Review of *”Annual mass changes for each glacier in the world from 1976 to 2023”* by *Ines Dussaillant et al.*

Earth System Science Data: essd-2024-323

This study combines in-situ glaciological annual mass-balance observations with remotely sensed surface 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 between 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.

1 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. Additionally, the standard deviation appears to be used later for other uncertainties (e.g., $\sigma_{B_{glac}}$, 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?

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 standard 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).

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 uncertainties 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?

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.

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.

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."

As noted in lines 454-456, a major issue arises from the fact that the approach may work well for reference 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 Hintereisferner still leaves the nearby Kesselwandferner, which could skew the results. To provide robust estimates 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, Hintereisferner 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.

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

Regarding validation, if direct glaciological mass-balance observations were not included in the calibration 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.

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.

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.

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?

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.

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, indicating 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 periods 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.

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 contribution to hydrology and annual sea level rise."
- L391: "spatial and temporal impact of known glaciological trends and anomalies like, for example,

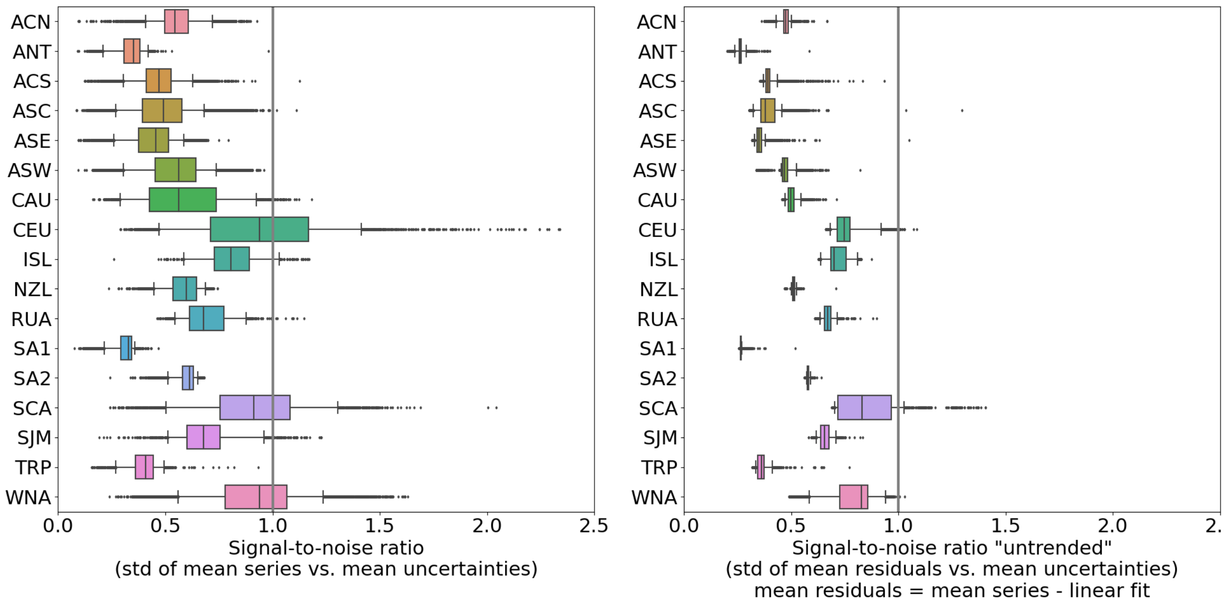

Figure 1: **Signal-to-Noise ratio analysis for the 20 regions of Dussaillant et al. (in review):** (left) Boxplots illustrating the signal-to-noise ratio, calculated as the ratio of the standard deviation of the mean interannual time series to the mean of the estimated total uncertainties for each glacier 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.

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 opportunities 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 a 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.

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 potentially 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 comprehensive "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).

- 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 described in Line 187, or one or two times the standard deviation (related to Sect. 1.1).
- 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.
- 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
- **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!
  - a bit confusing to have sometimes GLIMS_ids and sometimes RGI_ids ...

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.

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).
- 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.

2 Specific comments

To maintain relative conciseness in the review, specific comments have been provided without consistently using phrases such as "please reconsider" or "please change." However, many of these specific comments are intended as suggestions to guide improvements and offer constructive feedback, rather than as strict directives. Some comments made in the general comments section may be repeated in the specific comments section, potentially with additional elaboration or different descriptions. In the response to this review, you may disregard specific comments that you have already addressed in the general comments.

