We thank the reviewer 2 for his/her valuable comments and suggestions. The manuscript has benefited immensely from the review process. We appreciate the opportunity to provide further clarifications and add additional changes to improve the manuscript. We included below the reviewer's comments (italic black font) and our responses (plain blue font). Changes in the manuscript are highlighted in red.

1 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).

We acknowledge the reviewer's point of view in requesting statistical significance tests for model comparison. We performed two-sample Wilcoxon rank-sum tests to compare the correlation coefficients obtained from the three models : RF, naïve, and MLR for each flow regime. Their distributions are also plotted below with p-value included. We also modified the text to reflect this change, "Figure 1 shows the distributions of Pearson correlation coefficient (r) between forecasted and observed values obtained from the three models: RF, naïve, and MLR. Non-parametric, two-sample Wilcoxon rank-sum significance tests [Wilcoxon et al., 1970], which is used to assess whether the values obtained between two separate groups on are systematically different from one another, suggest that the pair-wise differences in r values between RF and the other two models are statistically significant ($p < 0.05$) in two flow regimes. RF is observed to outperform both naïve and MLR models in rainfall-driven and transient watersheds. Among snowmelt-driven watersheds, the three models yield similar correlation coefficients ($p > 0.05$). In Fig. 2a and 2b, we observe most points lie on the left of the 1-to-1 line, suggesting that RF outperforms naïve model at individual watersheds in rainfall-driven and transient regimes. 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, the data points lie on the 1-to-1 line, indicating that the three models show marginal difference in r values." We also performed the two-sample t-test, which assumes the data has a normal distribution, but decided to go with the Wilcoxon test due to its non-parametric nature. The two tests, however, yield very similar results where the Wilcoxon test is even more conservative (giving larger p-values).

Figure 1: Boxplots for Pearson correlation coefficient between forecasted and observed values for three models: RF, naïve, and MLR across three flow regimes. Pair-wise Wilcoxon rank-sum significance tests are performed and p-value (in black) are included for each pair of models.

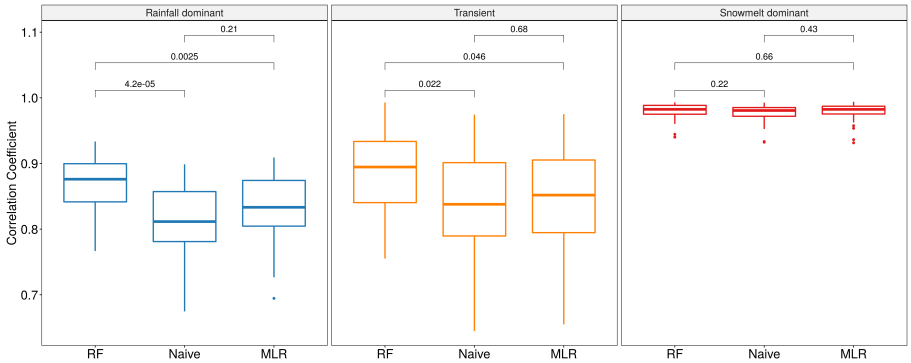
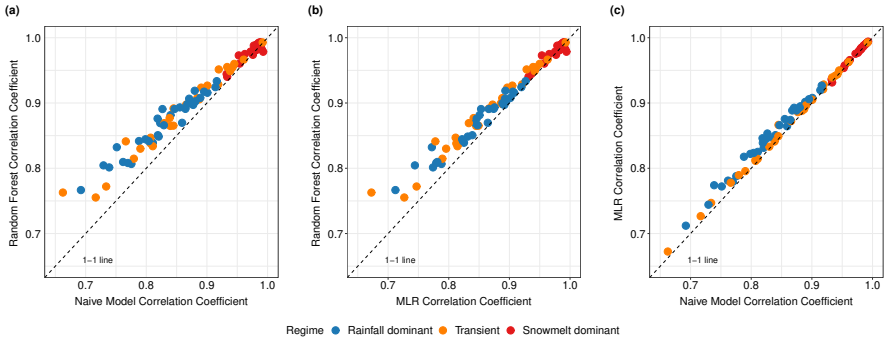


Figure 2: Pairwise scatter plots of Pearson correlation coefficient between forecasted and observed values for (a) RF vs. naïve model, (b) RF vs MLR, and (c) MLR vs. naïve model. Each dot represents a watershed (n=86).



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.

