Again, we thank the reviewer 2 for valuable comments and suggestions. Please see below our responses to the comments. The reviewer's comments are in black font and our responses are in blue font. Changes in the manuscript are highlighted in red.

1 Major comments

Major comment 1. To me, the discussion at the end of Section 4.5 is still very unclear. See inline comments.

Please see below our responses to comments related to Section 4.5.

Line 378: Does "in this group" mean "for snow-melt-dominated watersheds"? (I made this comment in round 2 as well.)

Yes, this is the case and we did miss this comment. We added the clarification to the text, "Precipitation does not seem to have significant contribution to the model's accuracy among the snowmelt-dominated watersheds."

Lines 394-396: Why would MDA, but not MDI, underestimate the importance of variables with a non-normal distribution? (I made this comment in round 2 as well.)

We addressed this comment in our response to the reviewer in Round 2 (page 8). Here was our response, "*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 caution when interpreting outputs from these two measures.*"

The following discussion was added to the text after Round 1 revision, "*In our precipitation data (both training and validation), at least 30 percent of the daily observations are 0 across the watersheds. There is a high likelihood that the day with zero precipitation ends up with the same value during the shuffling process, thus potentially affecting the randomness created to compute MDA. While we did not perform additional simulation to further confirm whether MDA and MDI measures are sensitive to highly-skewed and zero-inflated variables, this can be a topic of future research.*"

Lines 400-403: But you standardized all the predictors to range from 0 to 1, right? So the random forest saw predictors only in normalized units, not in physical units. So the "scale of measurement" should have had no impact on your importance measures.

Actually, "'scale of measurement" does not only refer to the numeric range but also the nature of the data (for example, ordinal vs continuous). You are correct that RF saw predictors only in standardized units, not in physical units. However, standardization and normalization do not change the nature of the data. For example, precipitation data is zero-inflated and the standardization does not make it more normally distributed like temperature data. These zero precipitation data points are just scaled into different values. We addressed this point in the previous revision (page 8). Here was our response, "*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.*" Again, the cited literature does suggest that the scale of measurement can have an impact on RF variable importance measures and we believe it's important that the readers are aware of this.

Line 401: What is a "potential predictor variable"? (I made this comment in round 2 as well.)

Sorry, we did miss this. "Potential" has been deleted from the text.

Lines 401-403: 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? (I made this comment in round 2 as well.)

We addressed this comment in our previous response (page 8). Here was our response, "*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) and the 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. Therefore, the key takeaway of this discussion was that, "*We suggest RF users to exert caution when interpreting outputs from these two measures.*" We added the following discussion to the manuscript, "In our study, temperature variables tend to have more correlation with other predictors than do the two precipitation variables. This is likely because temperature controls both the form of precipitation (snowfall vs rainfall) as well as the timing of snowmelt."

Line 403: I still don't know what you mean by "stability" here.

We added the following clarification, "There is also an ongoing discussion regarding the stability of both measures, in which the two variable importance measures can yield noticeably different rankings, in different simulated datasets (Calle and Urrea, 2010; Nicodemus, 2011; Ishwaran and Lu, 2019)".

Major comment 2. The authors neglected to make a lot of the recommended changes in round 2, and they did not make this clear. If you disagree with a recommendation, you are free to push to back at me. After all, that's part of the review process. However, I find it disrespectful that the authors just ignored a handful of comments. I put a lot of time into the review process, and I don't like having to sleuth around and figure out which comments the authors ignored.

We truly appreciate the reviewer's time in providing us with valuable comments and believe that the manuscript has improved thanks to the changes made in the two rounds of revision. It was not our intention to ignore the comments but we do acknowledge that we missed a handful of them in the revision process. We apologize for this. In this revision, we addressed all of comments point by point.

2 Other inline comments

Line 28: Replace with "less" (I made this comment in round 2 as well).

We agree that "less" is more appropriate and replaced "fewer" with "less" in the manuscript."

Line 35: Insert colon (I made this comment in round 2 as well).

Sorry we did miss this. The colon has been added.

Line 48: Do not capitalize (I made this comment in round 2 as well). The fractions are awkward. Please replace with "0.0625" and "0.05".(I made this comment in round 2 as well.)

We address to these comments in our response to the reviewer (page 7). Here was our response, "*We believe the current specification is consistent with the current literature on VIC model.*" In this paper published by HESS, Variable Infiltration Model (VIC) is capitalized and spatial resolution is in degrees (`https://hess.copernicus.org/articles/17/721/2013/`).

Lines 50-51: Insert comma (I made this comment in round 2 as well). Insert comma (I made this comment in round 2 as well).

We did miss these. The commas have been added.

Lines 81-82: The nested parentheses are awkward. Please rearrange the sentence in a way that gets rid of them.

We have modified the sentence," Both model-agnostic, such as permutation-based feature importance (Breiman, 2001), and model-specific, such as gini-based for RF (Breiman et al., 1984) and gradient-based for ANNs (Shrikumar et al., 2017), interpretation methods can provide useful insights into how the ML models make their predictions."