2.1 Title, Abstract & Introduction

- **L1** The title overstates the precision of the data and does not acknowledge the uncertainties sufficiently. Maybe change to something like "Uncertainties in extrapolating annual glacier mass changes from 1976 to 2023: Estimates for every glacier world-wide based on in-situ and geodetic data". At least the methods of your approach should somehow be incorporated in the title. For example, the phrase "from in-situ extrapolation" could be added to the title.
- **L19-20** In my opinion, you can not yet conclude that from the current leave-one-out cross validation (see Sect. 1.3). Also change: "in the conservative side" to "on the conservative side
- **L32-34** From these 500 glaciers, much less are actually usable time series. I think it would be valuable to rather mention how many you use (the glaciers with "glacier anomalies"). You also say, that nearly all glacier regions are represented. However, from these 500 glaciers, a lot are in Central Europe... related to Sect. 1 and idea of giving a clear overview of amount of used glacier MB anomalies and covered years-
- **L54** define FoG here (it is defined only later)
- **L54-58** long sentence, I don't understand the meaning of the sentence, specifically of "and evident..."
- **L67** "global glacier ice changes": mass or volume?
- **L68-71** Why are the geodetic observations of Hugonnet et al. 2021 giving you information on annual mass changes. For example from where do you get the density conversion information here? Rephrase the paragraph to clarify that you use the glaciological mass-balance observations.
- **L74** define "Fluctuations of Glaciers" at the first usage (L54)
- **L77-87** I am not sure if this paragraph is really necessary. You write here what you did, but not really the results. E.g. L78-80: strange to mention that here, as it does not tell the reader in which regions it works and in which it does not work so well...
  L85-87: the two sentences seem to be very similar, maybe combine in one sentence

2.2 Data and Methods

- **L93** do you apply any correction as the RGI dates are often different to year 2000?

- **L109** to which year do these outlines correspond?
- **L104** this should be in the caption, but actually the grey bars are almost not visible
- **L108** FoG already defined in database
- **L114-L115** "throughout their full hypsometry" : what is meant by that?
- **L136** Are these short-term geodetic records also excluded in the statistics of Fig. 1. Please clarify! Similarly, in Fig. 2, you show the short-term geodetic records, I would recommend removing or labelling them as you don't use them for the calibration.
- **L142-148** It is difficult to understand these steps if you haven't read the individual subsections. I would suggest to either put that at the end of Sect. 2.2.3, or instead move the important stuff into the captions, or somehow clarify that the individual steps will be explained later ...
- **L150** Fig. 2 caption: (see Fig. 2 comments below)
- **L160** What is $\sigma_{B_{glac}}$? I assume it corresponds to one standard deviation, but I think it is important to mention that. You often have the σ term in your equations, maybe clarify here directly if those represent always one standard deviation.
- **L169** "glacier annual anomaly" → see Sect. 1.8
- **L168-170** You choose a threshold for the amount of years within 2011-2020 for a glacier to be chosen to be used for your calibration approach. However, how many years outside of the 2011-2020 years are necessary that a glacier is chosen to be used? I think this is important to mention / analyse to understand the "number of glacier anomalies" of Fig. 1
- **L174** "more than two...": replace with "at least three glaciers with mass-balance anomalies". This threshold is only valid for the search radius, as there are many periods and regions where the anomalies come from less than three glaciers? (related to Sect. 1.4)
- **L180** the uncertainties of the anomaly are computed without any inverse weighting correctly? Maybe clarify that!
- **L182** Do you mean here $\sigma_{B_{glac}}$? If yes, please clarify and maybe refer to eq. 1?
- **L183-186** These are almost the same sentences as in line 162-164! Remove one of them. Also typo: replace "ire" with "are"
- **L186-189** Why do you use two times the standard error here, why not the standard deviation. More about that in Sect. 1.1
- **L191** Why can you assume here random error propagation? Couldn't it be that on a specific year, glacier mass balance is under/over estimated for several glaciers because of a specific "climatic" phenomenon? (see also Sect. 1.2)
- **L194** "low confidence glaciological series" → how do you define "low confidence"? Are only high confidence glaciological series included in the statistics of "glacier anomalies" in Figure 1?
- **L196-197** What do you mean by that? I guess you mean by that that you want them to have anomalies until then? This sentence does not say something about how much glaciers with "glacier mass-balance anomalies" are used, but when I first read over the sentence I thought that you want to say that you somehow only select glacier mass-balance anomalies that cover the entire period 1976 to 2023. Please rephrase.
- **L197-200** How do you define "climatically similar" (see comment to Table 2)? Here you describe

gaps in mean calibrated glacier MB annual-anomalies that result from the aspect that no glaciers with observations for these gap years were found within the search radius? But do you also explain how the uncertainties are estimated for these gaps?