We agree that saying one ML model is more interpretable than the others may be controversial and made the following modification to the manuscript, "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: mean decrease in accuracy (MDA) and mean decrease in node impuritiy (MDI). These two 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]. The interpretability of a ML model, however, can be a controversial subject and remains an active area of study [Ribeiro et al., 2016, Carvalho et al., 2019]. Both model-agnostic (e.g., permutation-based feature importance [Breiman, 2001]) and model-specific (e.g., gini-based for RF [Breiman et al., 1984], gradient-based for ANNs [Shrikumar et al., 2017]) interpretation methods can provide useful insights into how the ML models make their predictions. ~~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 and streamflow response.

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.

You are right that only training data should be used to compute the min-max values for variable scaling. This was our approach. First, we computed the min and max values from training data sets for each of the predictor and response variables at each watershed. These min and max values were then used to standardize both training and validation data sets. The training data, which were used to compute min-max values for standardization, therefore have values between 0 and 1. We added this clarification to the text.

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

We added the following justification to the manuscript, "While all considered parameters might have an effect on the performance of RF, we chose to focus on two parameters: `ntree` and `mtry` for a number of reasons. First, these two parameters were originally introduced by [Breiman, 2001] in

the development of RF algorithm. Second, `ntree` in a forest is a parameter that is generally not tunable but should be set sufficiently high [Oshiro et al., 2012, Probst et al., 2019] for RF to achieve optimal performance. Furthermore, empirical results provided in previous works suggest that `mtry` is the most influential out of parameters in RF [Bernard et al., 2009, Van Rijn and Hutter, 2018, Probst et al., 2019].

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.

We acknowledge that ROC curve may not be used to diagnose and detect systematic overestimation. In the revised manuscript submitted after round 1, we removed this part of the discussion (please see lines 354-356) for the following reasons. First, as you suggested, detecting systematic overestimation requires plotting the performance diagram, which is not commonly used to evaluate hydrologic models, both data-driven and physically-based, based on our review of the literature (we further discussed why this may be the case below). Second, it also requires introducing both false alarm ratio and false alarm rate and their differences, which you clearly explained and we appreciate it. However, this may prove confusing to the readers considering it's not a major focus of the paper.

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.

We apologize for the confusion. The "no-skill line" in the ROC curves is not misplaced. You are right that the no-skill line should be the x = y line (or POD = false-alarm rate FAR) and this was the case for our figure. In Figure 7 of our manuscript, we showed a partial ROC curve (please note the scales of the x-axis and y-axis are not 1-to-1 here) with FAR < 0.065, which is slightly larger than the maximum FAR value for the three threshold across all watersheds. When we set xlim = ylim, the data points are clumped together. We provide the two plots below to illustrate where we believe the bottom plot, which is included in the manuscript, presents the data better.

Figure 3: The probability of detection (POD) plotted against the false alarm rate (FAR) for three extreme thresholds: 90^{th}, 95^{th}, and 99^{th} 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.

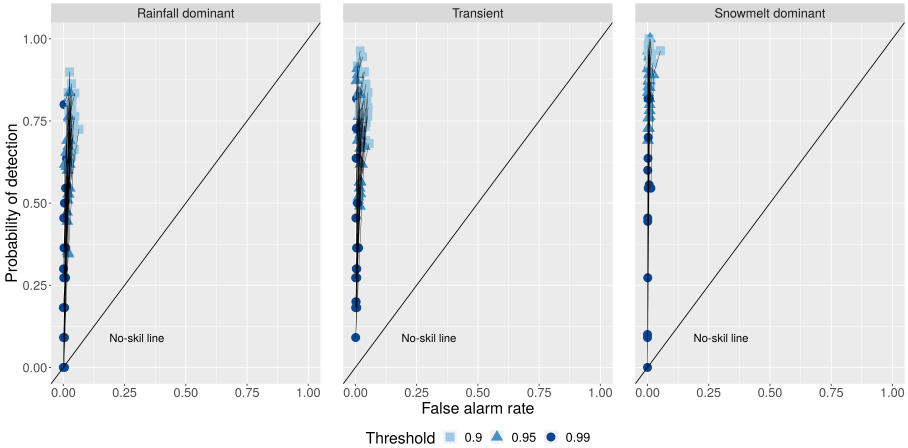
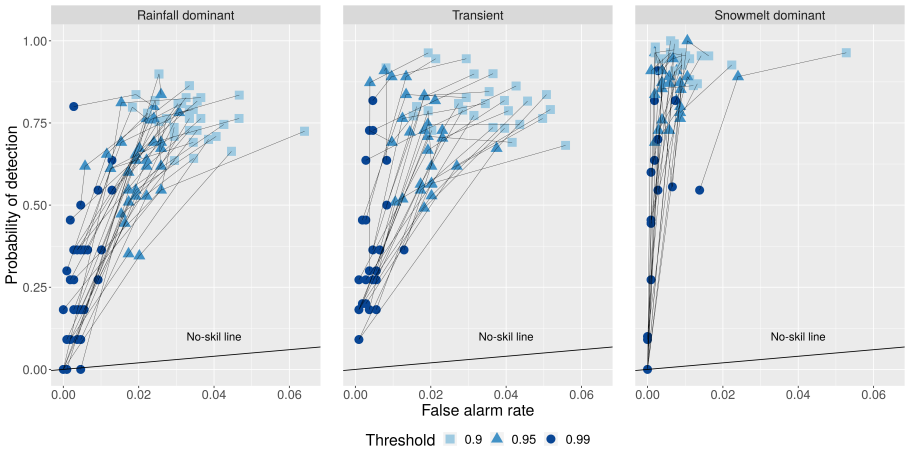


