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.

Author response to reviews of "Decadal re-forecasts of glacier climatic mass balance" by van der Laan et al.
(https://doi.org/10.5194/egusphere-2024-387)

Dear editor and reviewer 2,
We would like to thank you for the thorough review and detailed comments. This has been a great help in understanding where our manuscript needs improvement. 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

This paper presents an analysis of glacier mass balance forecasts on multi-year to decade timeframes, using the mass-balance module of the Open Global Glacier Model. The authors compare forecasts made with climate forcing from Global climate model historical simulations, observation-initialized reforecasts from decadal prediction models, and a simple persistence forecast. Comparing their simulated mass balance re-forecasts to both in-situ and geodetic glacier mass balance observations, they conclude that using the initialized climate reforecasts provides an improvement in skill over GCM-based forcings. The predictability of glacier mass balance on short timeframes is an important problem, and the overall approach of comparing these forcing strategies using the mass balance model and assessing them against observations seems sound. However, there are some significant issues with the clarity of methods and results. It is not clear to me

that the authors have demonstrated a meaningful difference in skill between reforecasts and GCM-based forcings. My main comments are detailed below, with some additional minor comments later on.

We thank the reviewer for this concise summary and evaluation of our work. We hope the answers below and the edits of our manuscript will improve the clarity.

Major comments:
1) Significance of skill improvements
The principal finding is that reforecasts provide an improvement in skill over using GCM output as forcing, and they emphasize the improvements for decadal mean and cumulative balance (since the skill for yearly forecasts is low; e.g., Fig 1 and line 244). However, there appears to be a huge spread in the Mean Absolute Error (MAE) metrics, such that the 1-sigma ranges substantially overlap. For example, in table 2 for decadal means, the authors report MAE of 0.29 +/- 0.32 mw.e. using reforecasts, and 0.27 +/- 0.31 mw.e. using GCMs. Overlaps are even greater for cumulative balance: 1.33 +/- 3.21 and 1.58 +/- 2.96. Presumably the standard deviations correspond to the distribution of errors across the individual WGMS glaciers. It's hard to see how this decrease in MAE is significant, given such wide distributions. The authors do not really comment on the wide spread of these error statistics. At the very least, they need to be discussed and the overall conclusions put in the context of these wide distributions.

We acknowledge the large spread in the error metrics and that this means it may be difficult to demonstrate significance for individual glaciers. To make the overall improvement more evident, we performed a binomial test. Out of the 279 glaciers we analyze, 174 showed improved skill using decadal re-forecasts for decadal mean mass balance, and 186 showed improved skill for cumulative mass balance. Using a binomial significance, this suggests that the *overall* improvement is significant at the 5% level.

At a more technical level, there are some other issues with reported statistics that I find puzzling and not well explained.

We agree that the explanations for these points were lacking and will add clarifications accordingly.

To return to the MAE metric in table 2, the +/- range in many cases exceeds the central value reported, implying negative values, which don't make sense for Mean Absolute Error which should be positive definite. If the standard deviation of a positive-definite distribution is so large, does that imply a very long tail and some very large errors?

The large standard deviation indeed implies a long tail and very large errors here. Below is an MAE histogram example of the 2000-2010 decade from the GCM Historical experiment, which illustrates the spread of the errors. Especially the few very large errors (above 1 m w.e.), which are also present for the other experiments, affect the standard deviation. In the revised manuscript, we will change all statistics to use quantile ranges instead of mean and standard deviation to reflect that the distributions are not gaussian.