- **L212** add the unit of the density conversion factor (maybe best to make it in kg m-3 and then divide by 1000 in eq.5)
  it should also be 0.85+/-0.06 (not 0.60)
- **L215** here the σ_ρ is in units of kg m-3, please be consistent. The acronym is also different to line 212.
- **L213** eq.6 is a bit confusing, as it is unclear for what this is done. Maybe add that this uncertainty is later used for the weighting algorithm of the geodetic samples in eq. 8 (if I got it correctly).
- **L218-222** I had to read this several times to hopefully understand it. You create for every geodetic sample of >= 5 years, an individual time series which goes over the entire considered period (not only the period of record of the geodetic sample), correct?
- **L222** here you write "only geodetic observations larger than 5 years" are used (also in L230), but in L155, you wrote longer or equal to 5 years. Which option did you choose?
- **L224-226** You explain the uncertainty estimate, but there is no equation to it. I have the feeling that exactly this is already done in eq. 9-12, maybe? If yes, please merge this sentence with the explanations there? Or is this where you basically convert $\sigma_{\bar{beta}_{g,y}}$ to $\sigma_{\bar{beta}_{k,y}}$?
- **L231-234** Maybe add the respective acronyms (such as W_t for the second uncertainty part) at the end of the descriptive sentences. Like that it is easier to understand eq. 8, and otherwise the acronyms of eq. 8 are not explained.
- **L239** replace considering with "assuming", because you do not show that,or?
- **L241** I don't really understand why you call it "error" separation. You aggregated the errors beforehand for the weighting, but you have the errors already indivudally, so you don't need to separate them again, or? What eq. 9-11 do is basically averaging the uncertainties from the individual k mass change time series of the individual geodetic samples. If yes, maybe instead add something like that instead of saying error separation. You could also maybe explain everything from line from 239 to 244 before explaining the weighting, and then at the end explain how you get to the total uncertainties (L245). Then you don't need to say you look again at the errors separately...
- **L245** Why can you assume independence and add up the square roots of the different squared standard deviation uncertainty sources to get to the total uncertainty? If I get it right these three uncertainties are the ones available in the per-glacier files. In the data files the total uncertainty is not available, maybe consider adding the total uncertainty as dataset, or somehow clarify that in the manuscript
- **L240-246** What kind of errors do these uncertainties represent? Standard errors or standard deviation? I am just wondering because at least one of them apparently represents the standard error (i.e. the one calculated via line 192?)
- **L248-249** I find this sentence a bit confusing. Maybe clarify that this is what you did, i.e. rephrase by saying.... We calibrated a mean annual mass change for ...

- **L249** Explain what you mean with unobserved. If I understood correctly, it is a glacier that is not available in Hugonnet et al. 2021 (and also does not have any other geodetic observations or in-situ observations). Correctly?
- **L250** replace "Individual" with "individual"
- **L254** why is it only of the observed glaciers? Don't you use all glaciers where you created calibrated data and then in addition the "unobserved" glaciers?
- **L255-257** It is great that you account here for spatial correlations, and have identified that some error sources are significantly correlated spatially, such as elevation change, density conversion and annual anomaly prediction. Are there any figures/analysis for that?
  I am not an expert here, but I am wondering if you need to account for a similar spatial correlation when estimating the total uncertainty of an individual glacier mass balance time series? The reason is that the glacier anomalies are estimated from time series that are coming from different glaciers.
- **L263** Huss et al., in preparation: It would be good to add some details here. When looking into the density uncertainties of individual glaciers, they seem to be quite small.
- **L268** "errors to the real values": what do you mean by that
- **L269, eq.16** see Sect. 1.2
- **L270** "as independent". Why can you assume that?
- **L273** Fig. 2f does not exist
- **L276** Same as L268. Why can you assume that? you use different expression for the same aspect, i.e. assuming that the errors are independent, then write law of random error propagation... maybe stick with one thing
- **L279** Zemp et al. 2019: "et" is written in different text style
- **L281** regional mass loss uncertainties independent and uncorrelated : from where do you know that you can assume that...
- **L285** do you assume that all mass loss is above sea level? eq. 21-23 are not explained
- **L315** starts at the first year of mass change records, is a single mass change record sufficient to start from that year, or is is three glacier MB anomalies?

