Reply to reviewer #1

We are very thankful for the anonymous reviewer's detailed and thorough analysis of our dataset and constructive review of our manuscript. Most of their suggestions are feasible within the time frame of the review. We acknowledge that the proposed changes will make the dataset (i) more robust in terms of uncertainty assessment and (ii) better express the strengths and limitations of the dataset to potential users.

Please find here our preliminary answers to their main comments. The resulting improvements and the response to the specific/minor comments will be detailed in our final response letter and a fully revised manuscript if the editor considers that our manuscript is appropriate for Earth System Science Data.

The document is color coded as follows:
Black: reviewer general comment
Green: answers to reviewer

This study combines in-situ glaciological annual mass-balance observations with remotely sensed sur- face elevation data to provide an annual time series of individual glacier mass changes over the last decades to a century. Based on the assumption that glacier mass-balance anomalies are similar be- tween neighbouring glaciers, annual observed glaciological mass-balance anomalies are extrapolated to all glaciers globally. These anomalies are combined with several geodetic samples to calibrate a mean annual mass change time series and its respective uncertainties.

This extensive data merging and extrapolation study presents an interesting analysis. However, major revisions are necessary before the publication of this dataset to avoid potential misuse by future users.

First of all, there are several major issues, questions and obscurities in the approach of how the uncertainties were estimated (see Sect. 1.1, 1.2) and how the cross-validation was done and analysed (see Sect. 1.3). Another critical aspect is the missing discussion about uncertainties at the glacier scale in the paper itself. The large per-glacier uncertainties become apparent only by checking the dataset itself. Many data users might not be able to use the per-glacier dataset as the uncertainties surpass the signal in many cases (see Sect. 1.4, Sect. 1.5).

Another essential aspect is to communicate clearly that this dataset is not a purely observed dataset since it was created by extrapolation. When extrapolating, predictions are made about unobserved glaciers/years based on an underlying assumption or rule. This dataset is based upon a model of belief of how the system behaves. Upon publication on the WGMS website, the dataset might be misused and falsely interpreted as observations. Most data users neglect or do not include the uncertainties in their frameworks. Therefore, it is important that data providers clearly state the limitations of their dataset. It may imply, for example, adapting the title and the analysis (see Sect. 1.6), and also "flagging" respective regions or glaciers by adding some "metadata" to the data (see Sect. 1.7). In general, the manuscript should focus much more on the uncertainties that vary regionally and temporally and showcase for what use case the data can be used and for what the data might not be useful.

There are several steps in the manuscript that I find unclear, particularly concerning the statistical analysis. I believe these issues need to be addressed by the author team and eventually reviewed by a statistical expert (if not already done). Additionally, the paper and data require a substantial rewrite before they can be reviewed properly. Consequently, I am only able to partially evaluate the manuscript and the dataset at this time. Only a revised version that incorporates or addresses my comments will enable me to fully assess the study and the dataset's added value.

My major comments are summarised in the 'General Comments' (Sect. 1). Line-by-line comments are in the 'Specific comments' (Sect. 2). After each manuscript section, the respective figures and tables and their captions are commented there as well.

General comments

1.1 Standard error

The manuscript uses two times the standard error as the uncertainty measure for the glacier mass balance anomalies. The standard error of the mean describes the uncertainty in estimating the mean, essentially providing the precision of the mean. In contrast, the standard deviation describes the variability of individual points around the mean, indicating the spread of the mass-balance anomalies. This distinction is visually explained on the following website: https://seaborn.pydata.org/tutorial/error_bars.html. In this context, the standard deviation would indicate how much the glacier anomalies deviate from their mean across different locations and years, which is likely of primary interest to data users. Using the standard error because it "allows" years with more observations to have smaller uncertainties is, to my knowledge, an uncommon approach.

As the reviewer correctly defines, the standard error of the mean describes the uncertainty in estimating the mean, while the standard deviation can refer to any measure of sample variability or spread. However, it seems unclear to the reviewer that the two terminologies are not mutually exclusive. By definition, the standard error is also a standard deviation, namely the standard deviation of all possible means. During uncertainty analysis, a standard deviation implicitly refers to the associated estimate (which can be a mean, or other), and is thus always a measure of uncertainty. We confirm that this is also the case everywhere in our study.

We agree with the reviewer that the justification of our uncertainty methods was sometimes convoluted, and that the language used was inconsistent. We will better explain those in the revised manuscript, either by relating the uncertainties to their physical meaning for a new source, or by justifying how they propagate from other sources.

Additionally, the standard deviation ap- pears to be used later for other uncertainties (e.g., σB_{glac} , density conversion factor uncertainty). The uncertainties are combined by adding the standard error and standard deviation together, which is probably also not a standard practice. Are there references that justify the approaches described in this paragraph?

We refer to the answer above on the fact that standard error and standard deviation are not mutually exclusive, and instead refer to the same concept in uncertainty analysis. Standard deviations always refer to the variability of the estimate described (whether it is a means of samples, a modelling estimate, a temporal or spatial sum for aggregating to larger regions).

For instance, the uncertainty in density conversion of Huss (2013) mentioned by the reviewer is a standard deviation of estimates of density conversions, derived from a modeling exercise. As we do not have any other prior knowledge and use a constant mean density conversion globally, also corresponding to Huss (2013), this standard deviation corresponds to our uncertainty in density conversion at the glacier scale.