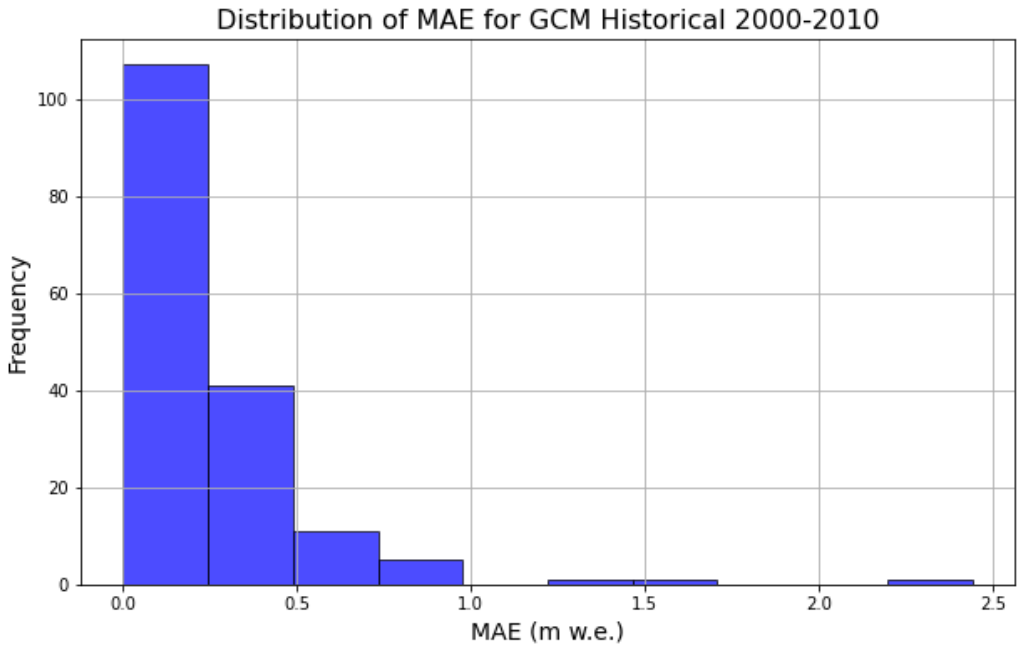


Fig 1. Histogram of the MAE for the GCM Historical experiment 2000-2010

Also in Table 2, what is the Model Error statistic? As far as I can tell this is never explained.

Thank you for alerting us to this. We use "Mean Error". We will add the following text to explain the Mean Error statistic: "To quantify model skill, we look at the mean error (ME), which is the average difference between observed and simulated values (sometimes called "mean bias"), the mean absolute error (MAE) and the Pearson Correlation Coefficient."

At line 250 it is stated that the period 2000-2020 gives 11 full decades. This is technically true in a moving-window sense but these 11, 1-decade windows are not statistically independent samples. Why are these reported as different decades? There is a lot of potentially independent information across different individual glaciers, so why is Fig. 2 plotted in terms of these 11 heavily overlapping windows?

Together, these make it hard to interpret the significance of the overall conclusions.

It is true that there is significant overlap between the decades. Unfortunately, because of the limited availability of decadal re-forecasts, and wanting to assess the re-forecasts in the manner that future forecasts would be used (i.e. forecasts outside the bias correction time frame 1971-2000), we can only use the time period 2000-2020.
However, because within the relatively short window of a decade, a year's difference - or the choice of the specific decade, in other words - can significantly affect results. This is a hallmark of working on this time scale, but we agree this was not sufficiently explained in the manuscript. To illustrate this, we added the following to section 3.1:

The full decades, which have significant overlap, are all compared separately and depicted in Figure 2 to show how the choice of decade can impact skill statistics. For

example, in the Decadal Re-forecast experiment, the decade 2001-2011 has a mean model error of -0.022 m w.e., whereas the decade 2002-2012 has a mean model error of 0.11 m w.e.

2) ReForecast drift correction
I found the explanation of the reforecast bias correction to be confusing. The authors note that the bias correction is lead-time dependent, but do not really explain why (lines 215-16). This would seem to be an important point to explain thoroughly in order to compare GCM to reforecast-based glacier simulations.

This is now explained better and we refer to relevant papers for more in-depth insights and limit our explanation to the following:

Decadal re-forecasts experience a bias which grows with lead time and is referred to as drift (Kharin et al., 2012; Manzanas et al., 2020). The drift is lead time dependent because the model drifts away from the initial state as the prediction progresses, towards a state more consistent with the model's climatology, which can lead to significant error (Pasternack et al., 2021). Re-forecasts are therefore bias corrected to counter this error. Our correction adheres to recommendations in Boer et al. (2016), who recommend an overarching bias correction method, regardless of the initialization type of the forecast. The reasons behind these recommendations are discussed in-depth in e.g. Boer et al. (2016), Kharin et al. (2012) and Hossain et al. (2022).
As explained above, the bias is model dependent and lead time dependent. Because of the assumption that the bias is different at each lead time, subtracting a mean drift over all times would lead to over-compensation at some lead times, and residual drift at others. For this reason, we create lead-time based climatologies for each model. These are then used to create anomalies relative to the baseline climate.

