Anonymous Referee #3

The manuscript provides an in-depth evaluation of the uncertainty in glacio-hydrological modeling of the Patagonian Andes using the Open Global Glacier Model (OGGM). Through an extensive series of simulations (1920 in total), the authors have quantified how various model choices influence glacier melt and runoff projections. They examined six different model choices (referred to as sources of uncertainty), that relate to both historical datasets (different glacier outlines, glacier volumes, and reference climate data) and future climate forcings (various general circulation models, emission scenarios, and bias correction methods). Additionally, the relative importance of each source of uncertainty was evaluated using a random forest regression method. The analysis revealed that reference climate data is the most critical source of uncertainty, even for metrics related to future projections.

The authors have caried out a rather comprehensive assessment of model uncertainty, which is adequately documented in this manuscript. In response to the feedback during the first round of revision, the authors have reorganized the discussion section and improved the figures and general clarity throughout the text.

I deem the manuscript fit for publication after a minor, mainly textual, revision. Please consider the more detailed list of suggestions below.

R. Thank you for your positive feedback. We have carefully reviewed and incorporated your detailed suggestions in the revised version. We appreciate your input and believe the improvements have enhanced the overall clarity and quality of the manuscript

Abstract

L24: area > 1km2 → with area > 1km2

R. Changed as suggested.

L25: Here, I would suggest removing the mention of the number of catchments and hydrological zones. It does not add to the clarity of the abstract and it is not of key importance.

R. We have made the suggested revision and removed the mention of the number of catchments and hydrological zones from the abstract

L25: Consider replacing “We used different glacier […]” by “As sources of uncertainty, we used different glacier […]” to prevent confusion about what is meant by *each source* in L30.

R. Thanks for the suggestion. Changed as suggested.

L30: “We used the permutation feature importance of random forest regression models to assess the relative importance of each source on the signatures of each catchment.” → “We used the permutation feature importance of random forest regression models to assess the relative importance of each source of uncertainty on the signatures.”

R. Thanks for the suggestion. Changed as suggested.

Introduction

L53-56: Having limited knowledge of South American geography, these lines appear confusing to me. How exactly are the ‘Andean glaciers’, ‘Patagonian Andes’ and, ‘Southern Andes’ related? You mention that the Glaciers in the Patagonian Andes are the dominant source of ice loss in the Southern Andes, and that all Andean glaciers combined are one of the largest contributors to sea level rise. Am I right to assume that the Patagonian glaciers thus make up most of the total number of glaciers in the Andes?

R. The terms "Andean glaciers," "Patagonian Andes," and "Southern Andes" refer to specific geographical features within the broader Andes mountain range in South America. Here’s how they are related:

- Andean Glaciers: This term refers to the glaciers located throughout the Andes mountains. The Andes stretch along the western edge of South America, from Venezuela in the north to Chile and Argentina in the south.

- Patagonian Andes: This term refers to the southernmost portion of the Andes mountains, located primarily in southern Chile and Argentina.

- Southern Andes: The Southern Andes (RGI region 17) typically refer to the segment of the Andes that lies south of approximately 25°S latitude. This region includes the Patagonian Andes and extends towards the southern tip of South America.

In answer to your question, yes, you're right. In this region, glaciers represent 82% of the total glacierised area of the Andes at the time of the inventory (see Study Area section). In the sentences indicated by the comment, we have added the latitudinal ranges.

L54: mass loss → ice mass loss

R. Thanks for the suggestion. Changed as suggested.

L60: in rivers with important glacier cover → in rivers in glacierized catchments. It is unclear what is meant by 'important glacier cover'. In addition, the introduction of the term 'catchments' in relation to rivers, at this stage, helps support your decision to study glacier evolution on catchment-scale rather than for each individual glacier.

R. This has been addressed in the reformulation of the sentence (see next comment)

L60-61: The subsentence 'some of which […]' downplays the rest of the sentence and seems contradicting to the first sentence of *Study Area* (L100). Consider rewriting to e.g., 'However, recent studies have reported increased flows in rivers in glacierized catchments (Masiokas et al., 2019; Vries et al., 2023), with a growing number of rivers showing significant trends ($p < 0.01$) in the last decade (e.g., Santa Cruz; Pasquini et al., 2021).

R. Thanks for the suggestion. We have rewritten the sentences following your suggestion: "Nevertheless, recent studies have reported increased river flows in catchments with important glacierized area (Masiokas et al., 2019; Vries et al., 2023), with a growing number of rivers showing significant trends ($p < 0.01$) in the last decade (e.g., Santa Cruz; Pasquini et al., 2021)".

L82: have examined the influence of → have simultaneously examined the influence of OR have compared the influence of OR have examined the respective influence.