The equation in line 193 is unclear and raises further confusion. The equation states that the uncertainty is two times the sum of the different standard deviations of the individual N selected annual glacier mass balance anomalies, divided by the square root of the number of observations. It appears these uncertainties correspond to the individual lines in Fig. 2b. However, it is unclear why a stan- dard deviation is calculated for each glacier anomaly (*i*), which is then summed. Additionally, the explanation provided in the text (line 188) does not seem to align with the equation on line 193. To clarify, the code was briefly reviewed at https://github.com/idussa/mb_data_crunching/blob/c9ab8e10198583d0cb2fc1de809e01e4bd5fbca3/2.1_spatial_anomalies/calc_global_gla_spatial_ anom.py#L505. Based on this, it seems the standard deviation is computed over the observations, not summed, which conflicts with the equation in line 193. The code then appears to calculate a mean over another variable. It also seems that the error is first calculated for every "line" shown in Fig. 2c, leading to an average in the script. However, there does not appear to be any summation applied, suggesting a possible discrepancy in the equation on line 193 or a misunderstanding of the correct line of code. Please clarify this process. Moreover, the rationale behind using a factor of two for the standard error is unclear. Please clarify the reasoning behind this choice (described further in Sect. 1.3).

There was an error in the notation of Equation 4b in the submitted manuscript. We corrected it and simplified it from:

$$\sigma_{var_{\beta_Y}} = 2 \cdot \frac{1}{\sqrt{n_Y}} \sum_{i=1}^{N} Stdev\ \beta_i \qquad \text{to} \qquad \sigma_{IDW_{\beta_Y}} = \frac{\bar{\sigma}_{\beta_{i,n_Y}}}{\sqrt{n_Y}} \qquad (4b)$$

where $\bar{\sigma}_{\beta_{i,n_Y}}$ is the per-glacier mean of the standard deviation of yearly glacier anomalies.

The sum in the previous equation was an error of notation, it was intended to show that we use the mean of the yearly standard deviations. In addition, we changed the subscript “var” for “IDW” to express that this term represents the uncertainty in the IDW spatial interpolation (the way the term is derived, whether it is a standard error of the mean or other, should not be reflected in the terminology as already showed by the equation).

We agree that the use of the factor of two in this equation is not correct. Our uncertainty analysis assumes, as is common practice, that the distributions of errors are normal (i.e. Gaussian). This is what allows us to propagate 1σ terms throughout, but this is not applicable to 2σ. We therefore modify all uncertainty equations to be at the 1σ level, and then report results at the 2σ level (i.e., approximately 95% confidence) in the text, figures and tables, by multiplying any given 1σ uncertainty estimate by 2. This will be corrected in the revised manuscript text and the code will be changed accordingly.

Another point of concern is that the current approach results in glacier mass-balance (MB) anomaly uncertainties that only depend on the amount of included glaciers and their differences in the anomaly. I suggest that a mass balance anomaly from a glacier located further away results in larger uncertainties compared to one that is nearer. Is this accounted for in the uncertainty estimates? Do the uncer- tainties increase if only distant glaciers are available? In some cases, this might occur naturally if the distant glaciers are not clustered, leading to significant differences in MB anomalies and, consequently larger uncertainties. However, if the available glaciers with MB time series are far away but clustered closely together, could the assessed uncertainties be underestimated? Is there any algorithm in place to prevent this potential underestimation?

We agree with the referee that our methodology lacks a way to capture the varying errors with distance to the measurement used for interpolation. At present, our uncertainty in mean glacier mass-balance (MB) anomaly depends only on the number of glaciers included, their individual uncertainties, the distance of the selected anomalies used (inverse-distance weighting spatial interpolation, for which there is no integrated error propagation) and the differences of the individual glacier anomalies as standard error or a measure of the uncertainty in estimating the yearly means (EQ 4b).

To solve this issue within the timing of this review, the best solution is to maintain the IDW interpolation. A potential solution is to replace the IDW spatial interpolation by a kriging spatial interpolation (that includes error propagation natively). Our only worry is that the kriging implementation might not be sufficiently efficient computationally to run on our dataset in a reasonable time. Moreover, since it is still an exploratory solution, we need to analyze first if it really makes a difference on the uncertainty assessment. This is something to explore in the future. For the sake of time we propose adding this in the discussion of the dataset limitations.

As the reviewer suggests, we will make sure users are aware of where our dataset is better and less constrained, not only through the uncertainties but also through adding the necessary metadata. Additionally, we will provide a clear illustration of the results in the manuscript figures and discussion to flag the periods and regions where the dataset is less constrained. As an example: the reduced robustness of the mean calibrated time series during gap-filled years where neighboring glaciers have been used, will be evident not only because of their larger uncertainties, but also clearly delineated in the manuscript figures with dashed lines.

1.2 Uncertainties/Error propagation

For the analysis of per-glacier mass balance uncertainties, the law of random error propagation is frequently used. It would be beneficial to explain, in each instance, why it is considered valid to assume that the errors are completely uncorrelated. Specific examples where random error propagation might not be valid or should at least be discussed are noted in the specific comments (e.g., **L191, L255-257, L269, L276, L281**). It may also be necessary to mention that assuming complete independence could lead to underestimating the actual uncertainties.

