Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-191-AC1, 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.

Alexander Pasternack et al.

alexander.pasternack@met.fu-berlin.de

Received and published: 15 December 2020

Answer to referee 1

Thank you very much for your informative and detailed comments.

General comments

"The manuscript builds on the post-processing procedure *DeFoReSt* proposed by Pasternack et al 2018 and presents a *boosted recalibration* of decadal climate predic-

tions. 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:

1. 18: "Significant advances could be achieved by recent progress in model development,data assimilation and climate observation." → has been made

   Answer: Will be corrected

2. 25: "unconditional, and conditional" → unnecessary comma

   Answer: Will be corrected

3. 37: "third/second" → why third before second?

   Answer: third before second because the ensemble mean is corrected via a 3rd order polynomial and the ensemble spread via a second order polynomial, which is described a few sentences before (31ff).

4. 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.

Answer: If the name "objective function" is misleading, we will change it to "cost function".

5. 87: "For the sake of completeness and readability these are presented in this section again." - Unnecessary sentence

   Answer: Will be deleted.

6. 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.

   Answer: We will replace the symbols Φ and φ for the CDF and PDF of the standard normal distribution with N_{S_C} and N_{S_P}.

7. 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.

   Answer: We will add such a schematic to the manuscript.

8. 185: Figure 1: name it consistent with Fig. 1 or rename all Figs to Figures.

   Answer: Will be corrected.

9. 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.

   Answer: We will add a schematic flow chart describing the boosting algorithm analogously to Messner et al. 2017.

10. 202: "R-function poly" please make it a proper reference

    Answer: Will be corrected.

11. 205: "R-package crch" please make it a proper reference

    Answer: Will be corrected.

12. 206: "http://cran.r-project.org/" should go into the references

    Answer: Will be corrected.

13. 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.

    Answer: We will add a better motivation to the manuscript.

14. 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.

Answer: We will add a more detailed description.

15. 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

    Answer: We will add a more detailed description and two figures showing the effect of the imposed systematic errors of the toy model. Moreover we will change the phrase '...we consider two extreme toy model experiments...' to '...we consider two toy model experiments with different potential predictabilities...'

16. 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.

    Answer: Showing this Fig. 2 is relevant for the toy model construction since it supports the decision to use the same magnitude for the coefficients of the start and lead time dependent systematic errors. However, since it is not used for any further evaluations we will put it to the appendix related to the table A1 which shows the final coefficients for the toy model construction.

17. 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 misunderstandable that in fact more than 100 values to estimate the percentiles are used. In that case the section has to be rewritten.

    Answer: Indeed, using 100 experiments is not enough for calculating the 0.025 and 0.975 percentile. We will repeat that with 1000 experiments and update the corresponding text passages and figures in the manuscript.

18. 283: (see 4) : What is referenced here?

    Answer: Will be corrected

19. 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?

    Answer: Regarding Figs. 3-10, particularly the ESS and the intra-ensemble variance omit a certain inverse U-shape. The reason might be, that *DeFoReSt* tends to be more underdispersive for the first and last lead year due to the missing additive correction term for the ensemble spread.

20. 288: It is not explained why the uncertainties of the ESS are not visible (either small or not calculable).

Answer: We have decided not to show any uncertainties for the ESS, since we just wanted to show the general effect of *boosted recalibration* and *DeFoReSt* and to ensure a better visibility.

21. 330ff: Two consecutive sentences start with "Here,".

    Answer: Will be corrected.

22. 334 Why is there a bootstrapping in this section but not in the section above?

    Answer: Unlike Sec. 4 we evaluate in Sec. 5 the CRPSS also w.r.t. a raw model. Thus, we decided to apply a bootstrapping approach to avoid any advantages of the post-processed models.

23. 334 Why is there a bootstrapping in this section but not in the section above?

    Answer: Unlike Sec. 4 we evaluate in Sec. 5 the CRPSS also w.r.t. a raw model. Thus, we decided to apply a bootstrapping approach to avoid any advantages of the post-processed models.

24. 340ff: Why is there no comparison to the coefficients in Fig. 2?

    Answer: The coefficients in Fig. 2 were used to derive the scale of the coefficients associated to 4th to 6th polynomials for the pseudo-forecasts. Here, unlike Fig. 11 and 13 no model selection was applied, i.e. a comparison is not very reasonable.

25. 348: "have also some impact." This should be analysed with a significance test and statements made accordingly

    Answer: We will change the statement "have also some impact" to "have also been identified by the boosting algorithm as relevant".

26. 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?

Answer: *DeFoReSt* and *boosted recalibration* have been developed within MiKlip project. Here, the NA as well as the global 2m-temperature are the key variables within this project. Moreover these regions distinguish themselves by their potential predictability. Thus analog to the toy model experiments we show the mechanisms of theses recalibration approaches to MiKlip predictions with smaller and higher potential predictability. Furthermore, regarding the different identified predictor variables for the NA and global 2m-temperature (Figs. 11 and 13) one can see that other processes are relevant due to a different spatial scale of these examples.

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

    Answer: Will be corrected.

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

    Answer: Will be corrected.

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

    Answer: We would like to keep these plots separate, since they are discussed in different sections. Thus, to ensure a better readability it may be better to show these figures separately.

30. Fig11: MiKlipl → MiKlip

    Answer: Will be corrected

31. 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?

Answer: We wanted to distinguish the toy model results optically from the results based on MiKlip data.

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