R. Thanks for the suggestion. Changed as suggested ("have compared the influence").

L88: Unclear what is meant by 'the climate model chain components', this could be more explicit by stating e.g., '[…] the main source of uncertainty was associated to climate forcing data […]'

R. Thanks for the suggestion. We have rewritten the sentence to directly mention the main sources of uncertainty (GCMs and SSPs).

L92-97: The scenarios were tested […] → The resulting 1920 scenarios were simulated […]

R. Changed as suggested

Study Area

L106-107: I suggest to mention here why you aggregate the catchments into nine hydrological zones. I.e., 'For this or that purpose/analysis, the catchments were aggregated into nine hydrological zones.' This could help prevent the confusion that was brought up during the initial revision. Otherwise, the reasoning needs to be more explicitly mentioned in the method section.

R. Following your suggestion, we have clarified the purpose of aggregating glaciers into the catchments and then catchments into hydrological zones: "To better analyse the spatial variability in hydrological dynamics and to provide a framework for aggregating projected glacio-hydrological changes, the glaciers in the study area were grouped into catchments, which were then aggregated into nine hydrological zones (Fig. 1). This catchment-scale aggregation is consistent with ongoing efforts to integrate global glacier simulations into hydrological models (Hanus et al., 2024; Pesci et al., 2023; Wiersma et al., 2022), which often operate at the catchment or river scale" This adjustment aims to prevent confusion noted in the initial revision and ensures the reasoning is explicitly stated.

L108: '[…], that showed a strong capacity to reproduce recent glacier changes.' Technically, 'that' refers to spatial patterns in precipitation and temperature, which cannot explain glacier changes. Either rephrase to '[…], that showed a strong capacity to reproduce recent spatial variability in glacier changes.' OR ''[…]. Precipitation and temperature showed a strong capacity to reproduce recent glacier changes.' or similar.

R. Thank you for your suggestions. Following the argument of aggregation to hydrological zones, we have clarified the spatial aggregation: "The zones were selected based on the spatial patterns of precipitation and temperature, which have previously shown a strong capacity to explain recent spatial variability in glacier change (Caro et al., 2021)".

Figure 1: It is unclear to my why the names of the main catchments are included in all the figures. In my opinion it makes the figures more cluttered and less obvious. If you are not explicitly referring to these names, I suggest to remove them.

R. Thank you for your feedback. We have followed your suggestion and removed the names of the main catchments from the figures (except in Fig. 1) to reduce clutter and enhance clarity.

Methodology

L128: In fact, you are using the same version at the model used by Marzeion et al. (2012), since you set Pf to 1 (i.e., ignore it) and include a Tspinup (which is similar to their Tbias), correct?

R. No, we use an adapted version of the mass balance model presented in Marzeion et al. (2012). In particular, we do not use a bias correction β* in our equation and the calibration strategy is quite different. Part of this new calibration strategy is a dynamic spinup, where we search for an initial past glacier state in 1980 from which the glacier evolves in a way that matches the RGI area. To find this initial state, we use Tspinup as described in point (iv) of L171. Therefore, Tspinup is only used to find an initial glacier state in 1980 and nowhere else. Finally, the model of Marzeion et al. was glacier-wide (with solid precipitation computed at the glacier top and melt at the terminus elevation) while the OGGM standard model is computed for all elevation bands.

L136: The term 'positive degree-months' has not been introduced with respect to Eq. 1.

R. Thanks for pointing that out. We have added the definition of the term in parenthesis.

L154: It is unclear what 'this' refers to. Consider rephrasing to 'Not considering frontal ablation is an acknowledged shortcoming of our study […]'

R. Thanks for the suggestion. Changed as suggested

L163: Although these lines are rephrased with respect to the initial submission, it is still unclear to me what is meant by the residual term. Based on Maussion et al. (2019), it appears that this is a time-based correction of the observed mass balance. Is that correct? I suggest to mention explicitly what you have done.

R. Thanks for pointing this out. The residual term is not time-based (either in our manuscript or in Maussion et al. 2019). In Maussion et al. (2019), the mass balance calibration relies on the concept of finding t*, a year in the past when the current static glacier geometry would be in equilibrium with the climatic mass balance of that particular year. However, we now use a completely different strategy to find the apparent mass balance (we simply add a constant offset to the mass balance

profile to find a mass balance in equilibrium with the static glacier geometry). We have tried to make this clearer in the manuscript.

L173: The Tspinup appears very similar to Tbias which is more often used in mass balance equations like Eq. 1. In fact, in Appendix A the term temperature bias is also used. For consistency and compatibility with other studies, consider adding Tbias to Eq. 1 and mention in step i that this parameter is initially set to 0 for reference period simulations.