We agree the text was not fully clear in showing where assumptions of correlation or no correlation are applied between sources, as we instead primarily focused on spatial correlations. The reason behind this difference in focus is that the assumption of correlation between sources has negligible impact compared to the spatial correlation of errors within the same source. This is because we only have 3 main sources of uncertainties, while we have 200,000 glaciers distributed spatially. For instance, for a sum (of volume changes), assuming for the sake of the example that all errors have the same magnitude, if 3 sources of errors are combined as uncorrelated or fully correlated affects the total uncertainty by a factor of 3/sqrt(3) = 1.7. For 200,000 glaciers, if errors are propagated spatially assuming they are uncorrelated or fully correlated, it affects the total uncertainty by a factor of 200000/sqrt(200000) = 447. This is why we focus on estimating spatial correlation to constrain errors more robustly in this study, while we give less attention to the correlation between sources that has little impact.

Nonetheless, we will edit the manuscript text to clarify the physical meaning of each term and the justification behind assumptions of error propagation as either uncorrelated, fully correlated, or correlated to a certain degree (for spatial correlations) at every instance. Below, our response and propose text to the five cases explicitly mentioned by reviewer #1 (L191, L255-257, L269, L276, L281)

Case 1: referring to L191 in submitted manuscript.
Proposed text:

We then estimate the uncertainty in the mean annual anomaly $\sigma_{\overline{\beta}_{g,Y}}$ by combining the uncertainty propagated from glaciological estimates $\overline{\sigma}_{B_{glac,Y}}$ and the uncertainty in the IDW spatial interpolation $\sigma_{IDW_{\beta_Y}}$. These two uncertainties capture independent sources (errors in interpolation and errors in glaciological measurement), and we thus propagate them as uncorrelated (Equation 4). We note that this assumption is conservative, because the variability of the glaciological estimates used to constrain the uncertainty in spatial interpolation is also affected by the uncertainties of glaciological estimates, which are therefore double counted. Uncertainties of glaciological estimates are largely independent spatially from one another, as they originate from differences in techniques, conditions, or locations in the field measurements. We thus combine them as fully uncorrelated (Equation 4a). The uncertainty in the IDW spatial interpolation is not directly provided by this method and thus delicate to assess, here we chose to estimate it using the variability of the sample (Equation 4b).

$$\sigma_{\overline{\beta}_{g,Y}} = \sqrt{\overline{\sigma}_{B_{glac,Y}}{}^{2} + \sigma_{IDW_{\beta_Y}}{}^{2}} \qquad \text{(4 - edited)}$$

Case 2: referring to L255-257 in submitted manuscript.
We keep this paragraph as is, since it explains that spatial correlation is considered between three error sources: elevation change, density conversion and annual anomaly prediction.

Case 3: referring to L276 in submitted manuscript.

Proposed text:

We propagate the uncertainty in the specific regional mass change, the uncertainty in the regional area (Paul et al., 2015) and the uncertainty in the area change considering them uncorrelated. Errors in the area stem mostly from remote sensing delineation errors, while errors in area change stem from a lack of multi-temporal outlines to constrain area change. They are largely uncorrelated with error sources described above on elevation change, glaciological measurements and anomalies. However, elevation change estimates usually already consider errors in area at the scale of each glacier, so we might conservatively be double counting these.

$$\sigma_{\Delta M_{R,Y}} = \left|\Delta M_{R,Y}\right| \sqrt{\left(\frac{\sigma_{B_{R,Y}}}{B_{R,Y}}\right)^2 + \left(\frac{\sigma_{S_R}}{S_R}\right)^2 + \left(\frac{\sigma_{\Delta S_{R,Y}}}{\Delta S_{R,Y}}\right)^2} \qquad (19)$$

Case 4: referring to L281 in submitted manuscript.
Proposed text:

To simplify the combination of annual values into long term trends or cumulative annual values, we assume the yearly uncertainty to be independent of other years. This is true for glaciological measurement, having an independent uncertainty estimation for each individual year of the time series, but not for the elevation change measurements, where uncertainties are correlated over the years of the survey period. This approach was chosen to make the dataset user friendly.

Regarding **L269, eq. 16**, it is stated that the errors are assumed to be completely correlated at regional scales, but the equation suggests that complete independence is assumed (as indicated by summing the square roots). Which assumption was actually applied in the results? This was not clear from the code.

Case 5: referring to L269 in submitted manuscript.
As explained above, we follow the assumption that correlation between sources has a negligible impact compared to the spatial correlation of errors within the same source. For this reason, at L269 we express that, after applying spatial correlation within error sources, we combine all sources of error propagated at the regional-scale as independent.

1.3 Leave-one-out cross validation

Applying a leave-one-out cross-validation is crucial, and it is great that this validation is performed by using geodetic data available for all glaciers. However, given the nature of the reference glaciers, there are concerns about the validity of the conclusions drawn, such as the claim in line 452 that the "leave-one-out cross-validation results prove that our algorithm can capture the annual variability of individual glaciers."

We will clarify these claims in the revised manuscript after the leave-block-out cross-validation analysis.

As noted in lines 454-456, a major issue arises from the fact that the approach may work well for ref- erence glaciers, often located in regions with nearby glaciers with mass-balance time series. Therefore, evaluating the metrics for these glaciers may not be representative. For example, removing Hintereis- ferner still leaves the nearby Kesselwandferner, which could skew the results. To provide robust esti- mates of the method's performance, a "data-denial/blocking" cross-validation approach is necessary. This involves analyzing how well the algorithm performs when assuming that, for instance, Hintereis- ferner has only one or two randomly selected glacier anomalies located far away, such as in the French Alps. Repeating this analysis across many glaciers and examining how the performance metrics change, as illustrated in Fig. 6, would provide a clearer understanding of the method's robustness. Additionally, evaluating how performance metrics vary with the number of considered glaciers would be valuable.