Figure 4: The probability of detection (POD) plotted against the false alarm rate (FAR) for three extreme thresholds: 90^{th}, 95^{th}, and 99^{th} percentiles. Thin black line connects values from the same watershed. (Vertical axis) Number of times RF *correctly* forecasted events that exceeded the 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.



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 1which 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/machinelearning/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.

We intend to use this partial ROC curve to facilitate our discussion of section 4.4, which focuses on the performance of RF on detecting extreme streamflows. We also justified our selection of the three thresholds on lines 185-192. As you mentioned that we could technically compute thresholds in the increments of 10%, plot the full ROC curve, compute the area under curve, and evaluate the skillfulness of RF using this quantity. However, we believe that it is more appropriate to limit the discussion of POD, FAR, and the ROC curve only to extreme events as this is related to flood forecasting (where floods are the results of extreme streamflow exceeding specific thresholds and clearly defined as binary flood/no-flood events). The thresholds we chose are often used to study flood characteristics and have physical interpretations. But this also depends on the watershed and the evaluation of historical floods. Flood forecasting, however, is not the focus of our paper. Our study and the ones you cited below and in round 1 (both focus on storm detection) have very different aims. The main outcome in these studies is binary and the events are clearly defined (e.g., cyclones). Therefore, the use of performance diagram and ROC provides valuable insights. In our study, the outcome is a continuous variable and it does not make sense to evaluate the performance of the model using these tools. For example, if the observed streamflow is 30 m^3/s, which also happens to be the 10^{th} percentile streamflow at a watershed, and our prediction is 32 m^3/s. The prediction, in this case, can neither be considered a "hit" or a "miss". Therefore, correlation coefficient r, R^2, KGE, MAE, and RMSE are better criteria and were used to evaluate the overall performance of the RF model. To make it more clear, we added the following text to manuscript, "In hydrological forecast, one might be interested in the ability of the model to capture more extreme events rather than the overall performance. This is particularly relevant in flood risk assessment and flood forecasting where floods are associated with discharge exceeding a high percentile (typically $\geq 90^{th}$)[Cayan et al., 1999]."

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=fulltextdisplay. 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.

We provided our reasons for not including full ROC curve and performance diagram above.

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

We addressed these comments below.

10. Mean-flow benchmark. On lines 331-332, the authors cite a previous paper to claim that a Kling-Gupta efficiency (KGE) > -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.

To clarify, the author derived and concluded that a Kling-Gupta efficiency (KGE) > -0.41 improves upon the mean-flow benchmark, which is the KGE achieved by always predicting the time-mean flow ("climatology") at *any* basin. We added the following modification to the text to make it less ambiguous, "Knoben et al. (2019) suggested that, at any given basin, a KGE score greater than -0.41 indicates a hydrologic model improves upon the mean flow benchmark. Therefore, RF can be seen to give satisfactory performance at all watersheds in our study based on the KGE scores.

2 Reviewer's response to our response after round 1

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)".

We have added your suggestion to the manuscript, " Artificial neural networks (ANN), neuro-fuzzy (a combination of ANNs and fuzzy logic), 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]."

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.

We made the following modification to the text, "They found BNN outperformed multiple linear regression (MLR) as well as ~~two other competing ML models~~the other two ML models.

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.

We defined what baseflow saeration as you suggested, "Tongal and Booij [2018] forecasted daily streamflow in four rivers in the United States with SVR, ANN, and RF coupled with a baseflow separation (i.e., separating the two different components of streamflow into baseflow and surface flow) method.