In particular, I can't tell how to interpret the increased skill from reforecasts, in light of the differences in bias correction when using reforecast vs. GCM data to force the model. The reforecast data are bias corrected using different lead-time-based climatologies over 1971--2000. Different lead times aren't considered for the GCM-driven forecasts, so the GCM data have a single bias correction step using CRU TS data from 1961—1990. Are differences in prediction skill (i.e., simulated vs. observed mass balance) related to different bias correction methods, or the fact that reforecasts start from an observed climate field? Either would be useful to know about, but the authors don't address whether the bias correction has an effect.

Thank you for this comment. It is correct that lead time dependent bias correction is a fundamental step used in decadal forecasting, and so is mean bias correction in future projections for impact modeling (e.g. Lange, 2019) . The aim of our study is to compare the standard methodologies commonly applied in the respective fields. The GCM Historical data are processed as they would typically be in glacier modelling (e.g. Zekollari et al., 2024, Rounce et al., 2023, and many more), and the re-forecasts are also processed according to the recommendations mentioned above. It must be added

that the drift correction following lead time (the length of time between the issuance of a forecast and the occurrence of the phenomena that were predicted) cannot be applied to the GCM Historical experiment where we have only one simulation.

We amended the text in several ways:

- Better clarify our goals as outlined above
- Better explain the drift correction, and explain why it can't be applied to the traditionally used GCM simulations
- In the discussion, mention and discuss the fact that drift correction does influence the results by a great amount and partly explains the performance of re-forecasts over traditional GCM-run simulations. The drift correction only impacts average-based metrics but not correlation or interannual skill.

These changes should hopefully help convey the main message of our study: decadal forecasts and the techniques developed to bias correct them should be preferred for medium range (e.g. decadal) predictions of glacier change over using historical simulations.

(also – why correct reforecasts over 1971—2000 and GCM data using 1961—1990 means? No explanation is given)

This is because the re-forecast DCPP project starts from 1961, so for 1961-1970, not all lead times are available. The GCM data is bias corrected as per OGGM standard, which is 1961-1990. To check that this doesn't change the results much, we ran an additional simulation with bias correction over the same period for the WGMS glaciers. A two-tailed t-test (significance level 0.05) reveals there are no statistically significant differences between the results.

3) Background on reforecasts and sources of skill
I think more background on initialized reforecasts is needed, to help the reader understand (i) the product being used to force the MB model, and (ii) where prediction skill might be coming from (if at all). I completely agree that decade timescales are of applied/operational importance, and this is an area worthy of focus. However, I found it puzzling that there is essentially no discussion of internal climate variability which is the main reason that forecasts on multi-annual to decade timeframes are challenging. The initialized climate models used for decadal forecasts are not summarized to much degree, or differentiated from GCMs, except for the fact that they are "initialized", but the authors do not really explain what is meant by "initialized". Again, this is key context when the main result is the relative skill of reforecast vs. GCM-driven mass balance predictions. Some physical reasoning for why an initialized forecast introduces more skill would be important for making sense of the results.

We agree, and hope that the following text and references help in this regard:

Decadal forecasts lie between numerical weather predictions and climate projections. Similar to numerical weather forecasts, they are initialized with current observations of the atmosphere, ocean, land surface and biochemistry, similar to numerical weather forecasting (Doblas-Reyes et al., 2013). The successful application of these methods to longer timescales, e.g. decadal, is an active area of research in operational climate prediction, which is a rapidly evolving field (Meehl et al., 2014; Smith et al., 2013). In particular at decadal timescales, initialization is expected to further improve our ability to