Please evaluate the approach with a larger glacier sample and the data-denial experiment to better demonstrate the dataset's robustness or non-robustness.

We appreciate the reviewer’s idea, and we agree that performing a so-called "data-denial/blocking" cross-validation approach will certainly add more insight into the robustness of our estimates. If the editor accepts to consider a revised manuscript, we will perform this analysis and add the results in the revised manuscript.

In more detail: For consistency with the literature on spatial statistics, we choose to use the term leave-block-out cross-validation analysis. The process will be similar to the leave-one-out cross validation, with the difference that, instead of removing only the reference glacier time series, we will remove all the spatially selected glacier anomalies surrounding the reference glacier at increasing distances ranges. The mean and standard deviation of the residuals will be calculated at every distance step, to assess potential systematic errors (with the mean), and the magnitude of random errors (with the standard deviation). Results will be plotted showing these errors as a function of the distance to the closer glacier anomaly considered. We will perform this analysis over our selected sample of reference and benchmark glaciers due to the reasons stated in the next answer.

Another consideration is the selection of glaciers for cross-validation. Why are e.g. Echaurren Norte and other WGMS reference or benchmark glaciers not chosen for the cross-validation? Including all glaciers with at least 10 years of observations could allow for a more comprehensive analysis, even if some glaciers have fewer years of data and are not validated. This inclusion would enable assessment in regions without reference glaciers and ensure that performance metrics are not skewed by a few well-sampled regions.

Glaciological time series are subject to biases inherent to the glaciological method. The WGMS highly recommends reference (+30 years) and benchmark (+10 years) glaciers glaciological time series to be reanalyzed every 10 years by calibrating them with long term trends derived from high resolution elevation change measurements (Zemp et al., 2013). We intentionally chose to perform the leave-one-out cross validation experiment with a selected list of reference and benchmark glaciers known to have been reanalyzed. These time series stand as the only ground truth available for validation of our global assessment. The decision of not using all glaciological time series in the experiment is justified by reducing the risk of validating over potentially erroneous “truths”.

Regarding validation, if direct glaciological mass-balance observations were not included in the calibra- tion due to the lack of data over the baseline period 2010-2019, it would be beneficial to use these observations for additional validation if possible.

We disagree with the referee’s comment for the same reasons stated above. These time series do not correspond to reference or benchmark glaciers and therefore might be biased due to the lack of reanalysis. To reduce the risk of validation against biased measurements, we intentionally exclude these glaciers and all non-reanalyzed time series from our cross-validation experiment.

Finally, the claim that cross-validation shows the uncertainty estimates are on the "conservative" side and that the dataset has realistic uncertainties needs clarification. The assessment of whether the cross-validation errors are sufficiently small is based on comparing them to the assumed uncertainties of the dataset. However, this approach may allow for "inflating" the uncertainties until they encompass the cross-validation errors.

The reviewer's comment is somewhat unclear, and we interpret that their statement “this approach may allow for inflating the uncertainties” refers to the practice of iterating (i.e. making changes) on the uncertainty calculation until they agree with the cross-validation results. If this is what is meant, we disagree with the reviewer that this is a potential issue. Iterating to improve theoretical uncertainty quantification until it matches empirical uncertainty estimates from the cross-validation is a good scientific practice, and the very purpose of cross-validation. It helps identify potential gross errors (mistakes in implementation) and ensures a realistic estimation of uncertainties. This is true as-long-as the cross-validation is representative of the conditions in which the methodology is applied for the whole dataset. In this case, as pointed out by the reviewer, we did not sufficiently discriminate estimates spatially during the leave-out process. The addition of the new leave-block-out cross-validation proposed by the reviewer should further improve this.

We note however that the cross-validation cannot identify some sources of systematic errors already present in the estimates used (as they are validated against themselves), only the ones that might be introduced by our methodology. We will add sentences in the text to clarify these points.

In relation to Fig. 6d, there is confusion about the comparison presented. If the y-axis represents $\sigma var_{\beta Y}$ from line 193 (i.e., two times the standard error) and the x-axis shows the mean absolute error, there seems to be a comparison of two different types of errors. The metrics being compared are different in nature: the mean absolute error is calculated differently from the standard error. It is unclear whether these two metrics can be directly compared. Should the x-axis not display the RMSE (Root Mean Squared Error, i.e., typically larger than the MAE), as it involves estimating squared differences, which aligns more closely with the standard deviation? The standard deviation is typically used to measure the spread of errors around the mean, and RMSE would be more appropriate for comparing with it. Comparing RMSE on the x-axis with the standard deviation from the calibration on the y-axis would allow for a more consistent evaluation of prediction error (RMSE) relative to the inherent variability or spread of errors (standard deviation). Please verify this approach (if possible with a statistician) and provide a clear explanation for the chosen comparison, including its validity.

The confusion of the reviewer here is completely acceptable. Thanks to this comment we were able to detect that there was also confusion among coauthors in terms of the best metrics to analyze the cross-validation results. We have now agreed upon using only the mean of residuals and the standard deviation of residuals as metrics to quantify potential systematic errors and random errors within the cross-validation results, respectively. We will not use the mean absolute error or the RMSE, since they don't provide any additional information. We will clarify the meaning of this parameter in the revised manuscript Figures and text discussion and correct the panels from Fig.6 accordingly.

1.4 Limited "glacier anomalies" for specific periods or regions

The manuscript mentions a threshold of at least three glaciers with mass balance anomalies as necessary. However, it appears that in regions such as the Southern Andes or Subantarctic and Antarctic Islands, only Echaurren Norte is used as a source of MB anomalies before the year 2000, and after 2000, only two to three glaciers are included. Are these sources truly representative for all the RGI regions in these areas?

