Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-191-AC2, 2020

Geoscientific
Model Development
Discussions
Open Access

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 2

Thank you very much for your informative and detailed comments.

1. General comments

"The authors present an extension to their previously introduced recalibration approach for decadal climate forecasts. The existing method is extended with a model selec-

Printer-friendly version

Discussion paper

tion approach using boosting to infer a parsimonious model from the data. Strengths and limitations of this approach are tested using synthetic data and an application to global mean and North Atlantic temperature forecasts is presented. While the boosting method presents a welcome addition to make the approach more generally useful across a diversity of applications (not limited to decadal forecasting) and therefore certainly merits publication, the article lacks in a few key aspects detailed below. Therefore, I suggest to accept the article subject to major revisions."

1.1 Interpretation of the results

"The authors focus on descriptive verification measures to discuss the results from *boosted recalibration*. In addition, I suggest the authors expand the discussion of the inner workings of the method and the configuration that is identified as optimal with boosting. From a methods perspective, I wonder if the *boosted recalibration* models are of lower complexity compared with *DeFoReSt* (i.e. if boosting actually manages to efficiently constrain the number of parameters). Also, the selected models appear still quite complex given the limited data at hand to train these. Have you explored early stopping rules for the boosting approach (generally skill improves rapidly in the first iterations and levels out afterwards, potentially another criterion for stopping provides better generalization ability through reduced models)? From an application perspective, some more discussion on the identified nature of the error that is corrected with *boosted recalibration* would be useful, *boosted recalibration* is less effective if the systematic error has very simple structure as appears to be the case here."

Answer: The basic feature of the boosting algorithm is to allow a priori for a complex structure of the model used for recalibration but use the complexity only as needed. Thus our procedure is able to adjust complexity according to the problem at hand based on out-of-sample prediction error. This is realized by the automatic selection of the most relevant predictor variables by iteratively updating the log-likelihood. For each iteration step only one coefficient (the one that improves the fit most) is updated and thus complexity is successively increased. Here, the maximum number of iteration steps must be specified beforehand. However, if the chosen iteration step is small enough certain model coefficients are remaining zero. In order to find the best performing model an adequate iteration step has to be identified (model selection step) using a cross-validation setup. For this purpose we split the data into 5 parts and for each part, recalibrated predictions are computed from boosting model at the corresponding iteration step that were fitted on the remaining 4 parts. Afterwards the log-likelihood over all 5 recalibrated parts were summed up. This procedure is repeated for every iteration step. The iteration step with the lowest log-likelihood is considered as the one which provides the statistical model with the best predictive performance. Due to this procedure predictor variables of the statistical model that are not relevant are remaining zero. This can be seen in Figs. 11 and 13 which demonstrate which predictor variables are identified as relevant. Here, one can see that both for the North Atlantic as well as for the global 2m-temperature the complexity of *boosted recalibration* is around 15 identified predictor variables whereas *DeFoReSt* uses 22 predictor variables. We will add a schematic overview of the boosting algorithm and further explanation of the cross-validation approach to the manuscript.

1.2 Link between the toy-model experiments and the application

"The authors quite clearly demonstrate the strengths and limitation of the *boosted recalibration* compared with the reference approach (*DeFoReSt*) using their toy model experiments. There is, however, no direct link drawn to the application of *boosted recalibration* with global mean and North Atlantic surface temperature forecasts. In particular, I would like to know if the lack of improvement from *boosted recalibration* compared with *DeFoReSt* is consistent with the adjustments that are applied (e.g. what errors are generally corrected)."

Answer: With the toy model experiments we show that *boosted recalibration* outperforms *DeFoReSt*, if the polynomial order of the systematic errors goes beyond the restrictions of the *DeFoReSt* design. If that is not the case, both recalibration methods perform equally. Regarding the global mean and North Atlantic surface temperature forecasts one can see in Figs. 11 and 13 that *boosted recalibration* mostly identified predictor variables with a polynomial order smaller than 3. Thus, the fact that *DeFoReSt* and *boosted recalibration* perform equally for recalibrating MiKlip temperature forecasts is in accordance to the toy model results. We will emphasize the connection between toy model and temperature results more in the manuscript.

1.3 Significance assessment

"The significance assessment introduced on L280 does not reflect that the scores between *DeFoReSt* and *boosted recalibration* may be highly correlated due to the same forecast observation pairs being used. The 2.5-97.5% interval on the mean scores therefore likely underestimates the significance of the results. Instead, I propose to use a Diebold-Mariano test or a t-test on the score differences. I expect that using such a more powerful test would allow you to demonstrate e.g. that *DeFoReSt* significantly outperforms *boosted recalibration* when the error dependency matches the assumptions in *DeFoReSt* at least for short lead times."

