Responses to Reviewer 2

This manuscript introduces a new type of postprocessing of numerical weather forecast using MOS random forests. It clearly describes the methodology, highlighting both the advantages and limitations. While the structure of the paper is closed to Schlosser et al. (2019), I consider the manuscript contains enough new results to be published in Nonlinear Processes in Geophysics. The manuscript can be considered as an update of Schlosser et al. (2019). I therefore recommend publication after minor revisions. Please find my specific comments and technical corrections below.

Thank you for taking the time to read our manuscript and for your constructive remarks. Please find point-by-point responses to your specific comments below.

Line 23: What is the meaning of "homogeneous" here? Please elaborate.

Here, homogeneous means that a single MOS model with constant coefficients can be used to effectively postprocess the forecast. That is, the systematic biases and miscalibrations of the numerical weather model are relatively constant within the dataset. We have clarified this in the text.

Line 37: Please cite some references of random forests used to perform ML-based postprocessing.

Three references are found in the second half of the sentence. We have now added a fourth reference as well.

Lines 107-108: What is the source of the citation, if it's a citation?

This is not a citation and the quotations were only used for emphasis. To avoid confusion, we have removed the quotations and instead italicized the phrase.

Line 179: It should be added that July 19, 2011 is missing for all the 95 stations in package RainTyrol (version: 0.2-0, date: 2020-01-13). This might not affect the results, but it is important to mention about it.

Thanks for noticing this! The missing day is in the last four years of the dataset that were used for model evaluation. Naturally no forecasts were made or evaluated for this day. We have mentioned the missing day in the text.

Lines 239: Please summarize the physical meaning of this, i.e., the mechanism, rather than only listing the variables. Or is it due to GEFS to generate more days with small amount of rainfall than observed? (which seems to be suggested by the authors.)

It is difficult to give a physical explanation for the particular MOS model in this terminal node. This is also not our objective since ultimately the postprocessed forecasts are obtained from a MOS forest and not the individual tree. Our goal here is to highlight the statistical interpretation of a MOS tree's terminal node: that is, to explain how a tree outputs a particular MOS model for the predictor values corresponding to the preceeding splits. We have now adjusted the text to emphasize this distinction.

Lines 264-267 and Figure 5: This is an interesting result. Particularly compared to Fig. 8 of Schlosser et al. (2019), which shows a less organized spatial distribution of best postprocessing. But the question of why the NE-SW distribution is not really discussed in the main text. Is it solely due to the topography? From the figure, it seems the terrain is lower to the NE and higher to the SW. Also, as the MOS random forest is weather adaptative, could this spatial distribution be linked to the main mode of variability of weather in July? (either in the real world or in the GEFS world). It would be interesting to discuss this possibility in the manuscript.

When compared to Fig. 8 of Schlosser et al. (2019), our results show a much more organized spatial distribution of the best postprocessing method. This is mainly because we also compared our method to quantile regression forests, which do not assume a parametric response distribution. In contrast, Schlosser et al. (2019) compared various methods that all assume the same parametric distribution. It seems that this assumption does not hold as well for the southern and western regions where the terrain is higher. We have now mentioned this in the revision.

MOS forests are not the only weather-adaptive model we compared: both the distributional forests and quantile

regression forests are also weather-adaptive. It is difficult to link the observed spatial distribution to a certain mode of weather variability in July since we cannot compare our results to any other months.

Line 288: "new stations (or measurement instruments) are installed all the time". Is "new" here equivalent to "additional" or to "in replacement of"? If it is additional, is it to document highler altitudes in the case of complex terrain, knowing the costs of installation and maintenance? Is it true at the globe scale? I am not sure this assertion is necessary (and accurate) here. On the other hands, if this correct, it might in fact introduce more biases in dataset (instrument drift, error in transcription, system failure...). Please elaborate.

