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.

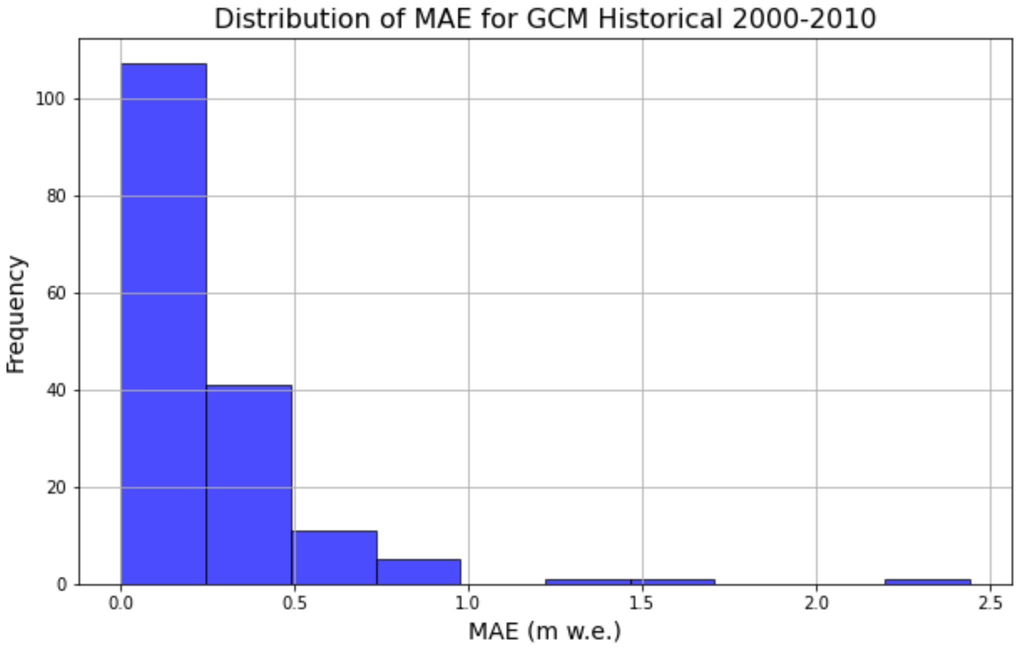


Fig 1. Histogram of the MAE for the GCM Historical experiment 2000-2010

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.

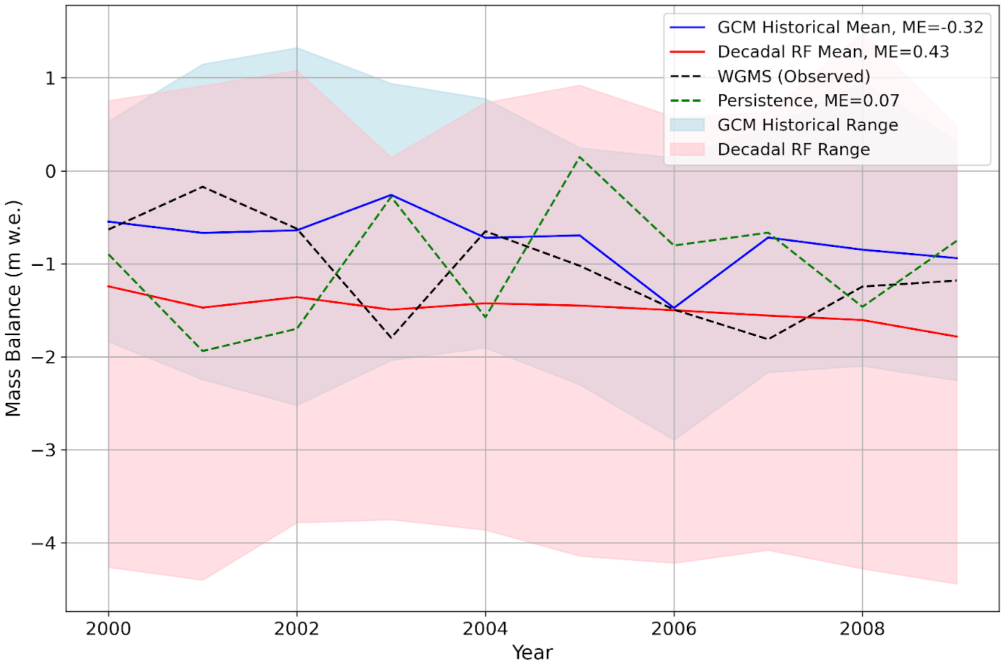
We will amend the text accordingly.

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.



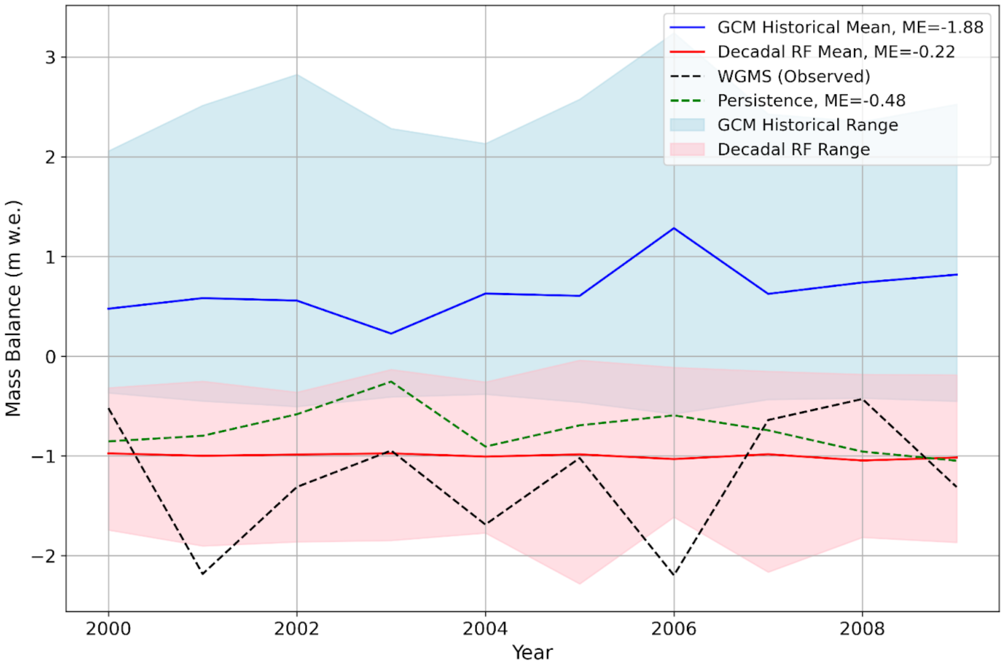


Fig. 3 Example for the Hintereisferner, Austria (above) and Langfjordjøkulen, Norway (below).
Mean errors (ME) for the different experiments, for mean mass balance over the decade, as in Table 2, are indicated in the legend. This means the difference between the mean observed mass balance and the mean 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.