range, particularly in the winter. Spatial trends and variations in annual mean temperature, mean precipitation, drainage area, and elevation of the watersheds are shown in Fig. 3.

Anonymous: Replace with "windward (west)".

Anonymous: Replace with "leeward (east)".

Anonymous: More moderate than what?

Anonymous: Does this statement apply to the whole domain, just the west side of the Cascades, or what?

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, and 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 performed additional screening by computing correlation coefficient between the respective gauge and mean basin streamflow and removed those with a correlation of less than 0.5. We also excluded small creeks with drainage area less than 50 km^2. In total, 86 watersheds were selected (Fig. 2).

Anonymous: Delete.

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~~ 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 SNOTEL Stations in ~~the PNW~~HUC 17. 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 %~~/~~, and 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) similar to the currently reported snow observations from National Water

Anonymous: Delete.

Anonymous: comma

and Climate Center (www.wcc.nrcs.usda.gov/snow/snow_map.html) 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 (Table 2 in the Supplement). 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.

Anonymous: comma

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. A hydrological year consists of 73 pentads where each comprises of five consecutive days and observation for each day is indexed with a pentad value between 1 and 73. Data preprocessing showed moderate to strong non-linear temporal correlation between daily streamflow and the 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.

Anonymous: Replace with "at each gauge".

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~~We standardized training and validation data at each gauge using min-max scaling. The standardize data for all variables have values between 0 and 1. A flowchart representing the input-output model ~~based on~~using RF is shown in Fig. 4.

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~~. We tested RF on training data sets of 30 randomly chosen watersheds and observed that the reduction in error is negligible after 2000 trees. We then set ntree=2000 for all 86 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~~OOB ~~M~~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. 5, 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.

Anonymous: comma

Anonymous: comma

4.2 Benchmark RF against ~~Multiple Linear Regression~~ MLR and naïve models

Figure 6 shows the pair-wise comparisons of r values for RF, MLR, and naïve models. In Fig. 6a, we observe RF mostly outperforms naïve model in rainfall-driven and transient watersheds. We also discern that large improvement, defined as the positive difference in r values between RF and naïve model, tends to occur with lower persistence. 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, it would be 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 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.

Anonymous: the three

Anonymous: Delete.

4.3 Evaluation of RF overall performance

We next evaluated the overall performance of RF across three flow regimes using four ~~metrics~~criteria R^2, KGE, MAE, and RMSE (Fig. 6). We observe similar trend reported in Fig. 6 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.

Anonymous: Insert colon.

Anonymous: ? Similar to what?

Because RMSE ~~gives more weight~~ is more sensitive to larger errors compared to MAE, the difference between the two ~~metrics~~scores 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 $0.5 > KGE > 0$~~ Knoben et al. (2019) suggested KGE score greater than -0.41 indicates that a hydrologic model improves upon the mean flow benchmark and RF can be seen to give satisfactory performance at all watersheds in our study. Our results are comparable to findings in (Tongal and Booij, 2018) where authors compare the performance of RF, SVM, and ANN~~s~~ to simulate daily discharge with baseflow separation at four rivers in California and Washington. Although authors did not ~~classified~~ classify these basins, it can be inferred that 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.

Anonymous: You could verify this by computing KGE for the mean-flow benchmark on your own dataset, no?

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. Ability of RF to correctly detect extreme flows exceeding 90th, 95th, and 99th percentile thresholds (defined as the POD) for each watershed are plotted against~~false alarm rate~~the FAR in Fig.7. A threshold point falling below the no-skill line indicates the model yields higher FAR than POD an is considered having no predictive power for that threshold. 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 (higher POD, lower FAR) compared to those in transient and rainfall-driven groups. At the 95th threshold, RF can forecast correctly at least 50 percent of the times (POD > 0.5) at most watersheds. At the 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~~ diurnal 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; Cho and Jacobs, 2020).~~We see in Fig.7b, False alarm rate (FAR) is in agreement with POD~~ In our study, we observe high POD is accompanied by low FAR for the same threshold. ~~and~~ This suggests that RF is ~~consistent~~ skillful in its forecasts of ~~rare~~extreme events.~~That is, a high POD value is not~~

Anonymous: How did you compute the "no-skill line"? The no-skill line in ROC curves is typically the line where POD = FAR.

Anonymous: and is considered to have

Anonymous: extreme events

~~a result of systematic overestimation. In such cases, we would observe both high POD and high 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~~measures: DMA and MDI (Fig. 8). In both ~~metrics~~measures, 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~~measures. We also observe the sum of 3-day precipitation tends to have more predictive power than than 1-day precipitation. Maximum temperature and minimum temperature share similar contribution where minimum temperature tends 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, ~~P~~pentad ~~Index~~ comes third in both ~~metrics~~measures. This supports the long-term snowpack memory of daily streamflow (Zheng et al., 2018) and can be useful in real-time 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~~measures in ranking of the variables in snowmelt-driven group.

In transient and rainfall-dominated groups, there are noticeable disagreement between the two ~~metric~~measures. 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~~ that 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), the 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. Temperature predictors receiving higher MDA can also be due to identified bias where permutation-based importance measures overestimates the true contribution of

Anonymous: I think you mean MDA?

Anonymous: According to Figure 8, for snow-melt-dominated watersheds, pentad is either the 3rd- or 4th-most important predictor.

Anonymous: Does "in this group" mean "for snow-melt-dominated watersheds"?

Anonymous: is

Anonymous: was

Anonymous: Do you mean the PDF (probability-density function)?

Anonymous: Why would MDA, but not MDI, underestimate the importance of variables with a non-normal distribution?

Anonymous: What is a "potential" predictor variable?

Anonymous: This sentence is a non-sequitur. In the previous sentence you talked about normal vs.

correlated variables (Gregorutti et al., 2017). There is also an ongoing discussion regarding the stability of both ~~metrics~~measures across different datasets (Calle and Urrea, 2010; Nicodemus, 2011; Ishwaran and Lu, 2019). Although results from MDI make more sense in our case, we suggest RF users to exert caution when interpreting outputs from these two ~~metric~~measures.

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. These watershed characteristics were compiled as part of GAGES-II dataset using national data sources including US National Land Cover Database (NLCD) 2006 version, 100m-resolution National Elevation Dataset (NED), and Digital General Soil Map of the United States (STATSGO2) (Table 1 in the Supplement). 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 (Fig. 9b). 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. 9a) where the RF performs worse in watersheds with higher hydraulic ~~conductivities~~conductivity (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~~streamflow 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 are 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 the watersheds in the transient group where streamflow is contributed by a mixture of runoff

non-normal distributions -- the normality of a distribution has nothing to do with the scale of measurement (e.g., if you multiply a [non-]normally distributed variable by 1000, it is still [non-]normally distributed). Also, why would you mention number of categories here? Both temperature and precipitation are continuous, not categorical, variables.