The Southern Andes is a special case because there is only one long-term and continued glaciological time series available for the Central Andes: Echaurren Norte (1976-2023, which is also the only reference glacier in the entire Southern Hemisphere) and only one sufficiently long glaciological time series for the Patagonia region: Martial Este (2001-2023). Both these regions are extremely different in climatology, and we decided to process them differently, considering the 2nd Order RGI regions for the Southern Andes, dividing Patagonia from the Central Andes at 46S. We intentionally tuned the Echaurren Norte anomaly as the mean annual glacier MB anomaly for the Central Andes, the Martial Este anomaly for the Patagonia Andes. The mean annual glacier MB anomaly uncertainty for both regions was calculated using the standard error of these two glaciological time series over their common period.

Past glacier annual mass change assessments (Zemp et al. 2019) used the full annual signal from Echaurren Norte "as is" to estimate glacier mass changes in the entire Andes, as well as all time-series in the Southern Hemisphere: New Zealand, Low Latitudes, Antarctic and subantarctic. In our study, we decided to include the Echaurren's full time series only for the Central Andes, where it belongs and where it is more likely to be representative of the climatology. For Patagonia, New Zealand, Low Latitudes, Antarctic and subantarctic we only include the Echaurren time series only for Gap filling of the past period (before the glaciological observational period of each independent regional sample). Furthermore, for each independent region, and to reduce the effect of possible climatic differences, the amplitude of the Echaurren glacier anomaly on these gap years is normalized to the amplitude of the mean glacier anomalies of the regional sample. The reduced robustness of the mean calibrated time series during these gap-filled years is apparent on the larger uncertainties in the time series. We will make sure to show this more clearly in the revised manuscript figures using dashed lines over the gap-filled years of the over regional time series.

We acknowledge that considerable uncertainty remains, but we are confident that our present approach is better constrained than past studies, and it is the only possible way to go back in time for these regions, considering the lack of past period observations in the Southern Hemisphere.

Similarly, in the Alps, the MB time series are extracted only from Claridenfirn and Silvretta. To my knowledge, these observations are based on very few stakes during the first 40 years (only two stakes?), which likely introduces higher uncertainty compared to more recent MB time series (e.g., Huss et al., 2021, https://doi.org/10.3929/ethz-b-000474039; Huss et al., 2017, https://doi.org/10. 3189/2015JoG15J015). Was this increased uncertainty in the past data accounted for in your analysis? The dataset and the estimated individual glacier MB time series show relatively small uncertainties for Central Europe in the period when anomalies are sourced from only two glaciers. Please clarify how these factors were addressed.

As explained in the answer above regarding our decision to use the standard error as measure of uncertainty, this increased uncertainty in the past data has been accounted for in our analysis. The standard error allows years with few observations to get larger errors than years having a larger observation sample (i.e. larger errors in past years where only a few series are considered, or in regions with few time series). This effect is apparent and well represented in our resulting uncertainties for, e.g. Central Europe, where uncertainties are two times larger before 1952 compared to the better constrained period after 2000. This is also apparent in other regions during periods where only few, or neighboring region glacier time series are used. In general, our results achieve a consistently good representation of the uncertainties across all regions, with larger uncertainties in regions and periods with small glaciological samples or where neighboring glacier time series are used for filling gap years. And vice versa, lower uncertainties in regions and periods with large glaciological samples.

1.5 Uncertainty analysis - signal to noise ratio

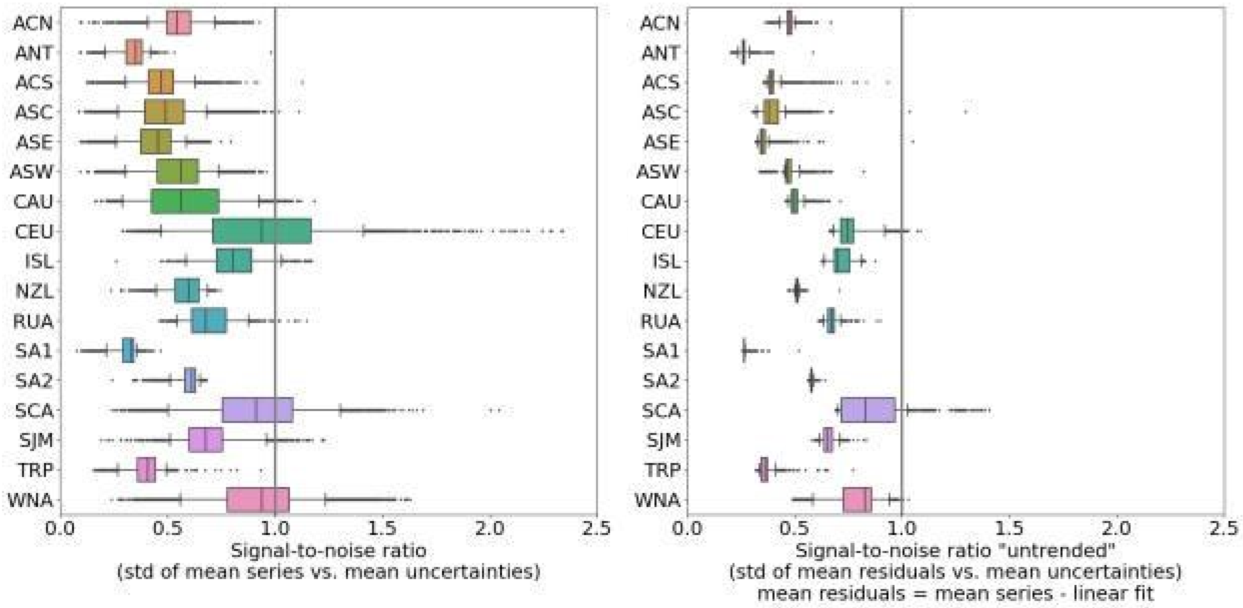
The manuscript would benefit from a more comprehensive uncertainty analysis that examines how uncertainties vary between regions, glaciers, and time periods. This analysis should include a review of the number of glacier mass balance anomalies used, the covered years, their distances, and the amount of geodetic samples. Such information is crucial for potential data users to assess whether the data are suitable for their purposes.

