Author response to reviews of “Decadal re-forecasts of glacier climatic mass balance” by van der Laan et al.

Dear editor and reviewer 2,

We would like to thank you for the thorough review and detailed comments on this second round of reviews. We appreciate the need for clarification and hope to offer it in the revised version. We propose to make changes to the manuscript in accordance with our replies below, in blue.

Thank you again for your time, kind regards,

Larissa van der Laan, on behalf of the author team

I have read the revised manuscript and appreciate the author’s efforts to address many of the comments raised by myself and the other reviewer. However, I do think further clarifications are needed, especially around the statistical metrics. The level of significance (statistical and in terms of general impact) is still difficult to assess, partly because some of the metrics reported are unclear, and partly because I see a mismatch between the differences in skill shown, and the conclusions that the authors make. I am not totally convinced that reforecasts are as promising as the authors argue in some places. I’ll grant that there could be reasonable differences of opinion on how promising one views reforecasts in light of these statistical results. But given that many results are found to statistically insignificant or marginal, at the very least the metrics for skill and the statistical tests used to evaluate them need to be much more clearly described, so the reader can make an informed interpretation of the results.

I elaborate on these major points of clarification below, and have some minor comments as well.

We appreciate this summary and will address both the major and minor comments below.

Major comments

1.) You say that you do these t-tests for individual glaciers. The wording is somewhat ambiguous – for a given comparison, are you doing

(A) 279 different t-tests (1 for each glacier)?

Or (B) a t-test based on the whole set of individual glaciers?

Wherever described (line 290, 304, 416), the wording seems to suggest (A), but then what are the samples being compared for each glacier’s decadal mean mass balance? The 21-member ensembles? (But then how is this done for the persistence forecast which is a single member?). (B) seems to make more sense to me as a way to test the significance of the difference between approaches, so maybe I am misinterpreting the wording. But if it is (B) then some discussion would be warranted as to why the differences are not statistically significant

with a t-test, but are with a binomial test. I get that different statistical tests address different things, but then some explanation is needed for what these differences mean. Either way this should be clarified.

You are right, this is confusingly worded. We indeed did B), a single t-test comparing the mean performance across all 279 glaciers (ensemble means for the re-forecasts and historical GCMs, and the single member for persistence). We acknowledge that our original wording was ambiguous and have clarified this explicitly: "The samples for these tests are the glacier-specific values of decadal mean and cumulative mass balance for each forcing type. For decadal re-forecasts and GCM Historical experiments, these values are ensemble means, whereas for the persistence forecast, it is a single-member simulation.", see line 289.

The differences between the tests are discussed below, in comment 2.

2.) When discussing the binomial test, which metric is used to assess improved skill? MAE, ME, or correlation? If multiple, how are they combined? This is especially important to clarify as in table 2, while MAE and correlation indicate improved skill for reforecasts over the GCM approach, the model error (ME) metric is smaller in magnitude for the GCM Historical category. So it would seem important to comment on these differences across skill metrics, and what this means for the binomial test of improved skill.

We agree this needs clarification. The binomial test was based explicitly on the MAE metric only, as it directly measures the magnitude of errors and is most appropriate for evaluating forecast improvement across glaciers. ME and correlation were reported for completeness but not used for the binomial test. The clarification "We also perform a binomial test, assessing improved skill by specifically evaluating reductions in mean absolute error for the decadal re-forecast relative to persistence or historical forcing. This shows that out of the 279 glaciers we analyze…" can be found in line 293.

The differences between the tests and their interpretation is clarified as follows, see line 298:

"We note that the t-test and binomial test can yield different interpretations because they test different aspects of statistical significance. The t-test compares the mean differences between forecast methods across the entire set of glaciers, assessing whether differences in the average performance are statistically distinguishable given the spread in the dataset. By contrast, the binomial test assesses significance based only on the count of glaciers showing improved performance (in terms of lower mean absolute error) when comparing one forecast method directly against another, irrespective of the magnitude of improvement.