Anonymous: Are you suggesting that the two temperature variables (min and max) have more correlation with other predictors than do the two precip variables (1-day and 3-day)? If so, have you verified this by computing the correlations in your dataset?

Anonymous: What do you mean by the "stability [of a measure] across different datasets"?

Anonymous: drainage

Anonymous: Replace with "land use and land cover".

from winter rainfall and springtime snowmelt~~, where~~ and the 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 State. Because we trained RF using the same input variables for all watersheds regardless of flow regimes and calculated performance ~~metrics~~criteria 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 watershed 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. Similarly, we provided RF model with a basin-average SWE from SNOTEL stations as an estimate of snowpack condition. Using a more spatially consistent SWE data such as the Snow Data Assimilation System (Pan et al., 2003) product would potentially improve model accuracy. As our results indicate that RF can produce reasonable forecasts, potential future research could explore the sensitivity of the model using satellite derived snow products ~~a station data or~~ and even include $t+1$ precipitation forecast as a predictor in the model.

Anonymous: Delete.

An inherent limitation of RF is the lack of direct uncertainty quantification in prediction. In our case, the forecasted streamflow using RF does not yield a standard error comparable to that provided by traditional ~~linear~~regression 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.

Anonymous: Methods to estimate confidence intervals

Anonymous: Replace with comma.

Anonymous: Replace with ", and"

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 forecasts at 86 watersheds in the Pacific Northwest Hydrologic Region. Key results are summarized below:

– Based on the KGE scores (ranging from 0.62 to 0.99), we show that RF is ~~able~~ capable of ~~to produce~~ producing ~~useful~~skillful forecasts across all watersheds.

– RF performs better in snowmelt-dominated watersheds, which can be attributed to stronger persistence in the streamflow time series. The 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~~measures should be coupled with understanding of the physical processes and how these processes are connected.

- Increase in 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.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resources Research, 36, 3663–3674, 2000.

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.

Cho, E. and Jacobs, J. M.: Extreme Value Snow Water Equivalent and Snowmelt for Infrastructure Design over the Contiguous United States, Water Resources Research, p. e2020WR028126, 2020.

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.

Gregorutti, B., Michel, B., and Saint-Pierre, P.: Correlation and variable importance in random forests, Statistics and Computing, 27, 659–678, 2017.

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.

Ishwaran, H. and Lu, M.: Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival, Statistics in medicine, 38, 558–582, 2019.

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.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, 2019.

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.

Pagano, T. C., Garen, D. C., Perkins, T. R., and Pasteris, P. A.: Daily updating of operational statistical seasonal water supply forecasts for the western US 1, JAWRA Journal of the American Water Resources Association, 45, 767–778, 2009.

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.

Safeeq, M., Mauger, G. S., Grant, G. E., Arismendi, I., Hamlet, A. F., and Lee, S.-Y.: Comparing large-scale hydrological model predictions with observed streamflow in the Pacific Northwest: effects of climate and groundwater, Journal of Hydrometeorology, 15, 2501–2521, 2014.

Salathé Jr, E. P., Hamlet, A. F., Mass, C. F., Lee, S.-Y., Stumbaugh, M., and Steed, R.: Estimates of twenty-first-century flood risk in the Pacific Northwest based on regional climate model simulations, Journal of Hydrometeorology, 15, 1881–1899, 2014.

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.

Sitterson, J., Knightes, C., Parmar, R., Wolfe, K., Avant, B., and Muche, M.: An overview of rainfall-runoff model types, 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.

Tohver, I. M., Hamlet, A. F., and Lee, S.-Y.: Impacts of 21st-century climate change on hydrologic extremes in the Pacific Northwest region of North America, JAWRA Journal of the American Water Resources Association, 50, 1461–1476, 2014.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resources Research, 43, 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

