**Review of "Decadal re-forecasts of glacier climatic mass balance" by *van der Laan et al*.**

This work aims to assess the applicability of forcing a glacier mass balance model with decadal re-forecasts that could help bridge the gap between seasonal and centurialmillennial timescales. The authors use the mass balance scheme of the Open Global Glacier Model and conduct three experiments (1) persistence runs using CRU forcing, (2) GCM historical runs with a 21-member CMIP6 ensemble, and (3) decadal reforecast experiment using the same ensemble from the Decadal Climate Prediction Project (DCPP). They conclude that decadal re-forecasts have similar or better fit to the observed mass balance as compared to the other two experiments.

Overall, this is a detailed assessment with adequately designed experiments to address the study objective. I have two major comments regarding the clarity of the methods and the presentation of the results. The manuscript will benefit from a restructuring of the Methods section for better understanding of the experiment design and the specifics of each of the three experiments performed. Second, the Results and Discussion section requires improvement as it currently lacks rigor in the presentation of the statistical analysis and the discussion of the results.

I have separated the major comments for the Methods and Results/Discussion section below, followed by a few minor comments. Hopefully, this will help the authors to revise and resubmit the manuscript.

Dear editor and Reviewer 1,

Our sincere thanks for the thorough review and detailed comments. We will make several revisions to address the concerns raised. Below, we provide detailed responses to each of the comments. Our replies to the comments are in blue.

Thank you again for your time, kind regards,
Larissa van der Laan, on behalf of the author team

## 1) Major comments:

### Methods:

- Section 2.2: The application and purpose of the two calibration approaches requires better explanation. Why is there a need to calibrate with WGMS data for the 279 glaciers only and how does that feed into the experiments? Why not just use the calibration of ~214000 glaciers with the geodetic MB for the experiments?

We understand why this can be confusing. In part, this has to do with the trajectory of our experiments and the development of OGGM. We began this study with the WGMS glaciers only, and the OGGM default calibration as of v. 1.5.3, which is to

calibrate the mass balance model with the WGMS glaciers. The global analysis was added later, as the global geodetic dataset became available for calibration and validation. Using both approaches gives us additional insights: WGMS MB data is relatively sparse, though has a yearly resolution and has data preceding the global geodetic data, while the geodetic data has a lower temporal resolution, but a global coverage. The two datasets are also not perfect and do not necessarily agree at the glacier scale. Therefore, the calibration with WGMS data ensures that we have the best available model for each glacier individually, to focus on the impact of each forcing dataset, and then similarly for the global analysis.

- Why does one calibration approach have ε and the other does not – are they both not done for individual glaciers?
- For 2.2.1, how are the two unknowns (μ* and ε) established with one equation?

In older OGGM versions, ε was introduced to apply a calibration where no WGMS data was available. For the purpose of our study, where we calibrate only for glaciers with data, ε is always close to 0. We refer to Marzeion et al. (2012) and Maussion et al. (2019) Sect. 3.3. for an in-depth discussion of the calibration procedure, but it must be noted that for the purpose of our study, the performance of the mass-balance model itself is only secondary, since only the change in performance when using various forcing products is investigated.

- Ln 115: What is the re-calibration step here that is done for the global run?

In essence, the addition of another analysis, global this time. This was awkwardly worded and will be changed to: “Note that the separate calibration for the global run…”

- Was the calibration with geodetic data also done with CRU (similar to the WGMS calibration)?

Yes

- Ln 116: Does this mean that once μ* and ε are established for each glacier (using CRU), the same values will be used for all experiments?

Yes, the parameters are held constant for each forcing product, allowing us to assess the impact of the forcing strategy alone, not the impact of calibration.

- 2.2.2 calibration was done over the 2000-2020 period, what about the 2.2.1 calibration?

  The calibration here was done over the years with observed data for the WGMS glaciers that fall within the CRU climate data period (1901-2020).

- Section 2.2.3 needs restructuring for clarification of the experiment design. For example, information in Ln 123 – 130 can be merged with the individual experiment information. It seems Ln 123 – 126 is describing the persistence experiment?