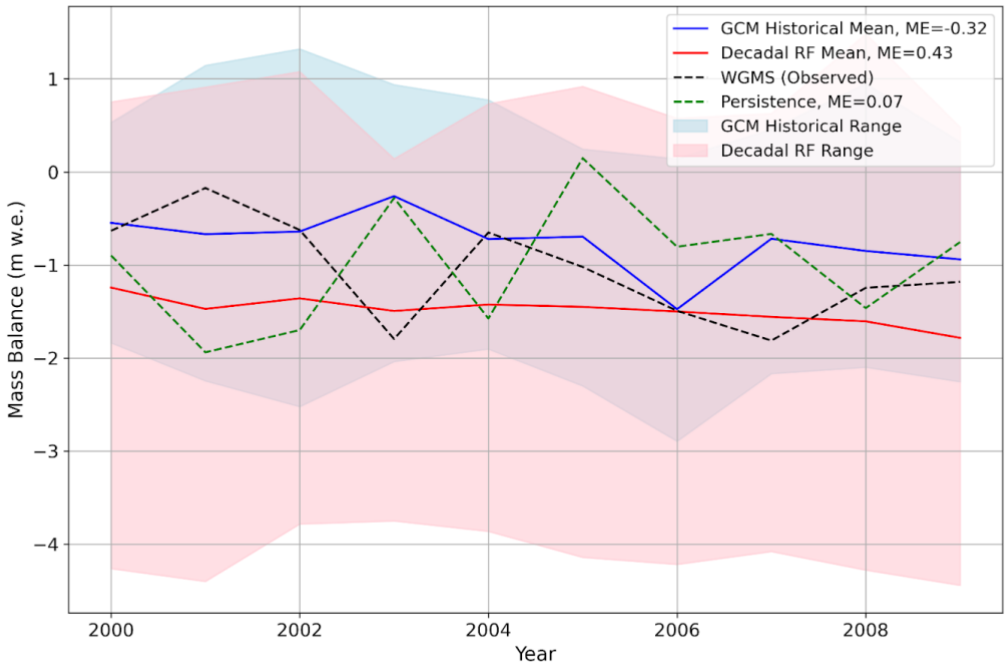
detect the impact of radiative forcing induced trends on the occurrence and frequency of the internal climate modes of variability, such as the El Niño Southern Oscillation (ENSO) (O'Kane et al., 2019). This will contribute to accuracy in forecast temperature and precipitation, not present at the decadal timescale in (uninitialised) GCM projections.
Over the last years, larger initiatives on decadal prediction development have been set up, such as WCRP Grand Challenge on Near Term Climate Prediction (Kushnir et al., 2019). One of these initiatives is the Decadal Climate Prediction Project (DCPP), a coordinated multi-model investigation into decadal climate prediction, predictability, and variability, contributing to CMIP6 (Boer et al., 2016). The Decadal re-forecast experiment makes use of this DCPP output.

We will amend the text accordingly.

Overall: thank you! These are all good points. We hope the background information above provides better context.

4) Visual examples of reforecasts
Finally, I think a figure showing some examples of the mass balance reforecasts would be helpful for understanding the method and results. The figures are largely aggregated statistics. Picking a glacier as a case study, and showing timeseries (perhaps individual members and ensemble means) of reforecasts under GCM vs. initialized forcing would help the reader immensely in understanding what the errors stats reported later would actually look like in terms of a forecast. It is great to draw on the wealth of WGMS data for validation, but I found myself wondering what these results actually look like for a given glacier. How quickly do the initialized reforecasts decorrelate due to internal variability? How do noise and trends compare? One or two examples would go a long way.

This is a very good idea and we have implemented it with an extra figure, also copied below. As can be seen, neither the GCM Historical nor the decadal re-forecast experiments perform very well when analyzed at the single glacier level, in this case the Hintereisferner and Langfjorjoekulen. In the Hintereisferner case, also cumulatively, neither the GCM Historical nor the Decadal re-forecasts have provided more skill than the very simple persistence experiment. The ensemble spread, in this case, is even larger than for the GCM Historical experiment, which does not speak for its benefits of initialization.
In the Langfjordjoekulen case, the decadal re-forecast experiment performs better than the GCM historical and persistence experiments. It does not follow the year-to-year variations, but captures the mean mass balance over the decade much better than the other two experiments.



