Although the reviewer has chosen to reject the manuscript in current form, the authors appreciate the reviewer for his/her valuable comments and suggestions. We would like to take this opportunity to provide further clarification and improve our manuscript. We included below the review's comments (italic black font) and our responses (plain blue font).

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 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.

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.

3 Inline comments

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

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."

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."

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.

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.

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

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.

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.

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].

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].

*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.

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.

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.

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

For the validation period, we calculated the 90^{th}, 95^{th}, and 99^{th} 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.

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.

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

We added the following map and table to Supplementary material.

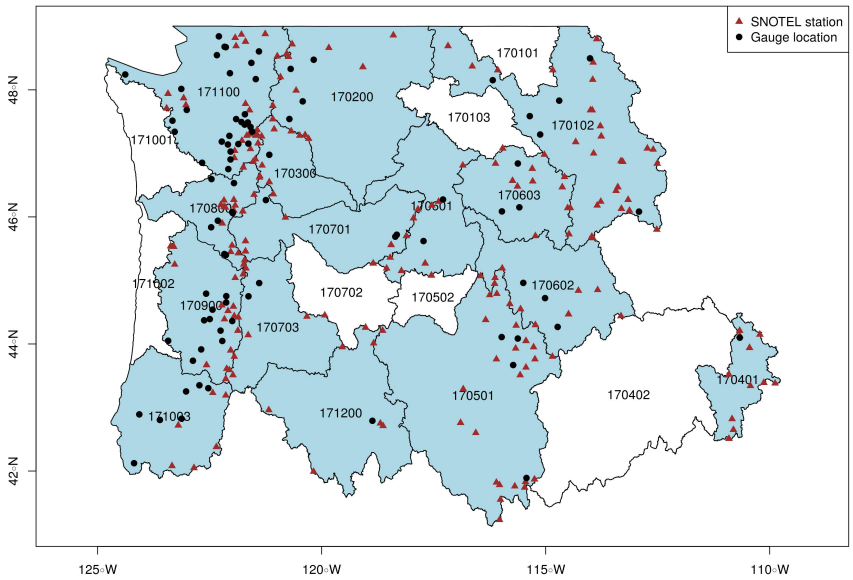

Figure 1: Map of basins within the Pacific Northwest Hydrological Unit. Blue basins contained at least one watershed and were included in the the study.

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.

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.

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?

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].

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.

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.

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: 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.

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.

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.

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.

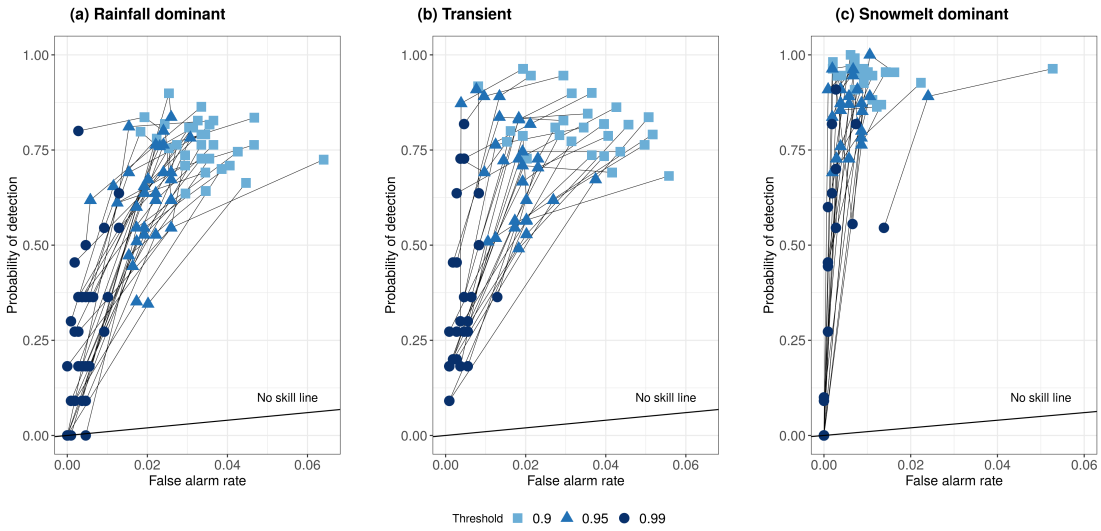

Figure 2: Probability of detection plotted is plotted against false alarm rate for three extreme thresholds: 90^{th}, 95^{th}, and 99^{th} percentiles.

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).

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.

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.

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.

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

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.