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 added the following explanation to the manuscript, "In [Rasouli et al., 2012], the authors forecasted streamflow at 1-7 day lead times using three ML models and data from combinations of climate indices and local meteo-hydrologic observations. The authors concluded that models with local observations as predictors were generally best at shorter lead times while models with 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. In our study, we focused on 1-day lead time forecasting and therefore did not include long-term climate information. ~~We focus on 1-day lead time because we assume only antecedent information for predictors are available at the time forecast is made.~~ "

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. 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. 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.

We addressed these comments in major comment 2.

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

We added the following clarification to the text, "The remainder of the paper is arranged as follows. Section 2 provides a brief introduction to RF, relevant parameters (which can also be referred as "hyper-parameters" in the ML literature), and selected evaluation criteria. Section 3 describes the study area, datasets, and predictor selection. Resultsand 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."

Please clarify this in the main text.

We added the clarification to the manuscript, "Non-parametric methods do not assume any particular family for the distribution of the data[Altman and Bland, 1999]."

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?

We used Caret to loop through the `mtry` parameter. This was stated on lines 117-118. "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]".

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

We have added your suggestion to the text.

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.)

We have added the following clarification to the text ,"MAE and RMSE values range between 0 and ∞ where a score of 0 indicates a perfect match between predicted and observed data".

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

We replaced "as predictors" with "as predictors for streamflow".

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

We explicitly stated and discussed this limitation under 4.7 Limitations and future research on lines 430-435.

Please put this in the main text.

This was added."Given that there is high temporal correlation in daily temperatures, TMIN and TMAX data can provide useful signal to our streamflow forecast."

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

We clarified this in the text, "We tested RF on training data sets of 30 randomly chosen watersheds and observed that the reduction in out-of-bag MAE error is negligible after 2000 trees. We then set `ntree`=2000 for all 86 watersheds.

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

We're not sure what clarification you're asking. M/3 and ceil(sqrt(M)) are only the same when M is divisible by 3. Since *mtry* cannot be fractional, we have the option to round up or round down. But we ended up tuning this parameter and decided to remove this line of text to avoid potential confusion.

Please put this explanation in the text.

The explanation was added to the main text.

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.

This clarification was added.

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

The following clairifcaiton was added to the test, "The MAE scores are heavily skewed towards 0 while RMSE scores are more evenly spread among snowmelt-driven watersheds."

*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.

We addressed this comment above and why we thought it might be confusing to readers to introduce both false-alarm rate and false-alarm ratio considering that it is not the main focus of our paper. You suggest that models with very good ROC curves often have poor performance diagrams and this makes it crucial to show both. While this may be the case, it's not our intention to argue that partial ROC in our study is "good". In fact, our main observation is that RF becomes "expectedly less skilful in its forecasts with increase in magnitude of the events." In the rest of the discussion in this section, we followed closely with the interpretation of POD and FAR, which we clearly defined.

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.

We addressed this comment above.

3 Reviewer's comments to the manuscript

We here address the inline comments. Simple typo fixes were made directly in the revised manuscript.

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

We believe the current specification is consistent with the current literature on VIC model.

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 replaced "the selection of input variables" with "the input variables provided" as you suggested.

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.)

We made the following modification to the text, "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 (Fig. 2). 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 (Fig. 2)."

Please explain this more clearly. (I also made this comment in round 1.)

This is a bit vague for us to provide clarification. Our current explanation, "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," is consistent with the principle of regression tree explained in the Elements of Statistical Learning text [Friedman et al., 2001].

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

We believe Breiman developed both feature importance measures: Gini-based in an earlier paper on classification and regression trees [Breiman et al., 1984] and permutation-based in the original RF paper [Breiman, 2001]. The following clarification was added, "There are two built-in measures for assessing variable importance in RF : mean decrease in accuracy (MDA) and mean decrease in node impurity (MDI). Both were developed by Breiman [Breiman et al., 1984, Breiman, 2001]."

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

Thanks for the suggestion. We replaced "average gain in residual error reduction" with "average error reduction".

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

We added "Pearson" to the main text.

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.)

As you suggested in round 1, Fig. 1 was first referenced in the Introduction (line 63) when the study area was first mentioned. We referenced it again here at the end of the first sentence of this paragraph in the revised manuscript, "In this study, we focus on watersheds in the Pacific Northwest

Hydrologic Region (Fig. 1)." We think it becomes redundant to keep referencing it after this point because we remain talking about the same region.

More moderate than what? Does this statement apply to the whole domain, just the west side of the Cascades, or what?

We provided the clarification to the manuscript, "For this region, proximity to the ocean creates a more moderate climate with a narrower seasonal temperature range compared to the inland areas, particularly in the winter."