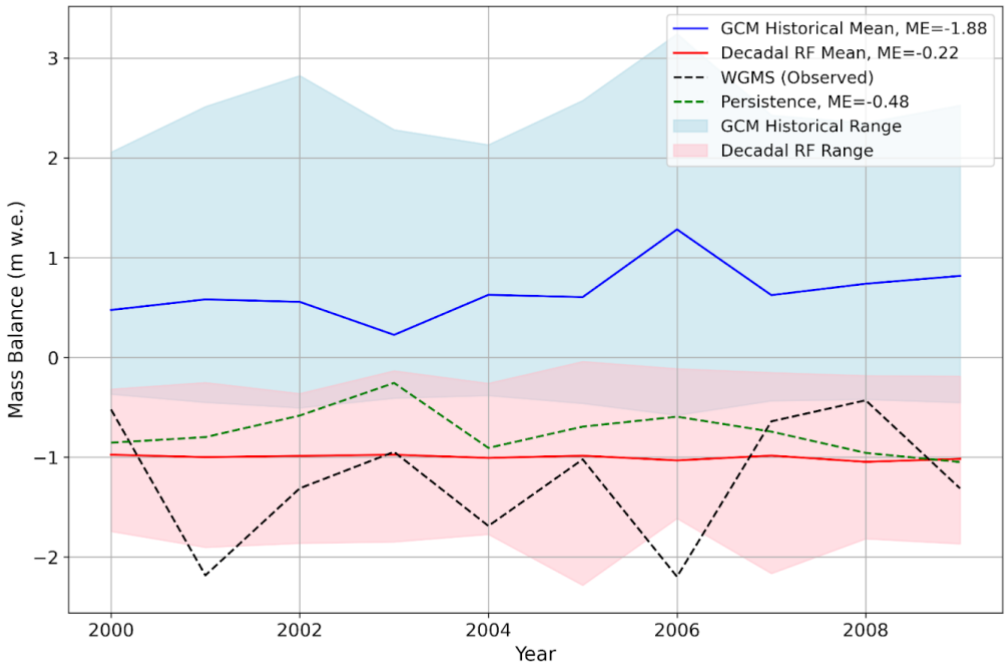


Fig. 3 Example for the Hintereisferner, Austria (above) and Langfjordjøkulen, Norway (below).
Mean errors (ME) for the different experiments, for mean mass balance over the decade, as in Table 2, are indicated in the legend. This means the difference between the mean observed mass balance and the mean simulated mass balance for the different experiments. For the GCM historical and decadal RF experiments, 'simulated' refers to the ensemble mean.

Minor comments

114: What explains the pre-defined range of 50-600 for the melt factor? If outside of this is deemed not “physically realistic”, what is assumed with an order-of-magnitude variation here? 50 kg m^-2 K^-1 seems very low – that’s 0.05 m w.e. K^-1? At least a citation would be useful. Also note typo: K^-1 not K^1.

The typo has been corrected, thank you. The range mentioned above is what was used in all OGGM versions until 1.6. For these versions, OGGM did not apply a temperature correction to the reference historical data (e.g. CRU or ERA5) as is commonly done in some large scale models (e.g. Huss and Hock, 2015; Rounce et al., 2019). The melt factor therefore played the implicit role of temperature bias correction (Maussion et al., 2019, Sect. 3.3). “Physically realistic” here is therefore misleading and we will amend the text accordingly.

123: “which represent forecasting from very simple to complex”. Some word is missing. Using simple to complex methods?

Yes, the words ‘methods’ was missing.

135: (and in general) If all of this is in terms of ensemble means, won’t much of the absolute error in comparing observations to reforecasts come from the observation (a single timeseries) having more interannual variability than the ensemble mean? I think it can be valid to focus on ensemble means, but might need to alert the reader to this.

This is a good point, and will be discussed

136: For persistence forecasts with multi-year lead times, are the X years just repeated? Or is the mean used for forcing? I was unclear on how this works.

The X years are repeated. We will change the text accordingly.

166: Drift correction is mentioned here but hasn’t been described yet, which may confuse a first-time reader.

True! We will refer to the section where it is explained

180: default correction factor – citation?

Citation added

235: errors in m w.e. – is that per year, or cumulative?

These are cumulative. Text will be changed.

244-45: “inability of a simple mass balance model to reliably simulate individual years”. I don’t think this has to do with the mass balance model… this is the inherent challenge of internal climate variability on these timescales.

Correct. See our response to the major comment above.