Table 1

- **L5** maybe add another row with the differences in the Uncertainty?

Fig 1

- It it strange that you show the location of all glaciological samples, although in this study you only use a fraction of these (i.e., just those with glacier anomalies). I think it makes much more sense to visualise those glaciers with a MB time series. I would prefer to see the hypsometry of those glaciers with "glac anomalies" instead of those from all glaciers with measurements.
- For me, the focus of this figure at the moment is to show the hypsometry of the glaciers of that region. For me it would be more interesting to see the statistics of the glaciers that are used for

the "glacier mass-balance anomalies" visually. For example, how many glacier MB anomalies are actually used for the individual years of the time series?

- maybe remove duplicate labels to make the figure less busy
- it is very difficult to see the hypsometry of the "glaciological" sample as the red is difficult to see
- you give the % of observed glaciers in terms of glaciological or geodetic observations. Is a single observation in one year that the glacier is here "observed"?
- The red shows the hypsometry of glaciers with any kind of glaciological observations (I count 468 glaciers there)? If I understand correctly the amount of glaciers with observed mass-balance time series are described by the "glac anomalies"? How many observations are necessary to be such a glacier? Maybe add that to the caption (it is later described in the text, but maybe good to explain it also here). Also related to that: change the wording of "glac anomalies" and describe that in the Fig. 1 caption, see Sect. 1.8).
- the grey "RGI6" glacier hypsometry is almost not visible (specifically if you print it)
- explain in caption the meaning of the circles (glacier region area). Location of the circles is sometimes far away from the region's glaciers. For example in CEU, SCA, NZL. In 17-SAN, there are two circles, probably from the two subregions, this needs to be documented further and if the two circles are kept the region should also be split up via the "black" lines.

Table 2

- Why did you exclude these specific glaciers?
- How did you choose the complementary glacier mass-balance anomalies?
- How did you choose the complementary normalized glacier mass-balance anomalies?
- typo: Hinteeisferner → Hintereisferner

Fig. 2

- It is a bit strange that you show the method for one of the best measured glaciers. Hintereisferner has an annual time series for over 60 years. I think it is very necessary to show at the same time the method for a less well sampled glacier, i.e. a glacier with no in-situ observations, with less and shorter available annual MB anomalies, and with only the Hugonnet et al. 2021 geodetic sample data. This second glacier corresponds better to most glaciers world-wide, I guess? You could add the other glacier in the same figure to have the comparison. Or move the Hintereisferner example in the supplements and add here another glacier.
- You show here the data of Hintereisferner only from 1952 onwards (i.e., the start of the Hintereisferner observational period). In the dataset that you want to publish, however, the new calibrated time series begins already in 1915. I think you should either mention this in the caption or show it in the figures. From the perspective of Hintereisferner, the period from 1915 to 1952 is the most interesting, as this is the period where your method actually creates new data.
- a: There is a big red cross in North Africa with the text B_{glac} search. I find that rather confusing.
- b: you mention 10 glaciers with anomalies, I guess one of the 10 is HEF itself, correctly? So, what we see are 10 thin lines together with the inverse-distance weighted average and the uncertain-

ties around it? And if I understand correctly, some anomalies are over the entire time period, and others are just over a period of time. It would be interesting to somehow visualise that. Probably this gets easier with a glacier with less glacier anomalies around.
I think it would be good to color the line showing the in-situ observed glacier mass-balance anomaly of Hintereisferner, I guess it will be near to the mean annual anomaly? At least it should be clarified in the caption or subplot that one of the lines represents the anomaly of Hintereisferner.
You do not explain what the grey shading is. You added the equation and from that I assume that the grey shading are the uncertainties from eq. 4, but I believe, both the grey shading and the equation need to be explained (e.g. by refering to eq. 4, and saying in caption that the shaded area corresponds to "two standard errors from the glacier MB anomalies and the glaciological sample uncertainties"?