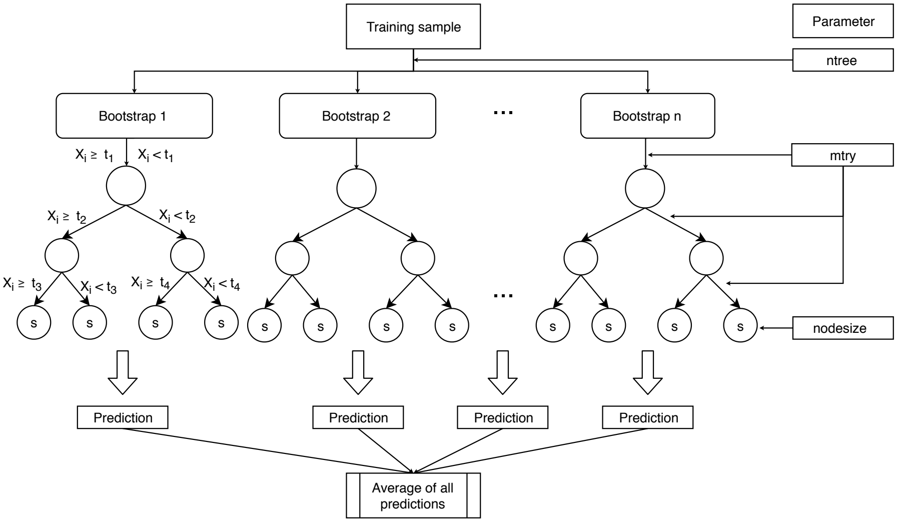
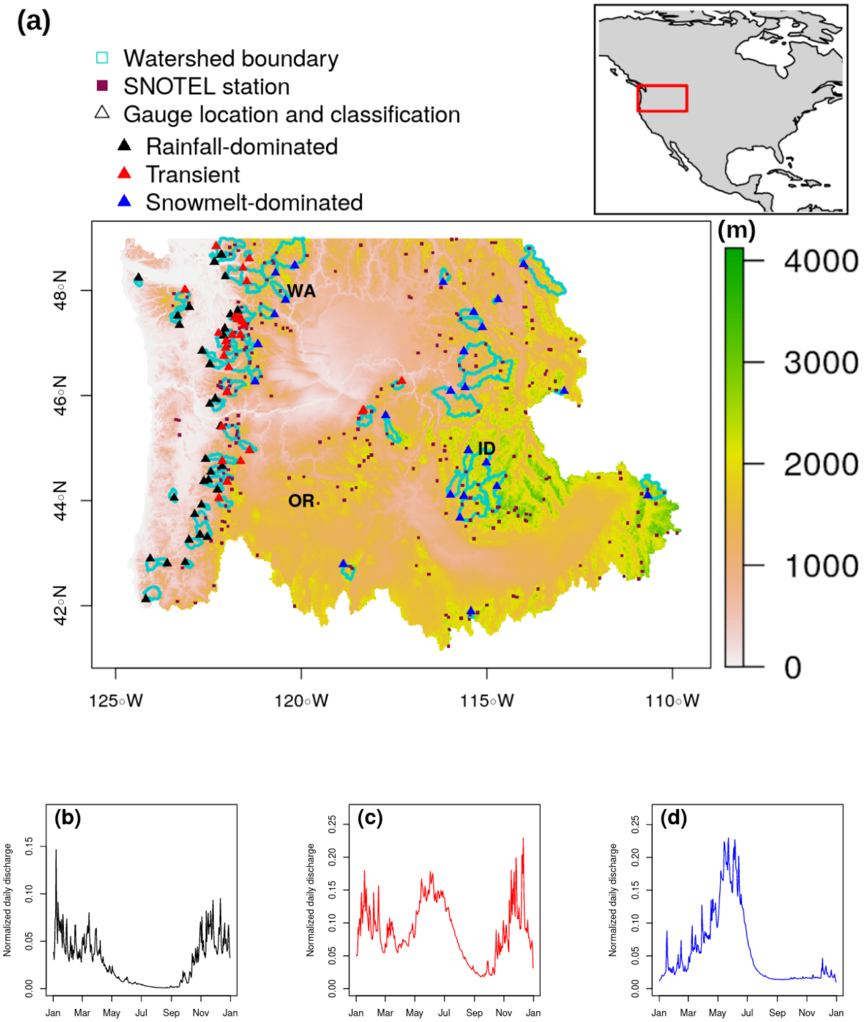
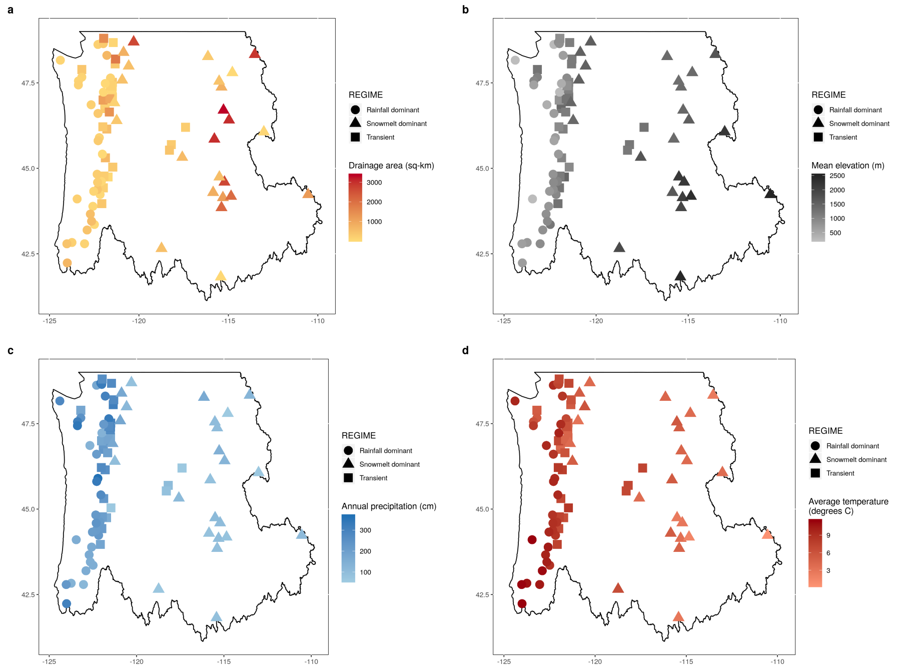


Figure 2. (a) Elevation (m) shading map showing the Pacific Northwest Hydrological Unit, 86 selected stream gauges (triangles), and their drainage area (cyan delination lines), and SNOTEL stations (brown squares). Examples of annual hydrographs of (b) rainfall-dominated, (c) transient ~~regime~~, and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



(Added plot)

Figure 3. Gauge locations with color gradient indicating variations in (a) watershed drainage area, (b)mean watershed elevation, (c) mean watershed annual precipitation, and (d) mean watershed annual temperature.



Anonymous: The font here is way too small. Please enlarge.

Figure 4. Flowchart showing the input-output model ~~based on~~using ~~Random Forest~~RF