- The manuscript talks about two components (e.g., Ln 123 and 127) and three experiments. It seems the two components refer to the two calibration approaches? Later, the results are separated for Reference and Global glaciers, and this was not clear in the objectives (Introduction) or the methods section. Ln 109 mentions about the 'global component of the study', but these components are never defined.
- The model run years require better explanation as well: the simulations are done over 1990 – 2020 period; Ln 117 states that "we will always run the model during the period it has been calibrated for"; Ln 112 states that the calibration with the geodetic MB is done for the 2000 – 2020 period; Ln 130 states that the validation is carried over the 2000 – 2020 period (calibration and validation here are supposedly used interchangeably?) – all this requires a clearer description.

We acknowledge the need for restructuring this section to improve clarity. The two components mentioned refer to the two sets of analyses/approaches — the reference glaciers and the global glaciers. The three experiments (persistence, GCM historical, and decadal re-forecast) were applied separately to these two components.

The manuscript will be revised to clearly define these components in the Introduction and Methods sections, with explicit mention of how they relate to the experimental design. Furthermore, the timeline for model runs, calibration, and validation periods will be clarified to avoid any confusion regarding the simulation years. The objectives in the Introduction will be updated to reflect these clarifications.

- I understand the authors created the separate sections on Experiments (2.2.3), Lead Times (2.3), and Climate Data (2.4) for clarity, but it made following the methods somewhat cumbersome. I recommend merging all the information on the three experiments under the experiment design section, including what climate data was used and how the lead times were defined. And then, summarize this information in a table.

Thank you for this suggestion. We tried various solutions for restructuring but concluded that the original structure still suits the logical flow best. We will however amend the text, making it more concise and hopefully more understandable.

**Results and Discussion:**

- Ln 232: Where is this N = 2676 coming from? The WGMS calibration approach has N = 279 and I presume calibration with geodetic MB has N = 214,000 (Ln 129)? I think these first few lines should be in Methods rather than results.

  I am also confused regarding the Fig.1 caption mentioning N = 279 reference glaciers for getting the forecast skill and then in the next sentence stating N = 2676 for *r* and MAE.

The number N = 2676 refers to the total number of annual observations across the 279 WGMS reference glaciers, not the number of glaciers themselves. This distinction was not clearly communicated.

We will move the explanation of N = 2676 to the Methods section and clarify that this number represents the total number of annual observations rather than the number of glaciers.

- Ln 235 – 245: I understand the authors want to share these results to highlight that year-to-year prediction is not practical with the current modelling scheme and is not the objective of the manuscript. But this paragraph is putting too much emphasis on the statistics without providing much context. For example, what does a MAE of 0.6 or 0.7 m w.e. mean, is this too high or too low? What are the annual mass balance magnitudes in general? Perhaps a metric like Mean Absolute Percentage Error (MAPE) will be more informative here.

  In general, I think the authors can remove this paragraph and Fig. 1 altogether and keep the focus on decadal timescales only (please see a minor comment as well regarding the definition of decadal vs annual timescales).

  We understand the concern that the emphasis on single-year predictions may detract from the main focus on decadal timescales. However, we do think this section provides important context on lead time and its importance in decadal-scale forecasting. In accordance with comments by reviewer 2, we have also added background information on lead time and decadal scale forecasting in general. We are unsure about how to apply MAPE in the context of mass-balance, which can have positive, zero, or negative values. We will change the text however to emphasize the relative differences between the various skill values, and discuss these values in the context of observational uncertainty for context.

- Ln 254: This statement needs better qualification; how is a model error threshold of <0.2 m w.e. established? Is this statement referring to the first row (ME) of Table 2? I suggest the authors establish more rigor in defining the statistical thresholds for good and bad results based on observed MB estimates and physical explanations. I am struggling to understand whether an error is too small, too large, or just right for the arbitrarily selected N = 279 (or 2676) glaciers.

We agree that this needs more rigor. The 279 glaciers, however, are not selected arbitrarily, they are the WGMS glaciers with observations over a time period longer than 5 consecutive years and are land-terminating. This information will now be included in the manuscript.
The <0.2 m w.e. threshold is indeed arbitrary and will be removed.

- Ln 265: I am not sure I understand this correctly, the decadal forecast MAE (0.29 m w.e.) is 7% larger than GCM Historical MAE (0.27 m w.e.), but this sentence suggests there is a 7% reduction in decadal as compared to GCM forecast.