- c: you write that you only use geodetic data with at least five years. However, in the plot, it seems like you also plotted the geodetic sample data for smaller periods? (even for single years, e.g. 2003). As you do not use them in the calibration, I would not include them in the plot, unless you somehow mark them in another color/style to clarify that these are just used for validation? Please also add how many grey lines there are, i.e. how many geodetic mass change observations were used
From Sect. 2.2.3, I understood that every of these "k" lines have their own uncertainty estimated from (b). Although too complex to visualise, you might mention that in the subplot or caption.
I would also prefer to see the uncertainties of the geodetic estimates instead of having red/blue filled ares to the zero line.
- d: What are the red and blue lines? I guess this is the same ones as in c? Not sure if it is necessary to keep them, but in any case, you need to describe them in the caption or in a legend.
You just write, that the grey is "uncertainty". Please clarify what kind of uncertainty it is (see comment in Sect. 1.1).
I would like to see here how the mean calibration time series changes to the actual in-situ HEF observations with that approach. I guess it is quite near as it is included. I think you should add a colored line with the actual Hintereisferner observations (similar as suggested for subplot b).
- add the corresponding equation numbers to b, c, d; consider adding legends into the subplots to clarify better what the lines mean
- Fig. 2b, d: "mean calibrated time series" and "mean annual anomaly" isn't it mean and "some kind of uncertainty" that you show?

table 3:

- "low confidence" glaciological/geodetic estimates : how is this defined?
- empirical function "of"...Hugonnet et al. 2022; Huss and Hugonnet (in prep) → called differently somewhere else, be consistent!

table 4:

- "uncertainty" : what uncertainty does that describe? Standard error/standard deviation/two times standard deviation?
- Dataset 1: there is no dataset with the total error. Interested people need to aggregate the uncertainties themselves, which is error-prone.
- Dataset 2: is the dataset really that large if you add all years together, it is easier to download just one file instead of many...
- Dataset 2: here you have mean time series and total error : again, what does total error represent?
- see table XX: -> ref. is missing
- time series start "of" hydrological year (of was missing)

2.3 Results

- **L326** What do the numbers represent? One standard dev. / std. error?
- **L333** Maybe refer here to Fig. 3; is it "m" or "m w.e", in Fig. 3 it is in m w.e.
- **L356** attention: your estimates are not "observations" anymore, as you apply basically a model to get the individual glacier MB time series. Consider rephrasing the word "observed".
- **L357-366** It is unclear to which figures or tables you refer to in this paragraphs
  You write volumes, but write "GT". I would suggest to replace volume everywhere by mass to coincide with the GT unit.
  Can you really say with confidence that these mass changes occurred on single years? More in Sect. 1.6
- **L359-362** I didn't understand the last part of the sentence, consider rephrasing the sentence.

Fig. 3:

- Global subplot: move that subplot up as the mass change time series axis is very near to the the Russian Arctic subplot. To make some space, move maybe the global pie to the center left part of the plot
- It is a bit confusing that the global plot is in a different style than the regional subplots
- Some of the timeseries are dificult to see due to the bright colors of that region. Consider using black insted and only colo
- Caption: the area of the pie charts: maybe clarify that you mean the size of the circle

Fig. 4:

- Why does RGI19 and RGI05 regional glacier mass increases from 1979 to 2000? Are there any physical explanations for that or other studies showing the same? It seems like Zemp et al., 2019 had less positive MB on these two regions.
- You describe here the meaning of m w.e. but this concept is used earlier. Maybe rather describe it somewhere in the methods?

2.4 Discussion

Fig. 5

- More in Sect. 1.6
- year 1976: (i) and (ii) look very different. On the individual time series, it looks like all glaciers lost mass, while on the gridded dataset it seems rather that they gained mass. Why?

- **L398-400** "...selected considering..": but that means you select glaciers where you know it works. Isn't that kind of a bias towards specific glaciers?
- **L403** How many of the 32 glaciers are in CEU? Does a typical "reference" glacier not have much more "glacier anomalies" in their search radius than a "normal glacier"? see Sect. 1.3
- **L431** "seven glaciers" : In Fig. 2b, you wrote that there are 10 glaciers used to estimate the "Mean annual anomaly" of HEF. If you remove the anomaly of HEF for the cross-validation, there should be still 9 available glaciers, why is it now seven? In Fig. 8a, there are also 9 glaciers listed.
- **L435-436** To my knowledge, the RMSE (combines variance and systematic errors) and STD-diff (variability in the errors) do not directly verify whether there is no systematic error for Ba. Without checking the bias, you cannot confidently rule out systematic errors. I think it would be important to include the bias in Fig. 6,7.
- **L441-444** Maybe clarify by writing sth. like ... For XX out of XX glaciers, the actual standard deviation is >XX larger than the standard deviation estimated by the cross-validation. Do you believe that in general the interannual variability is underestimated by your approach? This is very important to clarify, as e.g. glacier models interannual variability largely depends on the precipitation factor.
- **L448-454** I don't fully agree that you can conclude all of that. I am specifically confused about the comparison of two times the standard error vs MAE. (discussed more in detail in comments about Fig. 6d in Sect. 1.3).
- **L454-456** This is a major problem, and should also be accounted for in your cross-validation and uncertainty analysis. This is one of the main reason why I can not accept the conclusions from this paragraph. See Sect. 1.3 for an idea of a data-denial analysis.
- **L476** are these cumulative mean mass losses within the uncertainties of Zemp et al. 2019? Please add uncertainties to these numbers!
- **L486** Deviations of more than XX : please clarify that you compare here to Hugonnet et al. 2021 2000-2019 period?

