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 centurial-millennial 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 re-forecast 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.

1) Major comments:

<u>Methods:</u>

- 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?
- 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?
- Ln 115: What is the re-calibration step here that is done for the global run?
- Was the calibration with geodetic data also done with CRU (similar to the WGMS calibration)?
- 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?
- 2.2.2 calibration was done over the 2000-2020 period, what about the 2.2.1 calibration?

- 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.
- 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.

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.

- 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).

- 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.

- 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.

  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.

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

- 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).

- 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).

  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?

- 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.

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.

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.

Ln 47: In applications of?

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

Ln 57: What are “impact models”?

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?

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.

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

Ln 110: 94% of the RGIv6 glacier count?

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

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

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

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

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

Is Fig. 4 for 2000 – 2010 period?