The "new" here refers to either additional stations or modifications to existing stations (e.g., different sensors, orientation, changing surroundings, etc.). You are exactly correct that including such changes can introduce biases in the dataset. Subsequently, there is very little data avaiable for new stations and these greatly benefit from a *robust* postprocessing method such as MOS forests.

Line 289: How would you interpret this citation in the context of this study? Does it suggest that because the postprocessing is weather adaptive, it is constrained by the model's world weather?

The citation "data samples containing numerical model output are a perishable commodity" means that the datasets available for training postprocessing models often have a limited size. Weather-adaptive postprocessing methods are able to take into account many predictors during postprocessing, but therefore require more data for training. The advantage of MOS forests is robustness: the ability to account for these additional variables and issue skilled forecasts even when the data size is limited.

This study only focuses on July. It would be interested to see the robustness on this approach in other seasons, because the regional influence (main modes of variability) vary along the year. And could be linked to the comment on Fig. 5, i.e., if we see a change in the spatial distribution of best postprocessing.

We are sticking to the July data contained within RainTyrol to allow for a better comparison with Schlosser et al. (2019), who also considered weather-adaptive postprocessing based on generalized additive models for location, scale, and shape (GAMLSS). These GAMLSS models incorporate additional predictors into postprocessing using variable selection based on prior knowledge or with boosting.

From this, another question would be: Is it possible to objectively define the most appropriate postprocessing? From the main mode of weather variability? Please discuss the possible directions.

We are not sure what this question entails exactly. Do you mean whether it would be possible to establish a kind of "meta learner" which predicts which learner performs best for a certain setup (e.g., based on topography and climatology)?

If so, we think that this might be possible but is probably not straightforward. So far we have always performed some form of benchmark/validation study, possibly in combination with prior expertise, to select the best learner/model for a given area of application.

Throughout the manuscript, the author sometimes use the term "MOS forests" (for example l. 12 or l. 55) and sometimes the term "MOS random forests" (l. 43 for example). Please review the whole manuscript to homogeneize the terms (and maybe use an acronym).

We have kept the full name of the method "MOS random forests" in the title and the abstract, but used the abbreviation "MOS forests" in the body of the manuscript. Use of the abbreviation is now explicitly addressed the first time the method is mentioned in the introduction (second to last paragraph).

Line 8: "ML" is not previously defined.

We have written out "machine learning" and dropped the abbreviation.

Line 95: Maybe write "...the postprocessing literarure is the nonhomogeneous Gaussian ...".

Thanks, we have adjusted the sentence for improved readability.

Line 107: the sentence looks incomplete ("...a single MOS tree partitions the predictor space ...").

We think this is a complete sentence. The word "partitions" functions as a verb.

Line 180: I would a short sentence to tell how this number is defined, i.e., "median of all estimated power coefficient (Stauffer et al. 2017a)".

This has been added.

Line 184: It would be easier for the readers to mention that the authors are specifically referring to Table 1 of Schlosser et al. (2019).

Yes, that is true. We have now explicitly referenced Table 1.

Line 221: Please indicate where is the station of Axams located in Figure 5 (preferred). Or at least refer to Fig. 8 of Schlosser et al. (2019).

We have now referred to Fig. 8 of Schlosser et al. (2019).

Line 270: It would be better to define "PIT" and "PIT histograms" here.

You're right. We have added the full name "probability integral transform" and a noted that uniform histograms indicate good calibration.

Line 283: "very little data". Do the authors mean small sample size?

Yes, we have adjusted the text.

Figure 2: A "l" is missing in "Dashed and solid lines ..."

This has been corrected.

Figure 3: What is the meaning of "Location" and "Scale"?

Location and scale refer to the two distributional parameters (μ and σ, respectively). This has been clarified in the caption.

Figure 5: Shouldn't it be "CRPSS"? Also, the background is not described in the caption. And the size of the circles should be included in the legend.

We have modified the figure to show the CRPSS with respect to the second best method, rather than differences in CRPS, as was originally done. The size of the circles is now also included in the figure and the background shading described in the caption.