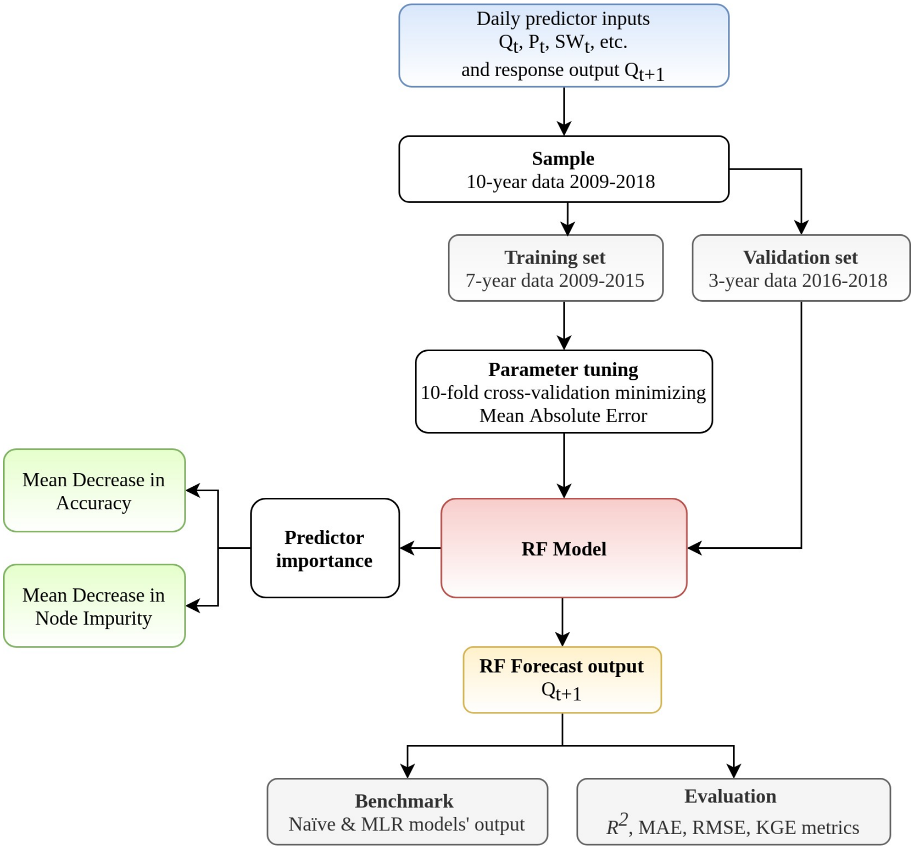

Daily predictor inputs
Q_t, P_t, SW_t, etc.
and response output Q_{t+1}
Sample
10-year data 2009-2018
Training set
7-year data 2009-2015
Validation set
3-year data 2016-2018
Parameter tuning
10-fold cross-validation minimizing Mean Absolute Error
Mean Decrease in Accuracy
Mean Decrease in Node Impurity
Predictor importance
RF Model
RF Forecast output
Q_{t+1}
Benchmark
Naïve & MLR models' output
Evaluation
R^2, MAE, RMSE, KGE metrics

Figure 5. Out-of-bag mean absolute error plotted against `mtry` during optimal parameter search at Carbon River Watershed (USGS site 12094000).

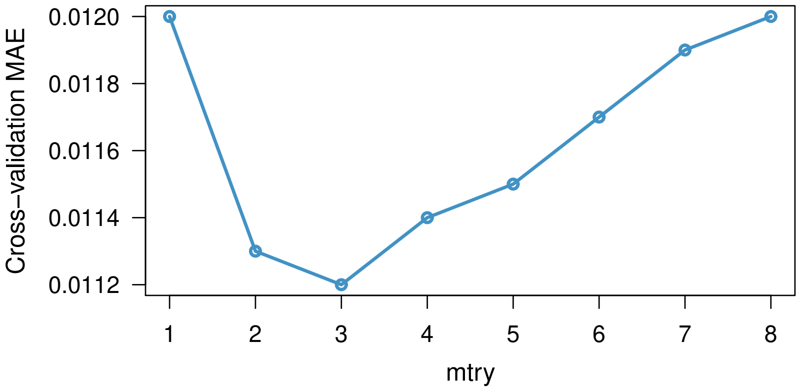
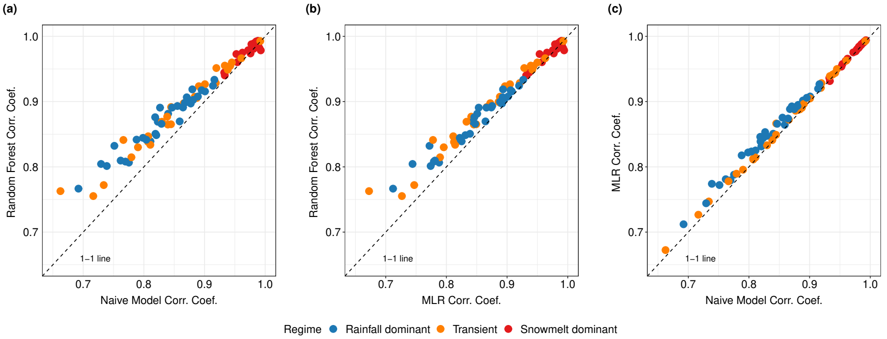
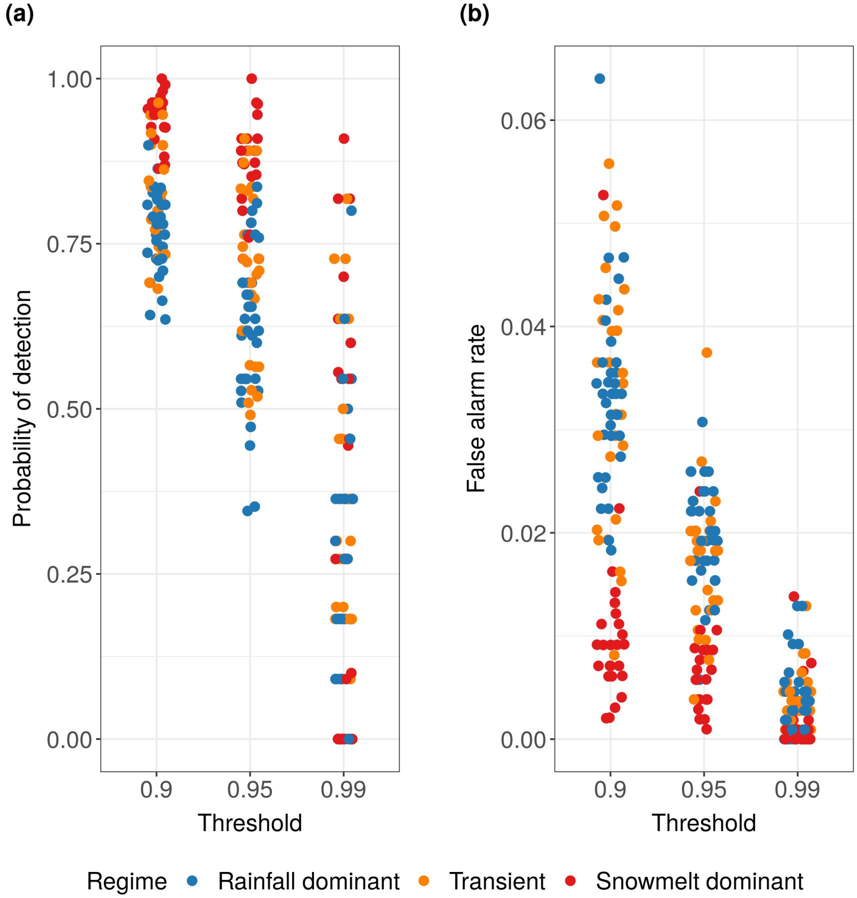


Figure 6. Pairwise scatter plots of Pearson correlation coefficient between forecasted and observed values for (a) ~~Random Forest~~RF vs. naïve model, (b) ~~Random ForestRandom Forest~~RF vs ~~Multiple Linear Regression~~MLR, and (c) ~~Multiple Linear Regression~~MLR vs. naïve model. Each dot represents a watershed (n=86).