R. Thank you for highlighting this issue. We agree that the similarity between "Tspinup" and "Tbias" could lead to confusion. To address this, we will revise Appendix A to consistently use the term "Tspinup". This discrepancy arose from changes in variable naming between the initial manuscript and the current version, and we apologise for any confusion caused. This interestingly also led to name changes in the OGGM codebase to avoid confusion [link].

L195: I suggest to explicitly introduce the distinction between historical and future uncertainty sources from the start and ideally use the same or similar headers as in 4.1.

R. We have addressed this suggestion by explicitly introducing the distinction between historical and future sources of uncertainty from the beginning. In addition, we have ensured that the headings in the Methods and Results sections are similar to maintain consistency and clarity throughout the manuscript.

L224: PMET outperformed ERA5 […] → In an earlier study, PMET outperformed ERA5 […]

R. Thanks for the suggestion. Changed as suggested.

L247: This description of the different bias correction methods is rather abstract. Although I understand that (explaining) these methods is not the focus of the study, I suggest to clarify. For instance (L246-248): '[…] method that combines quantile-based delta change and bias correction methods. Thus, it not only preserves the quantile changes predicted by climate projections, but also corrects the biases of modelled time series with respect to those of the reference time series.' Here, instead of explaining the method, the second sentence repeats the first sentence. In addition, 'but also corrects the biases of modelled time series with respect to those of the reference timeseries', is essentially the main aim of any bias correction methods, so it doesn't provide any real information on what distinguishes this method from the others.

R. Thank you for this comment. We appreciate the opportunity to enhance the clarity of our descriptions. We have revised the text to better distinguish the methods and clarify their specific characteristics. Specifically, we now emphasize that:

- Quantile Delta Mapping (QDM) not only corrects biases but also preserves the projected changes across the entire distribution of the climate variable, distinguishing it from methods like Mean and Variance Scaling (MVA).

- Multivariate Bias Correction with N-dimensional PDF transformation (MBCn) goes a step further by correcting biases in multiple variables simultaneously while preserving the relationships between these variables (in our case, precipitation and temperature).

L269: of → for

R. Thanks for the suggestion. Changed as suggested.

L282: I recommend to introduce the term 'commitment runs' here, as that term is used in the figures, but never explicitly coupled to these 16 runs.

R. Thanks for the recommendation. We have introduced the term "commitment run" in the specific sentence.

L284: You also corrected annual glacier area, volume and specific mass balance for the additional 16 runs.

R. The specific part of the sentence has been removed to avoid confusion about the variables extracted during the commitment runs.

Table 1: first DJF → (December – February) OR (December, January, and February)

R. Thanks for the suggestion. Changed as suggested.

L306: As indicated by one of the earlier reviews, the term RMSE is confusing because this usually refers to an error between predicted and actual (observed) values (hence 'error'), while you are comparing a number of predictions to another prediction using one of the selected options for each source of uncertainty. Is that right?

R. We recognize that the use of RMSE might be confusing in this context. Typically, RMSE refers to the error between predicted and observed values, as you noted. However, in the permutation feature importance, the RMSE is used to measure the difference between predictions from the original model and those obtained after permuting individual features, rather than comparing predictions to actual observed values. The complete procedure is as follows:

1. **Baseline Performance: First, the model's performance is measured using a chosen metric (such as RMSE) with all features intact. This provides a baseline performance score.**

2. **Feature Permutation: Next, the values of a single feature are shuffled, breaking the relationship between that feature and the target variable. The model is then run again using this modified dataset.**

3. **Performance Comparison: The performance of the model with the permuted feature is compared to the baseline performance. If the performance metric (i.e., RMSE) worsens significantly after shuffling the feature, it indicates that the feature is important to the model. Conversely, if the performance remains relatively unchanged, the feature is less important.**

4. **Repetition for All Features: This process is repeated for each feature in the dataset, allowing for the assessment of the relative importance of all features based on how much their permutation impacts model performance.**

We have clarified this procedure in the text.

In that case, does it matter which option (e.g., PMET as reference climate data set) is chosen as the initial selection? For example, imagine the projected glacier evolution to be very similar for ERA5, CR2MET and MSWEP, but very different for PMET: then the obtained total RMSE with respect to the PMET run would be larger than with respect to one of the others. I assume that the RMSE is always computed with respect to the previously selected option (instead of the initial one)? Maybe you can add a note on this in the text.