280-2: Isn't lower skill from single model ensembles here partly because the ensemble is smaller? Or are you comparing N=3 ensembles of either one of each, or 3 of the same? Please clarify.

Will be clarified in the manuscript. It indeed has to do with ensemble size

Fig 4 caption: What is decadal forcing? decadal *reforecast* forcing?

Yes, decadal re-forecast forcing

359: What is "mean cumulative mass balance"? Seems contradictory. Or "mean" as in net annual balance?

The word 'mean' should have been, and will be, removed

365: SSP's drifting apart over a few to 10 years strikes me as a tiny effect comparted to internal variability and other factors on these timescales. See e.g., Hawkins and Sutton (2009).
This is true, and will be removed from the discussion

370: Climate change increasing the amplitude of natural variability is rather contentious, I would be hesitant to assert it as "likely" here. And I don't think this is the conclusion of Nijsse et al., 2019 – they are looking at the magnitude of variability across different models with different equilibrium climate sensitivities - which is different than the variability increasing as warming progresses.

This is a very fair assessment, and we agree that this assertion is too strong. We have will remove it from the discussion and are editing the section in accordance with the comments above and those by reviewer 1.

Reference
Hawkins, E., & Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate dynamics*, *37*, 407-418.

References:

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., ... & Eade, R. (2016). The decadal climate prediction project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3751-3777.

Delgado-Torres, C., Donat, M. G., Gonzalez-Reviriego, N., Caron, L. P., Athanasiadis, P. J., Bretonnière, P. A., ... & Doblas-Reyes, F. J. (2022). Multi-model forecast quality assessment of CMIP6 decadal predictions. *Journal of Climate*, *35*(13), 4363-4382.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. (2013). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(4), 245-268.

Hossain, M. M., Garg, N., Anwar, A. F., Prakash, M., & Bari, M. (2022). Drift in CMIP5 decadal precipitation at catchment level. *Stochastic Environmental Research and Risk Assessment*, *36*(9), 2597-2616.

Huss, M., & Hock, R. (2015). A new model for global glacier change and sea-level rise. Frontiers in Earth Science, 3(September), 1–22. https://doi.org/10.3389/feart.2015.00054

Grieger, J., Smith, D., & Boer, G. (2016). Recommendations of the Decadal Climate Prediction Project for bias correction of decadal hindcasts.

Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F., & Lee, W. S. (2012). Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters*, *39*(19).

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., ... & Wu, B. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change*, *9*(2), 94-101.

Lange, S.: Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0) (2019). *Geoscientific Model Development, 12,* https://doi.org/10.5194/gmd-12-3055-2019

Manzanas, R. (2020). Assessment of model drifts in seasonal forecasting: Sensitivity to ensemble size and implications for bias correction. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001751.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., ... & Yeager, S. (2014). Decadal climate prediction: an update from the trenches. *Bulletin of the American Meteorological Society*, *95*(2), 243-267.

Mishra, N., Prodhomme, C., & Guemas, V. (2019). Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. *Climate Dynamics*, *52*(7), 4207-4225.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and forecasting*, *8*(2), 281-293.

O'Kane, T. J., Sandery, P. A., Monselesan, D. P., Sakov, P., Chamberlain, M. A., Matear, R. J., ... & Stevens, L. (2019). Coupled data assimilation and ensemble initialization with application to multiyear ENSO prediction. *Journal of Climate*, *32*(4), 997-1024.

Pasternack, A., Grieger, J., Rust, H. W., & Ulbrich, U. (2021). Recalibrating decadal climate predictions–what is an adequate model for the drift?. *Geoscientific Model Development*, *14*(7), 4335-4355.

Rounce, D. R., Khurana, T., Short, M. B., Hock, R., Shean, D. E., & Brinkerhoff, D. J. (2020). Quantifying parameter uncertainty in a large-scale glacier evolution model using Bayesian inference: application to High Mountain Asia. Journal of Glaciology, 66(256), 175–187. https://doi.org/10.1017/jog.2019.91

Smith, D. M., Scaife, A. A., Boer, G. J., Caian, M., Doblas-Reyes, F. J., Guemas, V., ... & Wyser, K. (2013). Real-time multi-model decadal climate predictions. *Climate dynamics*, *41*, 2875-2888.