? Similar to what?

We made the following modification, "Here, we observe a similar trend in R^2, KGE, MAE, and RMSE scores compared r values in Fig. 6 where RF performs better in snowmelt-dominated than in rainfall-dominated (higher R^2 and KGE, lower MAE and RMSE)."

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

We addressed this comment above.

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

We addressed this comment above.

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

Thanks for spotting this. We changed the text to the following,"Surprisingly, pentad comes third and fourth in MDI and MDA respectively."

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

Yes, and we modifed the text in the manuscript.

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

Precipitation data is generally zero inflated (at least 30 percent in our dataset depending on the watershed). As a result, there is a high likelihood that the day with zero precipitation ends up with the same value during the shuffling process used to compute MDA. While we did not perform additional simulation to explore this as it is out of the scope of our paper, we believe it is worth discussing and can be investigated in future research. This is the reason our conclusion is that ,"We suggest RF users to exert cautionwhen interpreting outputs from these two measures." We added this to the discussion.

What is a "potential" predictor variable?

We deleted "potential" in the manuscript.

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

We agree that both variables, precipitation and temperature, are not categorical variables and removed "their number of categories" from the text. However, among our 8 predictors in our study, pentad is considered an ordinal variable. Also, the scales of measurement of precipitation and temperature variables are slightly different. Precipitation is a flux variable and comprises discrete and continuous components in that if it does not rain the amount of rainfall is discrete whereas if it rains the amount is continuous. Temperature is a state variable and always continuous. Therefore, we believe the findings in Strobl et al. (2007) are relevant for our discussion on variable importance.

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?

Thanks for the opportunity to clarify this. Yes, temperature variables tend to have more correlation with other predictors than do the two precipitation variables in our dataset. This is likely because temperature controls both the form of precipitation (snowfall vs rainfall) as well as timing of snowmelt. However, due to the blackbox nature of ML models, we don't know for sure if this is directly related to the observed patterns in MDI and MDA.

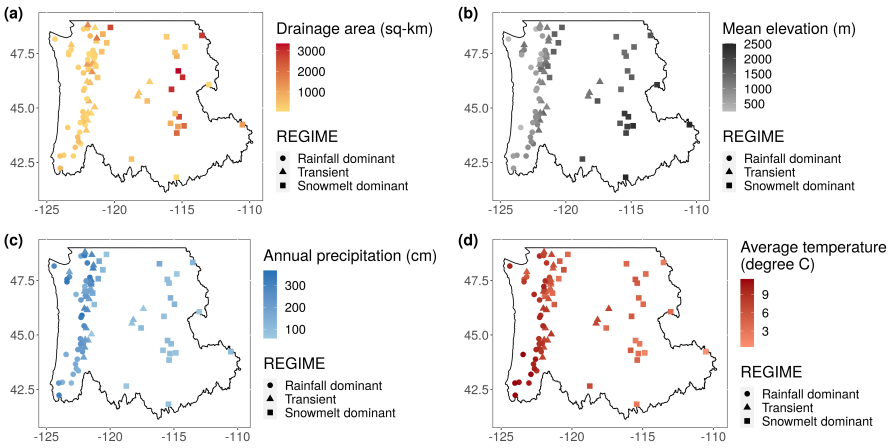
What do you mean by the "stability [of a measure] across different datasets"?

These separate simulation studies used different datasets and concluded that the variable importance ranks based on the two measures (MDA and MDI) can be unreliable.

The font here is way too small. Please enlarge.

Thanks for your suggestion. Please see the modified figure below.

Figure 5: 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.



The full range (from min to max)?
That is correct.

References

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

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

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

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

D. R. Cayan, K. T. Redmond, and L. G. Riddle. Enso and hydrologic extremes in the western united states. *Journal of Climate*, 12(9):2881–2893, 1999.

X. Chen and H. Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

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

G. James, D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning, volume 103 xiv of, 2013.

M. Kuhn et al. Building predictive models in r using the caret package. *Journal of statistical software*, 28(5):1–26, 2008.

X. Li, J. Sha, and Z.-L. Wang. Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal*, 64(15):1857–1866, 2019.

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

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

M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.

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

P. Probst, M. N. Wright, and A.-L. Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

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. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

J. E. Shortridge, S. D. Guikema, and B. F. Zaitchik. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7):2611–2628, 2016.

A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*, 2017.

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

J. N. Van Rijn and F. Hutter. Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2367–2376, 2018.

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

F. Wilcoxon, S. Katti, and R. A. Wilcox. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259, 1970.