Review of van der Laan et al. (2025) 10.5194/egusphere-2024-387

General comments

This manuscript by van der Laan et al. presents an impressive and relatively comprehensive analysis of the potential for global decadal glacier mass balance forecasts using bias-corrected CMIP6 data in OGGM. The authors compare the performance of OGGM in a re-forecast (also known as hindcast) setting when run with three different climate forcings: (1) an ensemble of initialized decadal re-forecasts extracted from CMIP6 DCPP-A, (2) baseline persistence forecasts that use forcing from previous years as a function of lead time, and (3) historical GCM simulations from an ensemble of uninitialized free-running CMIP6 outputs that represents the current state-of-the-art. Experiments are carried out on both a set of 279 so-called reference glaciers using high-quality WGMS glaciological mass balance data for calibration and for all $\simeq 214\times10^3$ land-terminating glaciers using geodetic mass balance data for calibration. With a fixed calibration routine it is convincingly shown that the decadal re-forecasts (1) can match or even outperform both the persistence forecasts (2) and historical GCM simulations (3) for ensemble-mean predictions of multi-year glacier mass balance. Crucially, by demonstrating the feasibility of global decadal glacier re-forecasts, this study paves the way for future work on global decadal glacier mass balance prediction that is vital for better capturing the near-term response of glaciers to ongoing anthropogenic global warming. This manuscript is both well written and structured with a clear experimental design. The results should be of great interest to the cryospheric science community, both for global glacier modeling and beyond. Moreover, it has already undergone an initial round of peer review with generally positive reviews. Although I was not involved in the earlier review process, overall I find myself in agreement with these mostly positive earlier comments. I recommend the publication of this manuscript once the minor and mostly technical points raised in the specific comments below have been addressed.

Specific comments

1. L2: Consider changing *"seasonal and long-term simulations"* to *"seasonal forecasts and long-term projections"* to be more precise. Simulations do not have to involve just forecasting/prediction, they could also be reanalyses that leverage historical observational data or scenario-based projections. Here, however, it becomes clear later that your focus is on demonstrating the potential of doing future decadal predictions through a re-forecasting exercise in the past two decades. I think this slight change in language would help situate your study and make your objectives clearer to the reader from the start of the abstract.

2. L37: Consider changing *"developmental"* to *"embryonic"* if you are trying to say that the field of decadal prediction (certainly for glaciers) is still early in its development.

3. L76: Consider changing *"the dynamical evolution of glaciers"* to just *"glacier flow"* since the term glacier dynamics is arguably vague even if it is commonly used in the field.

4. Correct Eq. 1 to

$$m_i(z) = p_f P_i^{\text{solid}} - \mu^* \max\left(T_i(z) - T_{\text{melt}}, 0\right) + \epsilon \tag{1}$$

   In LaTeXcode

```
m_i(z)=p_f P_i^{\text{solid}}-\mu^*\mathrm{max}\left(T_i(z)-T_\text{melt},0\right)+\epsilon
```

   Corrections: (1) Text in the superscript of P_i^{solid} and the subscript T_{melt} should be in text (e.g. `mathrm{}`) mode. These should also be corrected in the text (L96, L97) (2) The max operator should also be in text mode and include the crucial second argument 0 to make it clear that it is the ramp function.

5. L96 z should be in math mode as z (i.e. `$z$`).

6. L97: Avoid starting the sentence with a symbol change to (e.g.) *"Here the precipitation correction factor, p_f, is set to..."*.

7. Figure 2: This is a minor point and to some extent a matter of taste, but I would recommend transposing the axes here so that the observations (reference truth) are on the x-axis while the re-forecasts are on the y-axis. Such a change would ensure that *both* the sign and magnitude of the forecast−observation error correspond to the vertical direction and distance from the 1 : 1 line. That is, a positive error (overestimation) would coincide with a scatter point above the 1 : 1 line and vice-versa. To my knowledge, it also follows a (somewhat unwritten) convention for scatter plots when evaluating the performance of environmental models (Bennett et al., 2013; Pauwels and oters, 2019). Moreover, I would recommend going beyond the great first step of having equal axis limits by also making the axis aspect ratio equal so that the ticks on the x-axis and y-axis are equidistant.