In addition to this analysis, it would be valuable to include a metadata file for each glacier or grid point. This file should detail these statistics and clarify whether a glacier is "unobserved" and if the regional mean was used instead. Ideally, the metadata file would also list the glacier names used to extrapolate the MB anomaly for any given glacier.

While reviewing the paper and examining the data, several questions arise: Where is the annual time series valuable and usueful, and where should caution. A quantitative analysis with statistical tests would be useful for addressing these questions (more discussion on usage cases stated by the authors is in Sect. 1.6).

One potential approach could be a "signal-to-noise" ratio test, where the standard deviation of the mean interannual MB time series is divided by the mean uncertainties (also represented as a standard deviation). If this ratio exceeds one, it suggests that the data adds value; if below one, it implies that uncertainties might overshadow the signal. While this simple ratio is not a rigorous statistical test, it can provide initial insights into data usability. For most glaciers outside Central Europe, the estimated uncertainties are so large that the interannual variability appears smaller than the uncertainty, indicat- ing a signal-to-noise ratio below one (review Fig. 1 left), which raises concerns about data reliability. A more refined approach could involve detrending the time series and comparing the standard deviation of the residuals to the uncertainties (review Fig. 1 right). Repeating the analysis for different time peri- ods could further clarify the data's reliability. Please check with a statistician if this test or another test is suitable. This type of analysis should be included in the manuscript and referenced in the abstract and data documentation.

Figure 1: **Signal-to-Noise ratio analysis for the 20 regions of Dussaillant et al. (in review):** (left) Boxplots illustrating the signal-to-noise ratio, calculated as the ratio of the standard deviation of the mean interannual time series to the mean of the estimated total uncertainties for each glacier individually. A ratio below one indicates that the signal (interannual variability) is smaller than the noise (uncertainties). (right) Untrended signal-to-noise ratio, where a linear trend was removed from the time series to isolate the residuals. The ratio compares the standard deviation of these residuals (signal) to the total uncertainties. Values below one suggest that the residual variability is less than the uncertainties. (right) Untrended signal-to-noise ratio where a linear fit was applied to compute a trend, and then the signal was defined as the "residual" only. In both plots, values below one potentially mean that the signal is smaller than the noise (here assumed to be the uncertainties). The signal-to-noise ratios were estimated from the entire provided time series of each region. The total uncertainties were estimated by assuming complete independence of the three given uncertainty sources.



Performing a signal-to-noise analysis is a good suggestion, it is one way to give a measure of trust in the use of the data. However, a signal to noise analysis can be done in many ways depending on the use that the dataset will be given, it can be performed over various spatial or temporal scales: a specific glacier, a specific region, only during specific years, full time period etc. Further, what is considered as signal and what is considered as noise must be arbitrarily determined depending on the analysis (e.g., 1-sigma or 2-sigma uncertainties?). One might want to observe the signal over a given year related to the entire period, or a specific period and compare it to the mean uncertainty over that specific period or the entire series, etc. Further, there is also the problem that depending on the analysis performed, if something is not statistically significant it doesn't necessarily mean that is not true. Possibilities are endless and will ultimately depend on the specific use of the data. We prefer to put our effort into giving all the necessary information that individual users might eventually require to perform this analysis according to their specific needs.

In addition to this analysis, it would be valuable to include a metadata file for each glacier or grid point. This file should detail these statistics and clarify whether a glacier is "unobserved" and if the regional mean was used instead. Ideally, the metadata file would also list the glacier names used to extrapolate the MB anomaly for any given glacier.

We fully agree that a clear metadata file would be a valuable addition to the dataset and would benefit potential data users to assess whether the data is suitable for their purposes. We will provide for every region and on a glacier-by-glacier basis, a file with all the additional information that might be useful for users:

- The number of glacier MB anomalies used to calculate the mean annual glacier MB anomaly
- List of IDS glaciers MB anomalies used to calculate the mean annual glacier MB anomaly
- The period where regional glacier anomalies are used to capture annual variability.
- The IDS of any additional neighboring region glacier time series used to fill gap in time series

- The period where neighboring region glacier anomalies are used to capture annual variability.
- The mean distance of the spatially selected glacier anomalies
- The distance to the closest glacier anomalies
- The number of elevation change observations available for calibration
- The period with elevation changes observations available
- Clearly identify unobserved glaciers

1.6 Usage of the dataset as described by the authors

Among others, the following usages of the dataset are mentioned by the authors:

- L20: "new baseline for future glacier change modelling assessments and their impact on the world's energy, water, and sea-level budget."