(Removed plot)

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.~~



605 (Added plot)

Figure 7. The probability of detection (POD) plotted against the false alarm rate (FAR) for three extreme thresholds: 90th, 95th, and 99th percentiles. Thin black line connects values from the same watershed. (Vertical axis) Number of times RF *correctly* forecasted events that exceeded the threshold divided by the total number of exceedance. (Horizontal axis) Number of times RF *incorrectly* forecasted events that exceeded the threshold divided by the total number of non-exceedance.

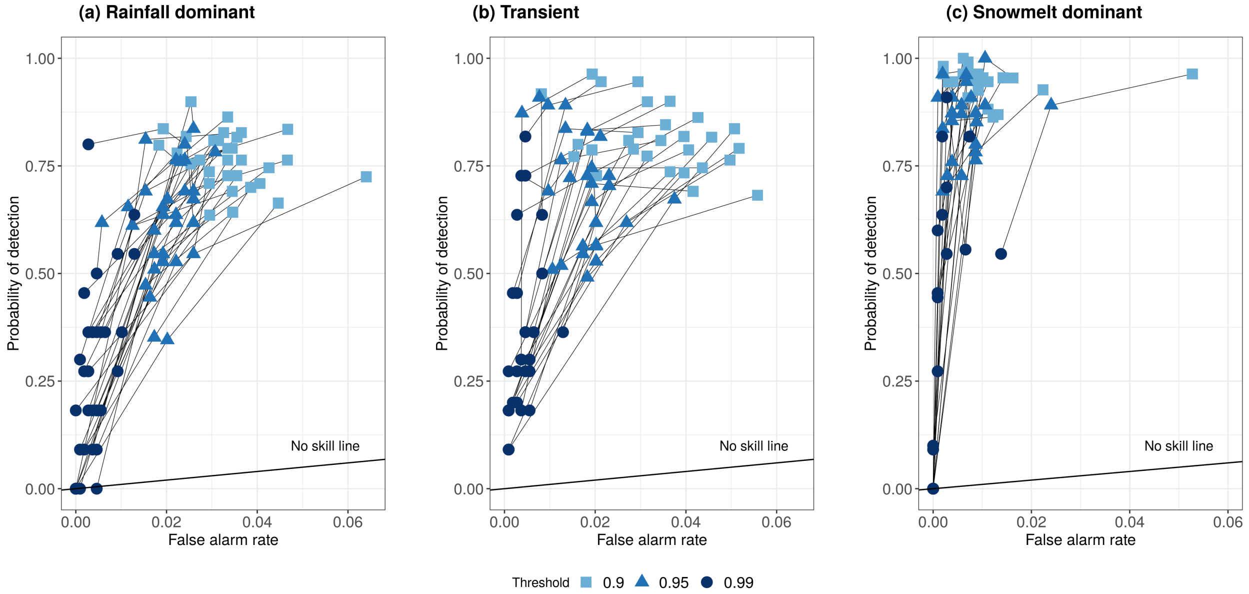
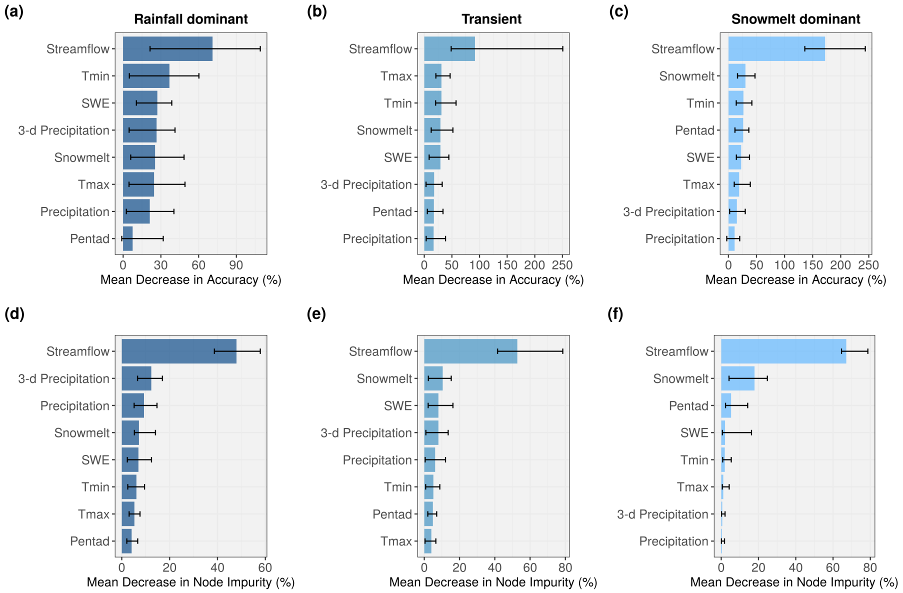


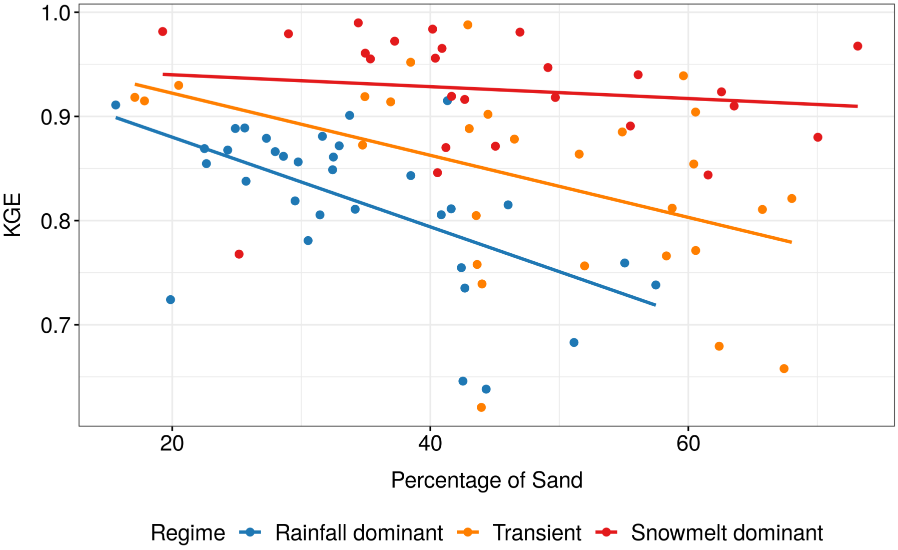
Figure 8. Barplots show importance of predictor variables using ~~(a-c) DMA and (d-f) DMI metrics~~(a-c) MDA and (d-f) MDI criteria. Length of the blue bars indicates the median value across the watersheds for each flow regime and the thin black bar represents the range of the values.