8. L119: While I understand the reasoning for using the multi-model ensemble mean as a point estimate relying on some kind of wisdom of the crowd to get rid of outliers, it is surprising that you would completely disregard the ensemble spread as a measure of forecast uncertainty. One could argue that you are 'throwing the baby out with the bathwater' since by trying to get rid of outliers you are also disregarding the valuable uncertainty quantification inherent in an ensemble. While it is true that the ensemble spread here might be under-dispersive (overconfident), by choosing only the ensemble mean you are making your results degenerate and overconfident by design. I am not asking you to redo any analyses or add ensemble spread statistics, but I think this point should at least be discussed later in the paper in the discussion or outlook. Otherwise it leaves the reader wondering why you made this choice. Note that you are not actually getting rid of uncertainty by focusing on the mean, you are just hiding it. Why not embrace uncertainty and use probabilistic scores (Hersbach, 2000)?

9. L130: Here too superscripts should be in text mode, i.e. 21^{st} not 21^{st}.

10. L138: Consider changing *"Per component of our study"* to *"For each component of our study"*.

11. L139: While I appreciate the emphasis on this point regarding the focus on forcing products rather than model calibration, in reality model calibration and the choice of forcing product exist in a state of strong entanglement. In particular, the absolute and relative performance of the 3 forecasting methods (re-forecast, persistence, GCM historical) may change as a function of the calibration method used. This would hold both in the case that the same calibrated parameters are used with each forcing product (but a different calibration routine is used) or if a different calibration is carried out for each forcing product. I understand that the focus of this study is *just* on the effect of the forcing product, but I would just like to push back a little by emphasizing that the fact that there is not really a clean separation between these two. In particular, the 'optimal' calibration parameters (two of which depend directly on the forcing) are conditional on the forcing product used. Ideally one would perform a sensitivity analysis of the joint (combined) effects of calibration routine and forcing product choice. Here too I am emphatically not asking you to redo any analyses or similar but I would recommend at least touching on this issue in an outlook section.

12. L153: Maybe I am missing something, but here you seemingly contradict yourselves. On the one hand, for component 2 you say that you do not need to use the residual ϵ since the dataset allows calibration with individual glaciers. On the other hand, for component 1 you do use the residual for individual glaciers (glacier-by-glacier basis). So which way around is it? Moreover, I wonder why you would ever not include a residual term in a calibration exercise without making strong assumptions. The residual represents observation and/or model errors and these are in reality never identically 0 even if this is sometimes assumed. Again, I am not asking you to change anything in the analysis but instead to clarify your assumptions.

13. L159: The wording here is somewhat unclear, presumably by 'perfect results' you mean that you are able to match the geodetic mass balance observations exactly. I guess this is not surprising if you have matched the number of parameters to the number of observations and if your model is flexible enough, but I would ask you to clarify what you mean by perfect in this sense do you fit the data on average or each datapoint (or are these the same). It is also not clear to me if you are fitting to a single data point estimating the geodetic mass balance from 2000 to 2020 or several within this period which section Section 3.2 seems to indicate. Perhaps this is obvious to frequent users of this dataset, but all readers should not be assumed to have this background knowledge. A final pedantic comment on this sentence is to consider changing *"mean bias"* to just *"bias"* since the term bias in statistics always refer to a mean (expected) error so the use of 'mean' here is redundant unless you are taking the mean of a mean but that would also require more explanation.

14. L187: Although this is already implicitly quite clear since you mention the reference height, I would nonetheless recommend specifying that this is *"air temperature"* to be more precise at least here when you introduce the climate forcing.

15. L241: Fix the LaTeX notation here, I guess something went wrong and $T't$ is supposed to be $\overline{T'_t}$ or similar.

16. L242: Change t to t (i.e. `$t$`) for consistency.

17. L257: Be more specific here and change *"ensemble"* to *"ensemble mean"* since you do not consider the entire ensemble as a whole (only the mean) or probabilistic error metrics like the CRPS (Hersbach, 2000).

18. L260: In future work it would also be interesting to compare the ensemble skill (not just ensemble mean) of the re-forecasts to the historical GCM simulations. Perhaps something to allude to in an outlook. I suspect that the decadal re-forecasts have better calibrated uncertainty than the historical GCM simulations and this is something that could be quantified with probabilistic skill scores.

19. 272: Again, 'mean error' is synonymous with just 'bias' the term 'mean bias' is generally nonsensical. This term does sometimes appear in the literature since some researchers treat error and bias as synonymous, but the latter is strictly a statistic of the former.