Line 87: Replace with "referred to".

We replaced "referred" with "referred to".

Line 99: Please explain this more clearly. I doubt that readers unfamiliar with random forests and ML will understand this. (I made the same comment in rounds 1 and 2, and it was ignored.)

We addressed this comment in our response in Round 2 revision (page 7). Here was our response in Round 2 revision, "*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)".* We reviewed other papers on the applications of RF and believe that current explanation in the sentence is accurate and intuitive in terms of how a decision tree regression works. This is also illustrated mathematically in Algorithm 1 Step 3. We added this reference to the manuscript, "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 (Algorithm 1, Step 3)."

Line 119: You need punctuation here. I suggest inserting a comma here, which means you will need to replace the colon before "ntree" with a comma.

The commas were added. We also removed the colon. Here is the new sentence, "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."

Line 121: What do you mean by this? Any hyperparameter is tunable.

The goal of tuning is to obtain the optimal value of the hyperparameter based on a criteria (lowest MAE for example). It has been theoratically proven (please see the cited literature) that more trees are always better (yielding lower MAE). In other words, optimal ntree value can go to infinity. The reduction in error, however, becomes negligible after a sufficiently large number of trees (we discussed this on lines 111-112). We added the following modification to the manuscript, "Second, `ntree` in a forest is a parameter that is tunable but not optimized and should be set sufficiently high (Oshiro et al., 2012; Probst et al., 2019) for RF to achieve good performance."

Line 121: I don't understand how this sentence justifies experimenting with ntree instead of the many other hyperparameters available.

In the manuscript, we explicitly stated that "*all considered parameters might have an effect on the performance of RF*" and we chose to focus on ntree and mtry. The main reason is that these two parameters were originally introduced by Breiman (2001) in the development of RF algorithm. In this sentence, we cited the literature which suggests that *"ntree should be set sufficiently high (Oshiro et al., 2012; Probst et al., 2019) for RF to achieve optimal performance"*. In other words, a high number of trees is essential for optimal RF model performance. Yet what "sufficiently high" means will differ from one dataset to another. This is why we needed to experiment with ntree in our study.

Lines 187-188: After standardization, both the training and validation data range from 0 to 1, right?

This is not the case. Only training data range from 0 to 1. Validation data are considered new data and therefore can have values outside of this range.

Line 310: are

This has been fixed.

Line 315: Replace with "at most" or "at most individual".

We replaced "at individual watersheds" with "at most individual watersheds"

Line 315-316: Is this shown anywhere in the paper, or does the conclusion from a separate analysis? I'm not asking for another figure, but if the conclusion comes from a separate analysis, please specify "(not shown)" at the end of the sentence.

Yes, this is shown in Figure 7a and does not come from a separate analysis. Watersheds that yield lower r values in the naive model are considered to have lower persistence (we defined persistence as the correlation between streamflow of day t and $t+1$ on line 156-157). We added the following clarification to the manuscript, "In Fig. 7a, we observe most points lie on the left of the 1-to-1 line, suggesting that RF outperforms naïve model at most 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 (lower r values from the naïve model)."

Line 328: Replace with "the r-value trend".

We replaced "r values trend" with "r-value trend".

Line 328: Insert comma after "Fig. 6".

We added the comma.

Lines 345-346: Why use someone else's mean-flow benchmark, instead of computing the mean-flow benchmark for your own dataset? I imagine that this wouldn't take a lot of effort, and the mean-flow benchmark could differ a lot between your dataset and theirs.

I just read your response to reviewers, where you said the following: "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."

My response to that: "So the mean-flow benchmark is truly independent of the basin? If so, please make this clear in the manuscript. Otherwise, it's unclear why you're using someone else's mean-flow benchmark, rather than computing the benchmark for your own data."

Yes, the mean-flow benchmark (such as the Nash–Sutcliffe efficiency (NSE and the KGE) are both independent of the basin. We made the following modification to the manuscript, "As observed mean flow is used in the calculation of KGE, Knoben et al. (2019) suggested that a KGE score greater

than -0.41 indicates a hydrologic model improves upon the forecast with mean flow, independent of the basin."

Figure 9: Please explain (either in the caption or in the main body where you introduce Figure 9) that these ROC curves are different than typical ROC curves. i.e., Since you plot only 3 thresholds, the x-axis does not go all the way from 0 to 1. For people used to looking at ROC curves, like me, this looked very wrong until you explained it.

We have added the following clarification to the caption of the plot, "It is noted that the scales of the horizontal and vertical axes are not 1-to-1 in the plotted partial receiver operating characteristic (ROC) curve."

The full range, from min to max? I also asked this question in round 2, and it was not answered.

We addressed this question in Round 2 revision (page 9). And that is correct. The bar represents the full range of values. We added the clarification to the manuscript, "Figure 10. Barplots show importance of predictor variables using (a-c) MDA and (d-f) MDI criteria. Length of the blue bars indicates the median value across the watersheds for each flow regime and the thin black bar represents the full range of the values."