Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-191-RC1, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development Discussions

Interactive comment on "Recalibrating Decadal Climate Predictions – What is an adequate model for the drift?" *by* Alexander Pasternack et al.

Anonymous Referee #1

Received and published: 19 October 2020

The manuscript builds on the post-processing procedure DeFoReSt proposed by Pasternack et al 2018 and presents a boosted re-calibration of decadal climate predictions. The manuscript describes a well thought approach on handling drift corrections and present it reasonable well for an statistical audience. The comparison between the boosted and the non-boosted calibration is excessive and described well, but lacks hypothesis testing to determine the actual differences between the two approaches outside of the argument that it is obvious.

While further work is required on the general presentation to make it more accessible to a wider audience, the authors might reconsider the choice of the journal, as the extreme focus on the statistical approach might be more appropriate for NPG. In its current shape the manuscript needs a much better illustration what has been done and

why it matters. Therefore, I recommend major revisions for the manuscript and would expect a rework of the figures and potentially the structure of the arguments.

Specific points:

18: "Significant advances could be achieved by recent progress in model development, data assimilation and climate observation." -> has been made

25: "unconditional, and conditional" -> unnecessary comma

37: "third/second" -> why third before second?

47: "objective function": objective has a specific meaning in statistics (see Jeffrey's prior) and would have to be individually proven. It is an unfortunate choice of word as it plays into the idea that statistics might be objective. As such, the word objective should be omitted in the manuscript completely.

87: "For the sake of completeness and readability these are presented in this section again." - Unnecessary sentence

124: By introducing the normal distribution with an calligraphic N and then use for the standard normal distribution greek letters, it gets quite confusing. As such this part needs to be rewritten. I would suggest to introduce N_S or similar for the standard normal distribution. As the authors work beforehand with large letters for CDFs, I would recommend to use a consistent approach for the nomenclature. I am aware that the equation for the CRPS is shown in this way often in statistical leaning literature, but as GMD is not such a journal I strongly recommend intuitive naming of variables.

138ff: I would strongly recommend a schematic on which basis the authors explain the mechanism of DeFoRFeSt. Equations are fine, but as they become extremely lengthy and hard to understand for the general reader (like eq 13), they need support and motivation.

185: Figure 1: name it consistent with Fig. 1 or rename all Figs to Figures.

202ff: The problem at this point is that the boosting algorithm forms an essential part for the understanding of the manuscript. I would strongly recommend the design of a schematic to make clear what exactly is done in the boosting process (apart from the equation, but the algorithmic strategy). This part of the manuscript needs effort to make it better understandable for the wider audience, especially as the authors do not publish here for a statistical, but a general model related audience.

202: "R-function poly" please make it a proper reference

205: "R-package crch" please make it a proper reference

206: "http://cran.r-project.org/" should go into the references

218: The way it is written the choice of nu requires a sensitivity test. So either it requires the motivation for choosing nu = 0.05 to be rewritten, or a demonstration and discussion of its effect.

226: The description of the cross-validation is not sufficient. A CV requires the statement on how the non-training data is afterwards evaluated (without taking into account the training data, otherwise it is not a CV but a Jackknife). The authors point to equation 21, but it is just the basis for the validation (which is described in line 216 with the Pearson correlation). So it would be required to state exactly what process is used for validation, which data is used for this step and which exact metric is applied to make the statement on a validated result.

238ff: Again the authors try in this section to explain everything by equations without explaining to the readers what consequences each of the decisions made have. The authors talk about extreme toy model experiments (l. 238), but do not state in what manner it is extreme. Then the authors introduce 5 parameters determining the experiments, but fail apart from short descriptions (like (un)conditional bias) to explain the reader what this actually means (and yes I am aware that most will know what it means in the direct community, but I think the authors should make the effort to explain it better

as it builds a foundation of their argument). So I would recommend here to create a figure explaining the consequences of each of the parameters to give the modelling community an entry point to follow the experiments to find analogues between the toy model and the usually used GCMs or similar (this has been done in Pasternack et al 2018, but perhaps a even more simplified/schematic version of Figure like Fig. 1 there will help). Giving the reader only an entry point by table 1 is not enough.

267ff: The authors show a very large figure with many elements in 4 main colours for the different parameters, but just spend three sentences without putting it in context and give the plot any meaning (e.g. comparison, interpretation apart from first three coefficients vs. last three). As such either the plot has not more information, then it is doubtful whether the plot has any use for the manuscript, or the many different whisker plots are important and it is not represented in the text. Just showing them is not enough, especially as later it is not referenced back to the figure when similar coefficient plots are made.

281: Estimating the 0.025 and 0.975 percentile from just 100 experiments is not a good way to demonstrate significances. The authors should either choose more experiments or go to alpha = 10. Or the description is so misunderstand-able that in fact more than 100 values to estimate the percentiles are used. In that case the section has to be rewritten.

283: (see 4) : What is referenced here?

285ff: Is there a reason, why in the DeFoReSt mode close to all metrics from Fig 3-10 show a U-shape over the lead years?

288: It is not explained why the uncertainties of the ESS are not visible (either small or not calculable).

330ff: Two consecutive sentences start with "Here,"

334 Why is there a bootstrapping in this section but not in the section above?

340ff: Why is there no comparison to the coefficients in Fig. 2?

348: "have also some impact." This should be analysed with a significance test and statements made accordingly

376: Are there significant differences between global and NA 2m-Temperature? Why is North Atlantic framed here as independent compared to the global and the comparison between those kept so short? It seems like it is written currently that one example would be sufficient. So why are the two not conclusively compared with each other in one section? So could there be a different story apart from just showing the statistical model applied to data?

Fig3-5 should be combined in one figure with 9 panels

Fig7-9 should be combined in one figure with 9 panels

Fig 6+10 potentially better to have them in one plot with 2 panels

Fig11: MiKlipl -> MiKlip

Fig12+14: Even when it is a stylistic choice: Why have the authors chosen a different colour-scheme compared to all the other figures in this manuscript?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-191, 2020.