Thank you for pointing out this error! This was mixed up in the table and should be 0.27 m w.e. For the decadal re-forecast experiment, not the GCM historical experiment.

> Are these differences significant, not just statistically but also in general terms (e.g., would these differences affect global or regional scale assessments to understand glacial mass loss or melt rates, etc.). This is important because the GCM Historical forecast metrics are similar to the decadal re-forecast ones in most cases, sometimes with slightly better results as well.
>
> For this entire paragraph, it would be helpful to understand the meaning behind the mean and cumulative mass balances and the ME and MAE, and where these differences are coming from for these select glaciers.

We hope that the following context may provide some clarity:

On average, the 279 WGMS glaciers lost 0.79 m w.e. during the decade 2000-2010. As a control against an ideal case, and to gauge the magnitude of errors, OGGM is also forced with reanalysis dataset CRU and run for the 279 WGMS glaciers. This data is also used in calibration and to create the persistence forecast. Comparing against observed WGMS data of mean mass balance per decade, the errors +- the half interquartile range are as follows:
Mean error -0.037 +- 0.16, mean absolute error 0.23 +- 0.17 and the Pearson correlation 0.72 +- 0.11.

Comparing to the experiment errors in table 2, this means the errors for both the decadal re-forecast and GCM historical experiment are within the order of magnitude of the error when forcing OGGM with 'ideal' CRU data. The forcing with CRU reanalysis data does, as expected, lead to better results than forcing with a forecast or projection. For all glaciers, in a binomial test, results when forced with CRU are closer to observed mean mass balance than in the forcing experiments.

The order of magnitude however, gives confidence in the skill of either experiment's forcing. For cumulative mass balance, the error statistics for the decadal re-forecast and GCM historical experiment are also within 10% of the error when forcing OGGM with CRU data.

To gauge the significance of experiment differences from a statistical point of view, we carry out a two-tailed t-test (significance level 0.05). For the decadal mean mass balance, the difference between the Decadal RF and Persistence experiment is statistically significant, as is the difference between Persistence and GCM Historical experiment. However, there is no significant difference between Decadal RF and GCM Historical experiments. For the cumulative mean mass balance, there are no statistically significant differences between the experiments.

In accordance with our response to reviewer 2, this emphasizes that the improvement in skill is notable but not necessarily significant. We will amend the text to make sure that our conclusions reflect these results accurately.

- Ln 275: Which figure or table are these results referring to?

Multi-model ensemble results on average show higher skill than single-model realizations. We will refer to a citation in the manuscript to make it clear that this is not one of our results.

- Ln 295 onward: It would significantly improve the narrative if the authors were to dissect the regional differences and provide a better explanation of where the "considerable variation in skill" is coming from. These results (Fig. 4, Tables 4 and 5) are the more interesting part of this study but the presentation of the results and discussion here is somewhat deficient (the text repeats the statistics in the tables, but their meaning and importance is not explained).

We appreciate the suggestion to delve deeper into the regional differences. These variations are indeed significant and warrant a thorough discussion. However, as noted in a response to reviewer 2, we aim to discuss the added value of using decadal re-forecasts, rather than their inherent skill at simulating (regional) climate. We will make clear that we think the regional SMB differences stem from differences in skill predicting temperature and precipitation in the respective regions, and refer to the relevant literature discussing this skill and its sources.

- Ln 300: Earlier a threshold of <0.2 m w.e. was used for 'good' results. These thresholds should be consistent across the analysis (and perhaps specified earlier in the methods section on how the metrics were established).

The earlier (quite arbitrary) threshold has been removed, so the thresholds are consistent throughout the manuscript now.

Also, the errors in the geodetic MB from *Hugonnet et al.* needs to be accounted for. For example, in Table 4, Region 10 has a MB of -0.38 ± 0.58. Why is -0.42 considered a good fit but -0.27 a reasonable fit based simply on the mean MB value?

We will amend the color coding/ goodness of fit criteria to reflect the Hugonnet errors. Thank you for this idea.

- Ln 323: Where are these results shown? In fact, shouldn't these be the main results to ensure that the three experiments are comparable by design and the results are not affected by the calibration/validation periods.

We agree that these should be the results shown for the persistence experiment instead. This will be changed in the manuscript, with the explanation of the calibration period.