Thus, while the t-test might indicate no statistically significant difference due to variability across glaciers and the relatively small magnitude of improvement on average, the binomial test can indicate statistical significance simply because more glaciers improve under one approach compared to another than would be expected by chance alone. Both tests provide complementary perspectives: the t-test assesses the robustness of the average improvement magnitude across glaciers, whereas the binomial test evaluates consistency in improvement across individual glaciers."

3.) More generally, I find some of the conclusion statements about using initialized projections somewhat misleading or at odds with the magnitude/significance of improvements being shown. For instance, the abstract concludes with:

"These findings highlight the operational feasibility and significant potential of decadal predictions in glacier modeling for hydrological applications, particularly in regions where near-term forecasts can inform water resource management and climate adaptation strategies."

That is a strong statement to make for difference that in most cases are statistically insignificant. I think the small magnitude of the improvements should at least be noted in the abstract for transparency, and maybe the statistical testing briefly summarized (beyond the binomial result).

And later, at line 398: "While this has not translated into marked improvement for mass balance prediction in the current study, decadal forecasts' narrower ensemble spreads, because of their constraining the initial conditions of key climate variables, may reduce future uncertainty in near-term predictions critical for glacier response modeling"

This statement seems self-contradictory – why should we expect they may reduce future uncertainty, if there are not marked improvements shown here? I can't tell what is backing up this statement.

We do stand by this point, as the narrower ensemble spreads of decadal forecasts vs. long-term climate projections are one of the main strengths in terms of uncertainty reduction. We have clarified that this 'narrower spread' refers to decadal forecasts vs. projections, see line 409: "While we do not observe substantial improvement in mass balance predictions in this study, initialized decadal forecasts inherently provide narrower ensemble spreads due to their constrained initial conditions. Future improvements in initialization techniques, e.g. regarding the initialization of the North Atlantic Oscillation (Nicoli et al., 2025), may therefore still offer potential for reducing uncertainties in near-term glacier modeling, even if the current benefits are limited."

Overall, I think the authors should consider whether these conclusions are really warranted given the significance of results shown, and whether some of these statements ought to be adjusted.

We acknowledge the reviewer's overall point. Given the limited magnitude of the improvement and statistical significance, we've adjusted the abstract and conclusions to more clearly reflect the modest nature of the improvements, highlighting clearly the small absolute magnitudes and statistical significance.

The abstract conclusion has also been adjusted to: "These findings demonstrate moderate improvements from using decadal re-forecasts, though statistical significance is limited. While improvements are modest, these results suggest decadal re-forecasts may offer potential for improved near-term glacier predictions relevant for hydrological applications, particularly in regions where near-term forecasts can inform water resource management and climate adaptation strategies."

4.) I am not convinced by the argument at the end of section 3 that the improvements in skill may be greater than suggested by the t-tests. I do see the point – that the improvement is diluted by the shared skill coming from the global warming trend in both GCM and reforecast forcings. But that doesn't mean the actual skill improvements are larger, just potentially less likely to be spurious. At least reword to clarify. But if you think this test is not properly indicating the statistical significance, is there a better solution? Can the amount of skill coming from the global warming trend at least be quantified? I am not sure what the reader is supposed to do with this argument – it seems speculative to introduce this issue without doing anything about it.

We agree the wording was overly speculative. We've clarified that the issue is that statistical tests might underestimate the practical relevance (not the size) of improvements, rather than actual skill being greater than measured. Clarification in the text, see line 435:

"While this does not imply larger actual improvements, it does indicate that even modest skill increases from initialization may be reliably attributed to improved forecast initialization rather than random chance."

Minor comments:

There is one comment from my earlier review that I'm not sure was addressed. In sec 2.1, it is noted that final outputs are ensemble means. Isn't much of the absolute error in comparing observations to reforecasts coming from the observation (a single timeseries) having more interannual variability than the ensemble mean? And for this reason, if the output for persistence experiments is a single time series, again with more variability than the ensemble means for GCM and re-forecast experiments, how are we to compare the absolute error metrics these cases?