Answer: Actually, we do not expected that *DeFoReSt* outperforms *boosted recalibration*, because the systematic error in the Miklip data is unknown and therefore does not have to be equal to the *DeFoReSt*-scenario. *Boosted recalibration* is able to cover systematic errors up to the 6th polynomial order, which also includes the the *DeFoReSt*-scenario, but is more flexible due to boosting. One can see in Fig. 11 and 13 that the identified polynomials do not go beyond the 3rd order, which is caught by *DeFoReSt* just as well. To compare these two post-processing methods we applied a bootstrapping approach. Within the applied bootstrapping approach, we calculate the score 1000 times, each with a different sample (replacements are allowed) from the original time series. The corresponding samples for the scores of *DeFoReSt* and *boosted recalibration* are not the same, i.e. a correlation between these scores is avoided. However, if these scores would base each in the same sample a high correlation between those is possible and a Diebold-Mariano test or a t-test would be meaningful, indeed. We will point this out more clearly in the manuscript.

2. Minor comments

1. L72: 1.5°and 40

   Answer: Will be corrected.

2. L74: The full-field initialization

   Answer: Will be corrected.

3. L151-2: the punctuation is somewhat weird, maybe this could be changed: "...drift adjusted ensemble mean forecast (i.e. a deterministic forecast without specific uncertainty quantification)."

   Answer: Will be corrected.

4. L192-4: now is used three times

   Answer: Will be corrected.

5. L209: Maybe mention that you chose maximum likelihood in the following for better readability.

   Answer: Will be corrected.

6. L310: toy model setup with low potential predictability

   Answer: Will be corrected.

7. L314: The ESS (see Fig. 8a-c) reveals that

   Answer: Will be corrected.

8. L325: Typo? Shouldn't this read "the low predictability leads to a increased CRPS" (not reduced CRPSS)?

   Answer: Actually not. In a setup with low potential predictability the benefit of *boosted recalibration* over *DeFoReSt* is smaller compared to a setup with high potential predictability. Thus the CRPSS is reduced.

9. L332: Repetition, use "We discuss..." instead

   Answer: Will be corrected.

10. L337: Typo. 10-year validation period

    Answer: Will be corrected.

11. L368: What fraction of the skill is due to the (linear) trend in global mean surface temperature?

    Answer: This is a very interesting question, indeed. It not possible to answer this briefly. We are currently working on a study where we use a recalibrated climatology as reference for the skill evaluation. The purpose is to analyze to what extent the predictive skill of recalibrated decadal predictions is superior to a statistical model with the same statistical properties as the applied recalibration strategy.

12. L402: Pasternack et al. (2018) show that

    Answer: Will be corrected.

13. L402: *DeFoReSt* leads to improved ensemble...or *DeFoReSt* leads to an improvement in ensemble...

    Answer: Will be corrected.

Printer-friendly version

Discussion paper

14. L409-: Long sentence. Maybe start with “Common parameter estimation and model selection approaches such as stepwise regression and LASSO are designed for predictions of mean values. Non-homogeneous boosting jointly adjusts mean and variance and automatically...regression.”

    Answer: Will be corrected.

15. L423: this is not supported by your figure. *Boosted recalibration* is not (significantly)superior to *DeFoReSt* if errors are ‘simple’ according to Figure 6.

    Answer: Will be corrected.

16. L438: equally

    Answer: Will be corrected.

17. Figure 1: Why not show all the initialization times? The figure would be easily readable even with many more lines and the alignment of the differently colored blocks may become more apparent.

    Answer: We will replace that figure with an new one showing all initialization times.

18. Fig. 3-5 and 7-9: Consider combining figures 3-5 and 7-9 each into one multi-panel plot to avoid splitting the figures across pages in the final publication. Also, the information shown is somewhat redundant and I encourage the authors to drop the sharpness plot for simplicity and for the following reasons: i) the sharpness of the raw model is of no use as it is not calibrated, ii) qualitative statements about the sharpness in *DeFoReSt* and *boosted calibration* can easily be derived from a visual comparison of the MSE and ESS plots. The legend should be shown only once for all 6 (9) panels of the multi-panel plot and axes should be labelled only once per row / column. Finally, consider using a square-root (or log) transform on the y-axis to take away the focus from large differences with large scores.

Answer: Will be corrected. However, we still would like to keep the sharpness figures. Indeed one could derive the sharpness from the ESS and the MSE but we think that is may be more convenient to have a visual impression of the sharpness.

19. Fig. 4: there is indication of extra overconfidence at the beginning and end of the forecast with DeFoReSt (with setups 1-3 and *DeFoReSt*). This appears to be an artefact of the method. Could you please discuss this?

    Answer: Regarding the ESS of the raw model, one can see that for lead year 1 and 10 particularly the setups 1-3 are strongly over- or underconfident. Thus we would explain the inverse U-shape of the pseudo-forecasts after recalibration with *DeFoReSt* with the fact 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. This example shows that *boosted recalibration* can account better for forecasts which are either strongly overdispersive or strongly underdispersive.

20. Fig. 6, 10: Excessive white space. Please adjust the y-axis to better focus on the available data.

    Answer: Will be corrected.

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