- L376: "This versatility enables identification of individual years marked by significant glacier changes and the detection of zones with varying impacts. For instance, it allows to pinpoint glaciers within a region that were affected by specific annual climate variations (e.g. droughts, floods, heat waves, etc.), as well as those with a larger or smaller influence on the yearly contri- bution to hydrology and annual sea level rise."

- L391: "spatial and temporal impact of known glaciological trends and anomalies like, for example, the Andes Megadrought (Gillett et al., 2006; Garreaud et al., 2017, 2020; Dussaillant et al., 2019) or the Karakoram anomaly (Farinotti et al., 2020; Gao et al., 2020; Ougahi et al., 2022) at an unprecedented yearly temporal resolution.

- L644: "... vast potential for applications in various fields within and beyond 645 glaciology. These include international cryosphere observation intercomparison exercises; multi-Essential Climate Variable (ECV) products; serving as invaluable resources for calibrating and validating climate models; and advancing our understanding of the broader implications of glacier melt on sea levels, freshwater resources, global energy budgets, and nutrient cycling. This work opens new oppor- tunities for future assessments of global glacier mass changes at increased temporal resolutions, fostering a more detailed examination of their climate and hydrological impacts worldwide."

The manuscript suggests that the dataset can be used for a variety of applications; however, there are concerns about the practicality and reliability of these uses, especially considering the uncertainties involved. Also, some of the examples provided are not sufficiently concrete, and it is unclear how uncertainties are integrated into these applications.

Fig. 5 presents an example from Iceland, but uncertainties are not shown. It raises questions about the reliability of pinpointing individual years when uncertainties are accounted for. Iceland benefits from relatively good coverage of mass balance time series and has unique conditions due to the presence of volcaninc eruptions, and is thus not very representative of other regions.

For regions such as the Southern Andes, Subantarctic, and Antarctic Islands, where annual data before 2000 are derived from a single glacier, the added value of the dataset compared to using data from that single glacier (or the few glaciers available) needs clarification. The dataset's ability to represent these regions accurately, considering the associated uncertainties, requires a more detailed discussion.

In lines **357-366**, the manuscript discusses mass changes for regions like the Subantarctic Islands and Periphery. Since these estimates are based on extrapolated data from Echaurren Norte and a few other glaciers post-2000, the confidence in these annual estimates may be limited. A more thorough discussion on how uncertainties impact the interpretation of mass changes should be included if these estimates are to be retained in the manuscript.

In the abstract, line 20 states: "...new baseline for future glacier change modelling assessments". Do the authors believe that glacier models should now calibrate their models to match the per-glacier annual anomalies? In my opinion, glacier models should not, because the uncertainties are way too large. Most calibration procedures just completely neglect uncertainties, and in that case, just calibrating to highly uncertain per-glacier annual MB time series would give a false estimate of confidence. While glacier modelers may benefit from having a more detailed MB time series to better constrain model parameters (such as the precipitation factor), the current dataset may not yet provide the level of precision required for direct application in glacier modeling due to its significant uncertainties. Some modeling approaches do incorporate uncertainties, such as the Bayesian calibration framework utilized by Rounce et al. (2023), which includes uncertainties from the 2000-2019 geodetic observations of Hugonnet et al. (2021). Once the uncertainty estimation approach is clarified and cross-validation is repeated with a data-denial approach, the MB time series and associated uncertainties may become valuable for such calibration methods. However, it is noteworthy that Rounce et al. (2023) did not incorporate the 5-year averaged per-glacier mass change observations from Hugonnet et al. (2021) due to the excessive uncertainties associated with these observations. A similar issue may arise with the current dataset.

Regarding the reviewer's concerns about the practicality and reliability of our dataset uses, especially considering the uncertainties involved, we argue as follows. We agree that the submitted manuscript and dataset was not clear enough to allow users to address this concern. However, we think that the changes suggested by reviewer #1, that have been addressed in the previous answers and will be considered on an updated version of the dataset and revised manuscript, will provide users with transparent information to allow them to define the practicality and reliability of their individual data usage.

To address the specific comments in this section:

Fig 5: The aim of this figure is to provide a visualization of the spatio-temporal resolution of the dataset (i.e. available for individual glaciers, gridded tiles, regions). Iceland was selected as an example for aesthetic reasons: it's the smallest region and easy to visualize fully in one figure. There is no intention of showing or analyzing the specific results or the uncertainties here. This is clearly shown on Fig. 3 and perfectly analyzable from the individual glacier time series and gridded product.

Regions like the Southern Andes and Antarctic and subantarctic islands: The issues regarding these regions have been discussed above. They will be properly addressed in the revised manuscript and figures and in the metadata of the dataset, so that users are aware of the periods where the time series are less robust.

Baseline for future modeling: We agree that the usefulness for modeling is unclear, as a large part of our estimates are extrapolated, rather than interpolated, due to the limited amount of glaciological time series available. We will modify our statements accordingly.

We agree that describing the dataset as a 'new baseline' is beyond our judgment. Data users are in a better position to make such a statement after testing the dataset. We will modify these statements everywhere in the revised manuscript giving them a more cautious tone as potential uses and advantages of the dataset for the modeling community.

1.7 Data and code documentation and availability

Firstly, it is great that the code and data are made fully available.

I have a few comments first on the provided data:

- Hosting the extrapolated / modeled per-glacier annual data on the WGMS website could po- tentially lead to misunderstandings. Given that this dataset is not purely observation-based, its direct availability at the WGMS website could result in misleading conclusions. If the decision is made to include the data directly on the WGMS website, it is essential to include a comprehen- sive "meta"-dataset and a flagging system to highlight glaciers/areas where the uncertainties are too large to extract a signal (as discussed in Section 1.5).