(The first response to this comment simply said it would be discussed but I did not see where it was discussed. I may have missed it, but please check)

This was indeed insufficiently addressed in the previous revision. The observational time series indeed contains greater interannual variability than the smoothed ensemble mean, with averaging multiple realizations naturally reduces variability by filtering. We acknowledge that this limitation somewhat favors ensemble-mean forecasts when evaluated via MAE compared to single-member persistence forecasts. However, ensemble means are commonly used in practice precisely because they aim to provide the most reliable prediction by averaging out unpredictable internal variability. Thus, comparing ensemble means to observations remains scientifically meaningful and justified, provided readers are clearly informed about this limitation. We hope to do this by adding the following beginning to the results and discussion section, see line 243:

"This section evaluates the results of the three experiments - decadal re-forecast, persistence, and GCM historical - by comparing them against observed glacier mass balance data for individual years and decadal averages. Throughout our analysis, we compare observational time series which inherently contain interannual variability to ensemble means from multi-member forecasts. It is important to recognize that ensemble means naturally reduce variability by averaging across multiple realizations, thereby smoothing internal fluctuations that occur in any single realization or observed series. This reduction in variability means that some portion of the absolute errors (e.g., the mean absolute error) used in the evaluation arises from the difference in variability between a single observational realization and the ensemble mean. Despite this limitation, ensemble means represent the most commonly used forecast product in practical applications due to their stability and reliability."

56 – what is "large scale glaciology"? I have a sense of what you mean, but not sure that's widely understood.

Clarified as "large-scale (regional to global) glacier modeling."

162 – specify – initial state of the climate/atmosphere? (As opposed to glacier)

Specified as "initial state of the climate system (atmosphere and ocean)"

181 – does a decreasing ensemble size have effects on the error metrics?

Explicity stated now that "Decreasing ensemble size at longer lead times typically increases forecast uncertainty, potentially affecting skill metrics slightly."

270 and Figure 2 – I still don't fully get what is added by comparing the 11 different overlapping decades. If those are getting averaged together in table 2, I think you would get the same #s by just averaging over 2000-2020, you are not adding any information by carving into 11 overlapping segments. And I still find the clusters of points in Figure 2 ambiguous to interpret since the individual points share a lot of information as overlapping decades. If it is just to demonstrate the effects of sampling error that is fine and useful, but not sure Figure 2 is needed for that.

We still believe it is important to show the overlapping decades, to illustrate variability in skill arising purely from changes in start and end years. These results highlight the sensitivity of decadal skill metrics to specific evaluation periods.
But figure 2 could in principle be moved to supplementary materials.

295 – why is the skill improvement different for decadal mean vs. cumulative? It seems odd that some glaciers would have improved skill for one and not the other. But as noted above, I'm not even sure which skill metric is used for the binomial test.

When presenting tables 2 and 3, are the skill metrics averaged across all glaciers? And across different decadal segments? Would be helpful to clarify.

This is indeed important to clarify, as we have done in line 272 : "The mean and cumulative decadal mass balances are distinct from each other as the cumulative mass balances only include the simulated years for which observations exist."

The skill metrics are indeed averaged across all glaciers and decadal segments.

382 – what counds as "good precipitation skill"?

Good precipitation skill refers to statistically significant correlation or low MAE between forecasted and observed precipitation in validation studies, such as Delgado-Torres et al. (2022)."

396 – I don't think there is any "predicting internal variability" – yes, there is still from starting with the right sampling of internal variability and capturing any climatic memory of that initial state… but I don't think predicting is the right word. And what is meant by regions where there is an externally forced response? I find this sentence confusing.

We hope this is more clear: "initialized forecasts better represent internal climate variability due to accurate initial conditions." See line 409

413 – I don't understand this sentence. What is the probabilistic context? And what are pathways? Ensemble members, or scenarios?

Probabilistic context refers to uncertainty arising from different socioeconomic emission scenarios (pathways) used in uninitialized projections.