**2) Minor comments:**

Ln 23: *"...glaciers were the largest contributor to sea-level rise..."* Is this specifically referring to glaciers outside the polar regions, in continuation to Ln 20? Can you please cite this.

The statement refers specifically to glaciers outside the ice sheets. We will clarify this.

Ln 38: It is best to keep the terms consistent. It does not make sense to use "decadal prediction" or decadal timescales for single years or durations <10-years.

This, although confusing at times, is necessary to remain consistent throughout the manuscript, since our decadal re-forecasts are e.g. clipped to hydrological years, hence not 10 full years. It is only clarified so explicitly here to avoid confusion later on.

Ln 47: In applications of?

Thank you for spotting this error. This should read “[...] into the application of decadal forecasts.”

Ln 55: The common time scales here are referring to centuries and millennia?

Yes, which we have also clarified.

Ln 57: What are “impact models”?

Impact models refer to models that assess the consequences of climate change on natural and human systems, such as glacier runoff or agricultural productivity (e.g. ISIMIP paper).

Ln 94: Can you please provide a justification for why the precipitation correction factor is set to 2.5 for all glaciers globally and for all forcing data sources? The *Maussion et al. 2019* citation alone is not adequate. Does this affect the MB computations for persistence experiments (using CRU) vs GCM historical or decadal RF experiments?

The precipitation factor is computed for historical data (here, CRU) by minimizing the error in variance of the mass-balance for all 279 WGMS glaciers. For another historical dataset (e.g. ERA5), the precipitation factor would be different indeed. The forcing climate datasets however (re-forecasts, historical GCMs, etc.) are then bias corrected to the historical data and therefore have a glacier specific correction depending on the bias correction method used for each product (re-forecasts or GCMs) according to practices commonly used in the large scale modeling literature.

Ln 101: What is the first component of the study? This was mentioned earlier in Ln 67 as well which needs clarification. The last paragraph of the Introduction can benefit from explicit enumeration of the objectives and the “components” of the study.

This will be clarified

Ln 106 – 107: Can you please clarify and rephrase this statement (on *“…parameters do not need to be transferred … and are therefore well constrained”*).

This was unclear indeed and refers to our reply to the comment above regarding the mu and epsilon parameter. Before the availability of global geodetic observations, parameters needed to be transferred to glaciers without any observation (Zekollari et al., 2024), leading to substantial errors. In our case, we apply the model to glaciers with either in-situ (WGMS) or geodetic observations, meaning that the MB model is calibrated to match observations. The statement “well constrained” however was not correct because of equifinality (e.g. Schuster et al., 2023). This sentence will be revised to convey the intended meaning: that the MB model is calibrated to match observations over the calibration period.

Ln 110: 94% of the RGIv6 glacier count?

Yes

Ln 133: “*All different realizations are downscaled to the glacier scale…*” What does this downscaling to glacier scale mean?

This is explained in section 2.4, and we will make sure the text references this section.

Ln 148: It is best to call it the persistence experiment only and not introduce a new term for this (i.e., naïve forecast).

We have included this term because it may be more familiar to readers and give added context to the term persistence.

Ln 240: What does remarkably consistent mean? These are just statistical results, so it is best not to use such superlatives.

We agree this can be misleading. We’ve revised the manuscript to avoid superlatives.

Ln 258: Can you clarify what the ten-year lag of warming means.

The "ten-year lag of warming" refers to the delay in temperature increases observed in persistence forecasts compared to actual observations.

Ln 289: Please rephrase "*slight but clearly noticeable*". In a tabular form, a difference in the third decimal place will also be clearly noticeable.

We will rephrase this

Is Fig. 4 for 2000 – 2010 period?

Yes

References:

Marzeion, B., Jarosch, A. H., & Hofer, M. (2012). Past and future sea-level change from the surface mass balance of glaciers. *The Cryosphere*, *6*(6), 1295-1322.

Schuster, L., Rounce, D. R., & Maussion, F. (2023). Glacier projections sensitivity to temperature-index model choices and calibration strategies. *Annals of Glaciology*, 1-16.

Zekollari, H., Huss, M., Schuster, L., Maussion, F., Rounce, D. R., Aguayo, R., ... & Farinotti, D. (2024). 21 st century global glacier evolution under CMIP6 scenarios and the role of glacier-specific observations. *EGUsphere*, *2024*, 1-33.