Anonymous: The full range (from min to max)?



Removed plot

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



Added plot

Figure 9. KGE scores plotted against (a) the average percent of slope and (b) the average percent of sand in soil at each watershed. Best-fit lines were determined using simple linear regression. Pearson correlation coefficients were computed with associated significance.

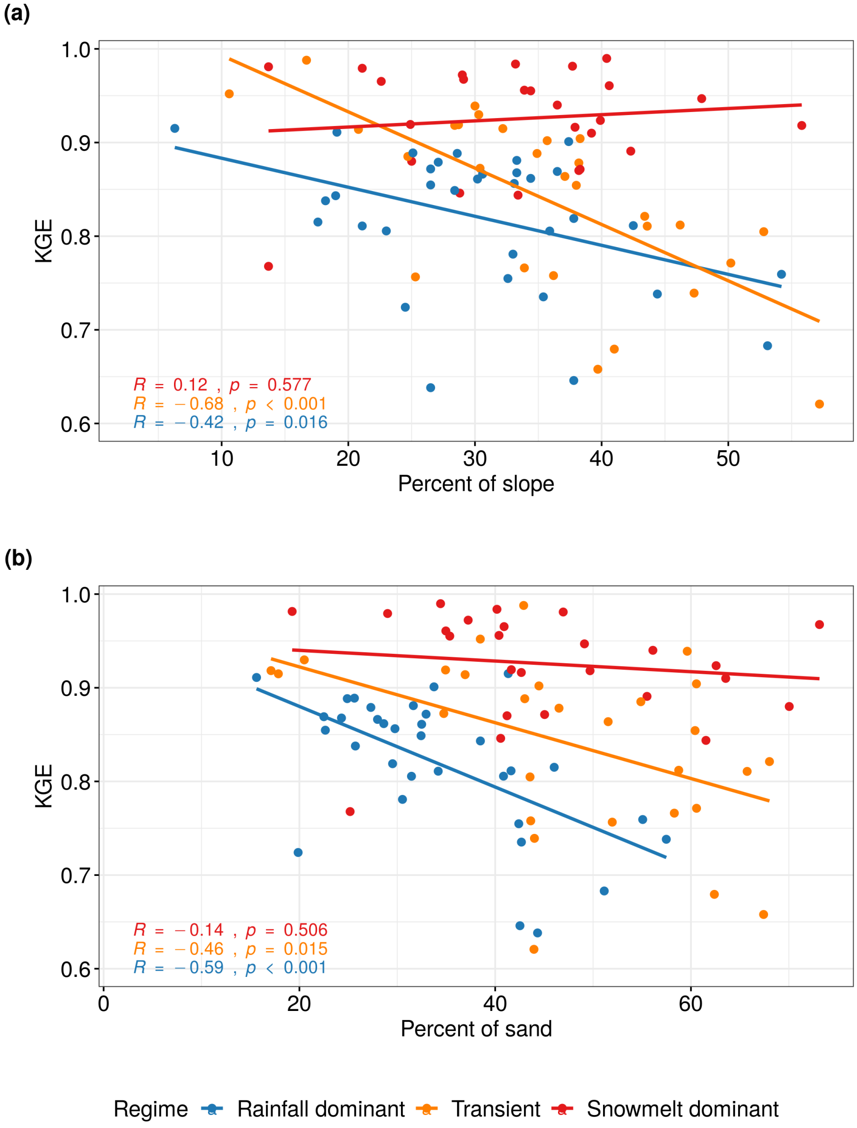


Table 1. Number of ~~streamflow~~ USGS gauges used in the study for each flow regime, ~~ranges of~~ mean watershed elevation, drainage area, annual precipitation, and annual mean temperature ranges.

Hydrologic regime	Number of gauges	Mean watershed elevation (m)	Drainage area (km^2)	Mean annual precipitation (cm)	Mean annual temperature (°C)
Rainfall-dominated	33	239 - 1207	58 - 703	122.0 - 367.0	5.4 - 11.5
~~Transitional~~ Transient	28	813 - 1477	58 - 1855	63.2 - 314.0	4.16 - 8.42
Snowmelt-dominated	25	1349 - 2509	51 - 3355	58.0 - 177.0	0.4 - 6.62

Table 2. List of potential predictors.

Anonymous: Delete. You use all these variables as predictors, so they are not just *potential* predictors.

No.	Predictors	Index	Unit	Source
1	Streamflow at day t	Q_t	~~ft^3/s~~ m^3 s^{-1}	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~~ mm	SNOTEL
5	Maximum temperature	$TMAX_t$	~~degree F~~ °C	SNOTEL
6	Minimum temperature	$TMIN_t$	~~degree F~~ °C	SNOTEL
7	Snowmelt (SW_t - SW_{t-1})	SD_t	~~in~~ mm	Derived from SNOTEL
8	Pentad ~~index~~	PEN_t	-	-

Table 3. The ~~achieved~~optimized 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

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

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

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

The following materials are submitted with the manuscript, *"Evaluation of random forests for short-term daily streamflow forecasting in rainfall and snowmelt-driven watersheds."*

1. Table 1. USGS Gaging stations used in the study, classified regime, and selected physical characteristics that were used to compute Pearson correlation coefficient presented in Table 5 in the manuscript.

2. Figure 1. Delineation of basins (USGS HUC-6) within the Pacific Northwest Hydrologic Region.

3. Table 2. HUC-6 basins, drainage area, number of available SNOTEL stations, and elevation range of the SNOTEL stations.

Table 1: USGS Gaging stations used in the study, classified regime, and selected physical characteristics.

