The authors would like to thank the reviewer for his informative and valuable comments. We appreciate the opportunity to provide further clarification and improve our manuscript for final submission and consideration. We included below the review's comments (italic black font) and our responses (plain blue font).

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.

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.

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.

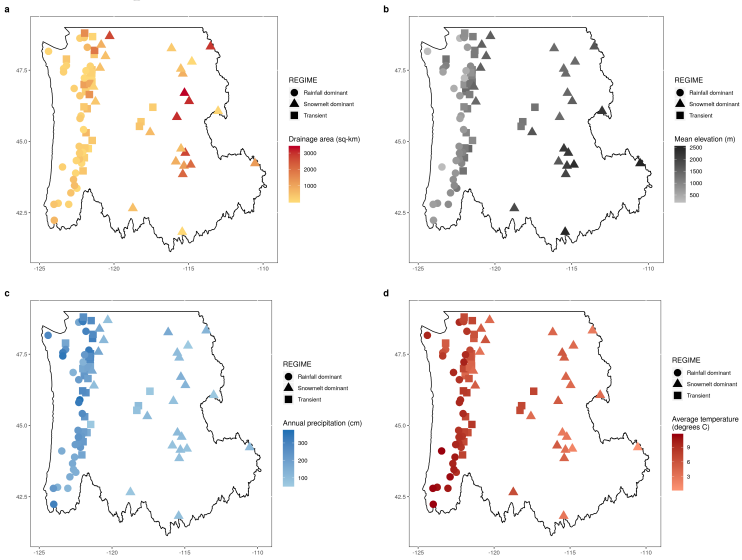


Table 1. Number of streamflow gauges used in the study for each flow regime, ranges of mean watershed elevation and drainage area. Complete catchment physical and hydro-climatic characteristics for each watershed can be found in Appendix A.

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.

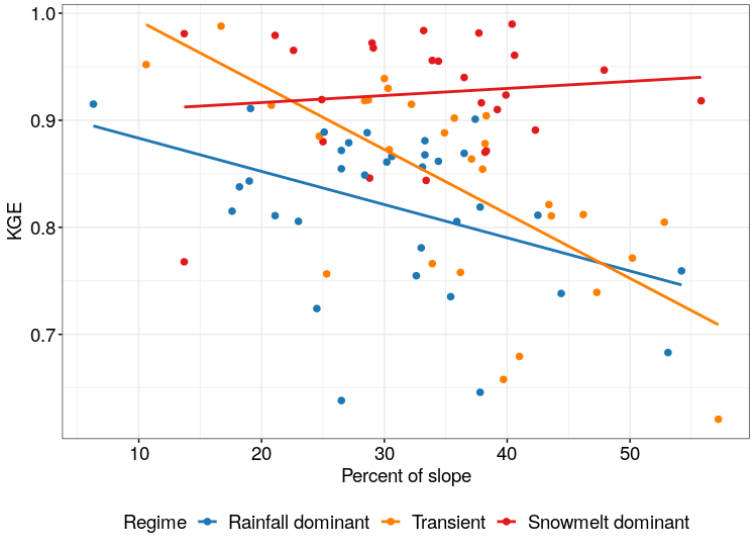
- Line 307: may these be due to rain on snow?

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)."*

References

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.

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.

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.

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.