As mentioned above in the answer to comment 1.5, we will provide for every region and on a glacier-by-glacier basis, a .csv file with all the additional information that might be useful for users.

- The type of uncertainty documented in the dataset requires clarification. The term "uncertainty" is used generically, but it is unclear whether this refers to two times the standard error as de- scribed in Line 187, or one or two times the standard deviation (related to Sect. 1.1).

All equations represent uncertainties at 1σ. Reported uncertainties in the text correspond to 2σ = 95% confidence. Therefore, the term "uncertainty" corresponds to 1σ when describing equations and 2σ for reported values.

- Currently, only individual uncertainties are provided, requiring data users to perform their own aggregation. It is strongly recommended to include a dataset with total uncertainties, as this will likely be the most utilized. Additionally, understanding the different sources of uncertainty and their origins took considerable effort. Enhanced documentation explaining these aspects would be beneficial for users.

We will add to the dataset a 4th file for total uncertainties for each glacier combining the individual errors from elevation change, annual anomaly and density conversion factor, as in EQ12:

$${\sigma^2}_{\bar{B}_{cal,g,Y}} = {\bar{\sigma}^2}_{dh,\bar{B}_{cal,g,Y}} + {\bar{\sigma}^2}_{f_\rho,\bar{B}_{cal,g,Y}} + {\bar{\sigma}^2}_{\beta,\bar{B}_{cal,g,Y}} \quad (12)$$

Total uncertainty file name:
RRR_ gla_mean-cal-mass-change_uncertainty_total.csv

One file per RGI 1st order region, where RRR corresponds to the RGI-region code

- To enforce people, to look into the uncertainties, consider creating a netcdf file that has the mean time series, the total uncertainties, and a "flagging" system

The gridded netcdf files already contain the mean time series and the total uncertainties per grid point and per year. As suggested by the reviewer, we will add to this file the additional metadata (as mentioned above in the answer to comment 1.5) as attributes per grid point. This will allow users to easily flag out the dataset to consider only the values that support their specific requirements. In this case, because the netcdf format allows us to have specific metadata for every grid point and every year, a metadata index can be applied to allow users to "flag out" fields depending on index value over specific periods. This will allow, for example, to flag out only specific periods within gridded time series that are not robust, but keep the years where estimates are more robust.

- **Issues found in the per-glacier annual time series**
  - no glacier ID for Greenland, everywhere NaN values as IDs. Please update the glacier IDs for Greenland!

True and well spotted. This bug in the code has been corrected now. Thank you.

  - a bit confusing to have sometimes GLIMS_ids and sometimes RGI_ids

The Hugonnet et al. 2021 dataset used the Tielidze and Wheate (2018) inventory available from GLIMS to calculate elevation changes for in region RGI-12 Caucasus and Middle East. This decision was made because the glacier outlines from the RGI06 inventory are to a great extent erroneous in this region. The Hugonnet et al. 2021 observations were ingested to the FoG database related to the GLIM-Id, to make sure that the calculations correspond to the GLIMS glacier extents. For consistency, in order to use the elevation changes from Hugonnet et al. 2021 for the Caucasus glaciers for this assessment's calibration step, we had to consider the Tielidze and Wheate (2018) inventory as well. We think this decision makes sense. We will make sure this is clearly explained in the revised manuscript.

Comments on the github/code:

- It would be beneficial to include a README document in the GitHub repository that provides a brief overview of the functionality of each script. Such a document would guide interested users on where to find specific processes or analyses within the codebase. While the code does not need to be meticulously documented, a general overview in the README would greatly enhance the accessibility and usability of the repository.

Agreed we will add a README document in the GitHub repository providing a brief overview of the functionality of each script.

1.8 Terminology

- The terms "(mean) glacier (annual) anomaly" appear to be unclear and could benefit from clarification. It is recommended to use more specific terminology, such as "(mean) glacier (annual) MB anomaly" or "glaciers with glaciological MB time series". This issue is particularly evident in Figure 1, where the term is not yet explained. The phrase "glacier anomaly" may imply that the glacier itself is unusual or deviates from expected behavior, rather than referring to mass- balance measurements. Including the term "mass-balance" would help clarify the meaning and ensure consistency throughout the manuscript (e.g., line 169 and other mentions).

Agreed we will replace "(mean) glacier (annual) anomaly" with "(mean) glacier (annual) MB anomaly" everywhere in the revised manuscript text.

- What is the difference between GTN-G regions and RGI6? For instance, in Line 102, GTN-G regions are mentioned, yet later references seem to align more closely with the "usual" RGI6 regions, with the exception of the Southern Andes, which is split differently. It would be beneficial to review the references to GTN-G and RGI6 throughout the manuscript to ensure consistency. If possible, it is recommended to use only one of these terms to avoid confusion.

Agreed we will refer only to RGI regions in the revised manuscript text.

We will also change the generic term “geodetic” to more specific “DEM differencing” or “elevation change” to avoid confusion of dataset users outside of glaciology, since geodetic is a generic term with signification beyond glaciology.

Specific comments

As pointed out before, at this stage we will provide answers to the general and major comments by reviewers. The resulting improvements will then be further described in a complete and detailed response to this review, with individual answers to the following specific comments and a fully revised manuscript. Most of these specific comments can only be properly answered after the dataset has been reprocessed, figures updated, and posterior analysis completed. This is why they are not listed in the present initial response.