  Why do you have these differences in the regional trends w. Hugonnet et al. 2021 in the period 2000-2019? Do they only come from additional geodetic data used over that period and in these regions? Or do the glacier anomalies (in-situ observations) also influence the regional trend over the period 2000-2019?

  In the entire paragraph, the mentioned differences are within the large uncertainty ranges of the regions. If you mention the differences, I believe you should also mention that uncertainties are larger than the differences.

- **L535-537** You apply a "model" by extrapolating the glacier anomalies from reference glaciers to another glacier. For example, you assume that the anomalies are similar for nearby glaciers, you even select a glacier with the most similar climate for glaciers in regions without glacier anomalies. All these choices are somehow like a model. Therefore, you should not call this product a purely observation-based product.
- **L549** From where do you know if the estimated uncertainties are sufficient. The cross-validation that you applied can not tell you that as none of the glaciers of the cross-validation are e.g. in Southern Andes or Subantarctic and Antarctic Islands.
- **L550 onwards** You clearly state the problem, but I am wondering if it is then really valid to still give an annual time series for these regions and periods of extremely high uncertainties
- **Sect. 5.5.1** I think adding an overview of which geodetic data sources were included before year 2000, and which after year 2000 would help. Which kind of additional geodetic observations were used to compare to Hugonnet et al. (2021) in the period 2000-2019? Did you use any photogrammetric data for geodetic observations in the past? L560: I assume, this is future work, so maybe clarify that.
- **Sect. 5.3.2** maybe also discuss potential usage of terminus location (e.g. more used in glacier runoff studies)
- **L592** please remove, your dataset is not purely observational

Fig. 6 (see also Sect. 1.3)

- maybe add in the legend the amount of considered glacier anomalies for each glacier and the region where they are located in!
- FIg. 6a-d: Please adapt all subplots to have the same scale on the x-axis and y-axis, with equal tick labels and lengths. E.g. in python, you can set: $ax.set_aspect('equal')$. I believe this would help a lot to correctly interpret the plots (and would help to understand that all grey dashed lines are 1:1 lines).
- does it work better for those glaciers that have a lot of glacier anomalies nearby?
- Fig. 6d: Can you please clarify clearly the meaning of the y-axis (more details on that in Sect. 1.3).
- caption L419: I find it hard to interpret "std-diff"? Does every glacier count the same? Does it basically describe whether the differences are similarly large for the different glaciers?
- caption L422: mass change "trend" (not std. dev)
- caption L423: mass change "standard deviation" (not trend)
- caption L424: x-y descriptions does not display the actual Fig. 6d. As all other figures represent on the y-axis the leave-one-out cross-validation results, it would be best to exchange x and y-axes in Fig. 6d.

Fig. 7

- maybe add the uncertainties of the "leave-one-out" time series
- caption L459: maybe add the word "observed" here. If I understand it correctly the time series that is actually used in the dataset (i.e., e.g. for HEF Fig. 2d) is not shown here, or? Or is it in case

of these reference glaciers the same?

- caption L459-460: caption description of right and left is reversed! Eventually show the other non-selected reference glaciers in the supplements or appendix

2.5 Data and Code availability

see Sect. 1.7

2.6 Conclusions

- **L628** Here again, I would prefer to remove that statement of "independence" and of "purely observational nature".
- **L629-630** With the current cross-validation analysis and figures that I have, I can not yet conclude that. Please recheck, once you refined the cross-validation (see Sect. 1.3).
- **L645-end** These are very broad use cases, maybe a bit more concrete and nuanced use cases would help to clarify what this new dataset can do (i.e., where it adds value) compared to other existing datasets.
- **L657** replace "The" with "the