Station ID	Regime	Drainage area ($km2$)	Compactness	Mean elevation (m)	% slope	Aspect eastness	Stream density	% sand in soil	% forested area	% impervious cover
10396000	Snowmelt dominant	528.9	1.92	1890.5	13.7	-0.78	0.81	25.18	14.31	0.12
12043300	Rainfall dominant	135.2	1.97	238.5	24.5	-0.98	0.78	19.87	58.83	0.1
12048000	Transient	405	2.22	1265.7	46.2	0.4	0.56	58.75	71.13	0.04
12054000	Rainfall dominant	171.7	1.69	1073.8	54.2	0.77	0.53	55.07	74.9	0.09
12056500	Rainfall dominant	147	1.93	990.6	53.1	-0.36	0.54	51.14	82.06	0.06
12060500	Rainfall dominant	198.3	1.46	601.9	42.5	0.21	0.74	41.62	78.88	0.63
12079000	Rainfall dominant	224.1	1.31	437.9	21.1	0.2	0.75	34.18	59.36	1.27
12082500	Transient	350	2.03	1182.5	33.9	-0.86	0.8	58.3	68.05	0.43
12092000	Transient	240.9	2.84	1420.3	39.7	-0.88	0.67	67.42	59.01	0.39
12094000	Transient	205.2	1.77	1263.9	41	-0.28	0.77	62.39	69.29	0.17
12095000	Rainfall dominant	205.8	2.8	680.3	23	-0.66	0.74	40.85	65.38	1.16
12096500	Transient	1142.8	2.34	812.6	25.3	-0.65	0.67	51.96	57.65	2.5
12097500	Transient	190.2	2.07	1214.1	35.7	-0.64	0.69	44.47	79.78	0.53
12097850	Transient	970.4	2.51	1262.7	38.2	0.21	0.75	46.52	72.32	0.51
12108500	Rainfall dominant	71.1	1.59	262.2	6.3	-0.86	0.52	41.33	28.96	6.79
12114500	Transient	66.6	2.93	1073.5	37.1	-0.98	0.81	51.54	77.27	0.75
12141300	Transient	401.5	2.05	1038.5	50.2	-0.99	0.7	60.57	76.88	0.06
12142000	Rainfall dominant	165.6	2.76	931.7	44.4	-0.99	0.67	57.48	77.76	0.61
12143400	Transient	107.9	1.9	1042.5	43.4	-0.92	0.64	68.02	69.04	1.52
12143600	Transient	165	1.47	966.3	43.6	-0.91	0.7	65.74	70.34	1.56
12144000	Transient	210.1	1.19	827.9	38	-0.48	0.68	60.42	70.66	2.33
12145500	Rainfall dominant	79	2.13	460.9	19	-0.57	0.87	38.49	73.84	1.23
12167000	Rainfall dominant	683.8	1.64	655.4	32.6	-0.89	0.59	42.41	78.89	0.54
12179900	Transient	128.3	2.38	1110.1	57.2	0.02	0.54	43.96	59.73	0.11
12186000	Transient	398.4	1.92	1176	52.8	-0.63	0.63	43.56	76.4	0.13
12189500	Transient	1855.3	1.88	1150.9	47.3	-0.78	0.58	44.01	74.44	0.25
12201500	Rainfall dominant	224.2	1.42	274.9	17.6	-0.92	0.73	46.03	65.32	1.82
12209490	Rainfall dominant	58.1	1.92	932.9	37.8	-0.95	0.62	42.53	66.85	0.62
12210000	Rainfall dominant	332.1	2.37	838.1	35.4	-0.94	0.64	42.68	71.42	0.35
12210700	Transient	1524.8	1.81	873.5	36.2	-0.91	0.69	43.63	69.46	0.5
12323670	Snowmelt dominant	102.4	2.27	2222.5	29	0.95	0.72	37.24	46.62	0.16
12354000	Snowmelt dominant	828.4	2.14	1383.4	34.4	0.45	0.68	35.36	85.62	0.38
12358500	Snowmelt dominant	2939.2	1.12	1723.7	40.4	-0.99	0.72	34.42	69.91	0.07
12374250	Snowmelt dominant	50.8	2.56	1408.3	24.9	0.98	0.84	41.65	94.87	0.14
12390700	Snowmelt dominant	470.2	2.23	1348.9	37.9	0.89	0.52	42.67	91.73	0.1
12392155	Snowmelt dominant	326.2	2.33	1389.8	38.3	-0.42	0.61	45.05	87.93	0.09

Table 1: USGS Gaging stations used in the study, classified regime, and selected physical characteristics.

Station ID	Regime	Drainage area ($km2$)	Compactness	Mean elevation (m)	% slope	Aspect eastness	Stream density	% sand in soil	% forested area	% impervious cover
12448500	Snowmelt dominant	2683.6	1.83	1591.5	38.2	0.1	0.79	41.21	51.58	0.42
12451000	Snowmelt dominant	830.6	2.12	1534	55.8	-0.61	0.67	49.69	48.34	0.1
12452800	Snowmelt dominant	526.4	1.7	1519.8	42.3	-0.36	0.7	55.51	64.63	0.65
12458000	Snowmelt dominant	499.4	2.59	1547.5	47.9	0.9	0.67	49.12	61.75	0.09
12488500	Snowmelt dominant	205.2	1.77	1465.8	39.2	0.97	0.77	63.55	81.66	0.27
13010065	Snowmelt dominant	1222.3	1.11	2508.5	13.7	-0.8	0.49	46.95	41.54	0.01
13162225	Snowmelt dominant	76.1	2.65	2474.2	40.6	-0.59	0.69	34.96	47.25	0.25
13185000	Snowmelt dominant	2154.4	1.9	1955.1	36.5	-1	0.79	56.11	43.97	0.11
13235000	Snowmelt dominant	1163.2	1.39	2079	39.9	-0.97	0.75	62.57	52.91	0.12
13237920	Snowmelt dominant	874.8	1.39	1658.2	29.1	-0.41	0.83	73.13	77.43	0.11
13296500	Snowmelt dominant	2090.9	1.17	2375.2	28.8	0.44	0.87	40.56	61.14	0.11
13309220	Snowmelt dominant	2696.6	1.55	2192.4	33.4	1	0.83	61.53	59.36	0.04
13313000	Snowmelt dominant	561.9	2.37	2180.4	25	-0.54	0.76	70.04	79.39	0.07
13331500	Snowmelt dominant	618.9	1.15	1736.3	37.7	-0.39	0.63	19.26	76.92	0.04
13334450	Transient	269.8	2.62	1271.8	30.3	0.89	0.82	20.5	51.92	0.21
13337000	Snowmelt dominant	3053.4	1.08	1584.1	33.2	-0.9	0.72	40.18	86.37	0.07
13338500	Snowmelt dominant	3027	1.28	1384.6	21.1	0.57	0.88	29.01	70.53	0.17
13340600	Snowmelt dominant	3354.6	2.18	1442.7	33.9	-1	0.72	40.39	75.57	0.05
14020000	Transient	341.4	2.22	1209.3	32.2	-0.79	0.65	17.85	73.37	0.31
14020300	Transient	456.4	1.53	1187.3	28.4	-0.87	0.79	17.1	67.74	0.42
14092750	Transient	57.5	2.64	1477.3	24.7	1	0.84	54.88	86.46	0.02
14096850	Transient	374.6	1.76	942.7	10.6	0.53	0.75	38.48	57.42	0.19
14107000	Snowmelt dominant	393.8	2.19	1428.8	22.6	-0.05	0.84	40.91	74.79	1.29
14137000	Transient	674.2	2.54	1006.4	30.4	-0.98	0.86	34.75	83.69	0.19
14141500	Rainfall dominant	59.9	1.45	724.7	18.2	-0.96	0.65	25.72	84.58	0.01
14150800	Rainfall dominant	113.3	2.26	765.5	26.5	-1	0.62	22.65	86.26	0.09
14154500	Rainfall dominant	546.8	2.41	857.5	33.3	-0.5	0.68	24.3	84.93	0.04
14158500	Transient	237.1	1.82	1253.9	16.7	-0.94	0.46	42.91	80.99	0.15
14159200	Transient	414.3	1.97	1280.8	28.7	-0.66	0.7	34.93	92.32	0.02
14161500	Rainfall dominant	62.4	2.7	982.7	33.3	-0.97	0.55	31.64	93.6	0.02
14166500	Rainfall dominant	226.5	1.73	253.8	19.1	0.18	0.58	15.62	57.02	0.94
14179000	Transient	272.5	1.73	1149.5	34.9	-0.88	0.85	43.02	85.41	0.05
14180300	Rainfall dominant	66.6	2.72	1027.9	26.5	-0.6	0.62	32.94	87.63	0.08
14182500	Rainfall dominant	286.8	1.73	818.3	35.9	-0.62	0.68	31.45	84.37	0.06
14185000	Rainfall dominant	458.2	2.4	889.5	30.6	-0.89	0.69	27.98	83.15	0.08
14185900	Rainfall dominant	258.2	2.11	918.7	37.8	-0.91	0.68	29.52	82.16	0.05