419 – "this common signal…" I think this sentence is too similar to the sentence in the Smith et al., 2019 reference (p. 3). Either paraphrase or make a direct quote.

You're right, and we have indeed used a direct quote now.

430 – what counts as skillful here?

Skillful here is a summarizing word of the results presented in the manuscript, including the discussions of the error metrics.

446 – "rather than continuation of interdecadal trends" – this seems different than the persistence forecast, which is just repeating the same decade. Can this be clarified?

Yes, this is now clarified: "rather than depending solely on the continuation or repetition of recent decadal climatic conditions (as in persistence forecasts)." (Line 466)

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.

Dear editor and reviewer 1,

We would like to sincerely thank you for your detailed, thoughtful review and your constructive comments. They have greatly contributed to improving the manuscript's clarity and rigor and given us ideas for future studies. We have carefully considered each suggestion and made adjustments in the manuscript as detailed in our responses below (in blue).

Thank you again for your valuable feedback and insights.

Kind regards,

Larissa van der Laan, on behalf of the author team

Specific comments

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.

Agreed and revised to "seasonal forecasts and long-term projections" to improve clarity

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.

We changed this to 'early' rather than 'developmental', to reflect the early stage of the field

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.

Agreed, done

Correct Eq. 1 to

$$mi(z) = pfPi\text{solid} - \mu * \max(Ti(z) - T\text{melt},0) + \epsilon \qquad (1)$$

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.

Done, and we thank the reviewer for being so considerate to add these precise suggestions to their review.

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

Done

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

Done

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.

We appreciate this comment and will take this convention into account for future manuscripts. For this case, we have ultimately decided to leave the figure as-is, as we do not feel the adjustments fundamentally change the figure's functionality.

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)?

We acknowledge this point regarding uncertainty quantification inherent in ensembles. We have expanded the

discussion to explicitly acknowledge that focusing solely on ensemble means indeed disregards uncertainty. We also in part address this in our response to reviewer 2, and have added a clarifying paragraph to the results/discussion section, see line 247. The outlook also proposes future use of probabilistic scores, see line 454.

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

Done

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

Done

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.

We fully agree that model calibration and forcing product selection are interdependent. We have added a discussion to explicitly acknowledge that the calibration parameters are inherently conditional on the forcing product used and clearly highlight the value of exploring their combined effects in future research. See lin 400. That being said, this effect will in part already be reduced by the fact that we downscale the climate products to the baseline climate (CRU).

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 (glacierby-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.

We do include a residual term for individual glaciers in component 1, but not in component 2, where individual glacier data allow for calibration without residuals: "[…] since this dataset consists of data for each glacier, removing the need for parameter transfer to glaciers without observations." (line 141) We also include a citation to Marzeion et al. (2012), which goes into more detail on the use of the residual.

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.

Clarified "perfect results" explicitly as exactly matching the geodetic mass balance observations used for calibration. "This means that when run with the baseline climate CRU, it provides 'perfect results': exactly matching observations over the calibration period (bias of zero)." (line 148)

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.

Agreed, done

—

L241: Fix the LATEX notation here, I guess something went wrong and $T't$ is supposed to be T_t' or similar.

True, thank you. Changed

L242: Change t to *t* (i.e. t) for consistency.

Done

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).

Done

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.

This is a good idea

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.

True, and explained in 272. It is only explicitly mentioned because it is used in the literature cited in our work.

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?

Clarified, see response to reviewer 2

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).

Clarified, see response to reviewer 2. Thank you for the literature suggestions! Will take note of this for future publications.

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

Done

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.

Thank you for the suggestions and explanation, we will definitely keep this in mind for future work!

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.

We fully agree with you, and this would be very interesting to include in future work. We appreciate your thinking and sharing these thoughts and ideas with us. It is now mentioned in the discussion as a potential future topic of research in this area. (see line 400)

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

Done

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.

Agreed!

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.

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.

Thanks for these thoughtful points

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

Thank you for this detailed, comprehensive and immensely constructive review.

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.