Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-191-RC2, 2020

Geoscientific Model Development Discussions
Open Access
EGU

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

Anonymous Referee #2

Received and published: 26 October 2020

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 selection 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.

Printer-friendly version

Discussion paper

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.

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).

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.

2 Minor comments

L72: 1.5° and 40

L74: The full-field initialization

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)."

L192-4: now is used three times

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

L310: toy model setup with low potential predictability

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

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

L332: Repetition, use “We discuss. . .” instead.

L337: Typo. 10-year validation period

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

L402: Pasternack et al. (2018) show that

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

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.”

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

L438: equally

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.

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.

Fig. 4: there is indication of extra over-confidence 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?

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

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

Printer-friendly version

Discussion paper