Table 1: USGS Gaging stations used in the study, classified regime, and selected physical characteristics.

Station ID	Regime	Drainage area ($km2$)	Compactness	Mean elevation (m)	% slope	Aspect eastness	Stream density	% sand in soil	% forested area	% impervious cover
14187000	Rainfall dominant	134.7	2.56	732.7	28.6	-0.09	0.63	24.89	71.81	0.02
14216000	Transient	594.6	1.84	1080.2	20.8	-0.91	0.75	36.93	83.55	0.97
14216500	Transient	349.5	2.33	921	30	-0.09	0.82	59.62	60.01	1.07
14219000	Rainfall dominant	167.4	1.79	687	30.2	-0.5	0.51	32.49	67.64	0.67
14222500	Rainfall dominant	323.9	1.85	573.3	25.1	-0.97	0.68	25.62	69.44	0.43
14231000	Transient	1378	1.94	1128.6	38.3	-0.99	0.82	60.56	76.23	0.5
14236200	Rainfall dominant	361	1.79	672.9	33.1	-0.84	0.58	29.76	61.65	1.02
14308990	Rainfall dominant	167.8	2.59	917.6	26.5	-0.36	0.75	44.33	84.29	0.04
14309500	Rainfall dominant	224.9	2.4	736.3	33	0.97	0.66	30.53	82.45	0.12
14316495	Rainfall dominant	79	3.22	1207.1	37.4	-0.5	0.75	33.75	83.85	0.03
14316700	Rainfall dominant	587.9	2.76	944	34.4	-0.36	0.69	28.62	87.24	0.04
14318000	Rainfall dominant	459.5	2.23	858.9	27.1	-0.26	0.71	27.3	85.85	0.04
14325000	Rainfall dominant	443.1	2.27	651.9	28.4	-0.31	0.65	32.43	75.87	0.31
14400000	Rainfall dominant	702.6	2.39	671.2	36.5	-0.88	0.61	22.5	57.15	0.1

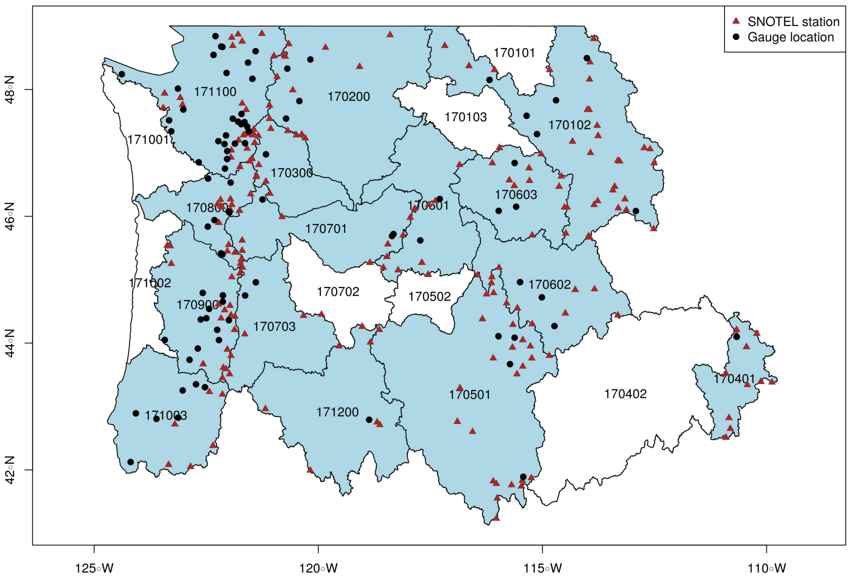


Figure 1: Delineation of basins (USGS HUC-6) within the Pacific Northwest Hydrologic Region. Blue basins contain at least one of the 86 chosen watershed included in the study.

Table 2: Basin names, drainage area, number of available SNOTEL stations, and elevation range of the SNOTEL stations.

HUC-6	Basin name	Area (km^2)	Number of SNOTEL	Elevation range (m)
170102	Pend Oreille	67598.70	30	4350-8250
170200	Upper Columbia	119755.57	10	3590-6490
170300	Yakima	15928.20	9	3430-5920
170401	Snake Headwaters	14812.20	11	6770-9820
170501	Middle Snake-Boise	85150.16	27	4800-8360
170601	Lower Snake	30198.02	7	4000-5760
170602	Salmon	36248.15	11	5350-9150
170603	Clearwater	24318.13	9	4600-6320
170701	Middle Columbia	29124.57	11	3310-5580
170703	Deschutes	27789.56	8	3810-5850
170800	Lower Columbia	16120.04	15	2140-5800
170900	Willamette	29697.66	15	2420-4950
171003	Southern Oregon Coastal	34510.01	6	3240-6050
171100	Puget Sound	52958.23	26	2250-5130
171200	Oregon Closed Basins	45143.34	6	5250-7660