20. L290: Consider clarifying what you are testing here. Is it the difference in the decadal mass balance estimates or the difference in skill (as measured by some metric) between the different experiments? I guess it is the latter, but if so, which skill metric are you comparing in the significance test?

21. L294: The procedure that you perform for the binomial test is also unclear. Did you do a binomial test for each glacier or a test based on statistics (e.g. the fraction with improved skill) across all 279 glaciers? More generally, consider dropping the overly dichotomous NHST framework involving thresholding in future studies and instead report the p-values (or statistics thereof) following recent recommendations McShane et al. (2019). More generally, be cautious of how to interpret the p-value and whether or not it is answering a statistical question that you are actually interested in (Ambaum, 2010).

22. L324: Change *"Analyzing"* to *"Comparing"*.

23. L330: This is a nice example of where reporting the p-values would make more sense rather than using the arbitrary traditional $p < 0.05$ threshold. In particular, the difference in significance here (or rather in the p-values) may itself not be significant (McShane et al., 2019). Maybe the p-values testing the difference in the decadal re-forecast and GCM historical experiment is not that different from the p-values testing the difference between the persistence and the other two experiments. The reader has no idea beyond the fact that the former are $p > 0.05$ (0.07 or 0.5?) while the latter is $p < 0.05$. Again, not something you have to change here, but in future work consider reporting p on a continuous scale rather than an arbitrary threshold introduced haphazardly by a defunct statistician.

24. L380: Regional variations in the skill of simulated precipitation help explain the results, but also raise questions about the choice of a (uncalibrated) constant precipitation factor $p_f = 2.5$. I understand that (1) you are not focusing on calibration and (2) you hope to partly address the lack of skill in the climate simulations of precipitation by using a bias correction based on CRU. However, these CRU data are also coarse and will not necessarily add much skill to precipitation simulations, especially in complex topography. Although this is somewhat speculative, a more explicit joint downscaling/bias-correction to the glacier scale through calibration of p_f could help add further skill to the re-forecasts. Several studies on seasonal snow (e.g. Fang et al., 2023) have shown that precipitation is the dominant source of uncertainty for seasonal snow storage in mountainous regions in global climate (reanalysis) products, and I would expect that this carries over at least partly to the glacier surface mass balance in some regions. I am not expecting you to redo any analyses at this advanced stage that are arguably outside the scope of the paper (since you do not want to focus on the calibration aspect), but I would nonetheless recommend touching on this precipitation calibration issue in the discussion as a potentially valuable topic to investigate in future studies.

25. L393: Consider being specific here and change *"historical simulations"* to *"GCM historical simulations"*.

26. L407: As previously alluded to, you had the opportunity to quantify whether or not there is an improvement in uncertainty quantification (both precision and accuracy) by (e.g.) comparing the CRPS (or some other score) of the glacier mass balance forced by an ensemble of re-forecasts to that forced by an ensemble of GCM historical simulations. As before, I am not asking or recommending you to do this in the present study but I am just highlighting the potential value of such an exercise in future work.

27. L422: I am fully in agreement. This further emphasizes the importance of reporting the p-values themselves rather than a binary significant/non-significant result.

28. L437: I agree that this could be an important step for future studies. As mentioned previously, I would also add (1) making full use of the uncertainty-aware ensemble of simulations (rather than just a point estimate from the ensemble mean) and (2) investigating the combined effects of calibration (also of p_f) and forcing data choice on the forecasts.

I would like to congratulate van der Laan et al. on their pioneering large scale glacier mass balance re-forecasting study and this well written manuscript which was a pleasure to read.
Kind regards,
Kristoffer Aalstad

References

Ambaum, M.: Significance Tests in Climate Science, J. Climate, https://doi.org/10.1175/2010JCLI3746.1, 2010.

Bennett, N. D. et al.: Characterising performance of environmental models, Environmental Modelling & Software, 40, 1–20, https://doi.org/https://doi.org/10.1016/j.envsoft.2012.09.011, 2013.

Fang, Y. et al.: Spatiotemporal snow water storage uncertainty in the midlatitude American Cordillera, TC, https://doi.org/10.5194/tc-17-5175-2023, 2023.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecasting, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

McShane, B. et al.: Abandon Statistical Significance, The American Statistician, https://doi.org/https://doi.org/10.1080/00031305.2018.1527253, 2019.

Pauwels, V. R. and oters: Evaluating model results in scatter plots: A critique, Ecological Modelling, https://doi.org/10.1016/j.ecolmodel.2019.108802, 2019.