R. Thanks for the question. It does not matter because this process is repeated for all features, and the choice of a reference among the different model options is solely for comparison purposes. The RMSE is consistently calculated relative to the selected reference, but since every feature undergoes the same permutation process, the results remain valid regardless of the initial choice. This ensures that the analysis accurately reflects the importance of each feature across all potential reference scenarios. This has been clarified in the procedure (see previous comment).

Results

L326: 'While […] both years).' This sentence reads confusing while you could just say that 84.7% of glacier area in RGI6 was acquired in 2000, the majority (~65%) of the data in RGI7 was acquired in 2001. There is no need to hide that on average, there is only one year difference between the acquisition dates in the two datasets. You come back to this in the discussion where you state that the inclusion of local data and different processing (correction) techniques are more important than the acquisition dates.

R. Thank you for your comment. Following your recommendation, we have adjusted the sentence to improve the clarity of the comparison.

L336: The use of the term normalized thickness appears confusing to me because you haven't considered real ice thicknesses. Perhaps you can refer to it as normalized or scaled volumes, or 'ice volume per catchment area'.

R. We have replaced the term normalised thickness with normalised volume, mentioning that the normalisation was performed using the catchment area (e.g. Fig. 4).

L349: spatial climate diversity.

R. Thanks for the suggestion. Changed as suggested.

L352: When you refer to glacier area, is this based on RGI6, RIG7 or does it refer to the entire study domain?

R. The calculation is based on RGI6. We have modified the sentence to clarify this point. This was only mentioned in the caption of Fig. 4

L383: The term 'future climate uncertainty' is not properly introduced. At first, it was unclear to me that this refers to the combined product of four other sources of uncertainty. This section can easily be misunderstood as a preview of the discussion of the respective importance of the six sources of uncertainty on glacio-hydrological modeling. I suggest to introduce this term/analysis in section 3.3.

R. Thanks for the comment. We have introduced the term in Section 3.3 as you suggested: "Finally, to assess the individual impact of each climate uncertainty source, we estimated the future climate

uncertainty, which we defined as the standard deviation across different reference climates (n = 4), GCMs (n = 4), SSPs (n = 4), and bias correction methods (n = 4), resulting in 480 possible combinations."

L405: The results suggest that ice loss will vary according to the different sources of uncertainty → The results indicate variable ice loss depending on model choices.

R. Thanks for the suggestion. Changed as suggested.

L407: Mention the term 'commitment run' in brackets.

R. Thanks. We have added this clarification.

Header 4.3: Use same header as in 3, Hydrological importance of sources of uncertainty.

R. Thanks for the suggestion. Changed as suggested.

L448: accumulated → explained

R. Thanks for the suggestion. Changed as suggested.

L455: 'Consistently, […]' Wasn't this to be expected?

R. We have replaced the "Consistently" with "As expected". This is explained by the fact that some signatures/metrics are calculated only taking into account the reference period and therefore the importance of future sources of uncertainty (GCM, SSP and bias correction method) is expected to be zero.

Discussion

L475-477: This sentence is rather long and the last subsentence is unclear.

R. Thanks for the comment. We have split the original sentence into two: "Despite the dependence of the specific mass balance (units: kg m^{-2} yr^{-1}) on the emission scenarios (Fig. S1), the ice melt component of runoff (units: m^3 s^{-1}) did not show a clear dependence on the emission scenario (Fig. 9). This is because the runoff is not normalized by glacier area, which decreases throughout the century (Fig. S1)"

L490: may have peaked (2021 +- 15) → may have peaked in 2021 (+-15 year).

R. Thanks for the suggestion. Changed as suggested.

L496: more important for → more important than the volume data source for

R. Thanks for the suggestion. Changed as suggested.

L519: what was the other methodology that Watanabe et al. (2019) used? Could any aspect of their methodology explain the different differences between introduced uncertainties?

R. Thank you for your question. Although both studies applied the iterative calibration process (adjusting a precipitation factor, followed by a melt factor, and finally a temperature bias parameter) as outlined by Huss and Hock (2015), it's important to highlight several key differences that complicate a direct comparison between the two. These differences include:

- The studies utilized different modelling systems: GloGEMflow in one and HYOGA2 in the other.

- Compagno et al. (2021) used three historical climate data products (E-OBS, ERA-I, and ERA-5), whereas Watanabe et al. (2019) employed six historical climates, derived from combinations of two air temperature datasets and three precipitation datasets.

- The focus areas differ, with Compagno et al. (2021) studying Scandinavia and Iceland, while Watanabe et al. (2019) focused on High Mountain Asia. These regions may exhibit varying discrepancies between climate products. Additionally, Compagno et al. (2021) modelled 3,411 glaciers in Scandinavia, compared to the 28 randomly selected glaciers analysed by Watanabe et al. (2019).

For the sake of brevity, we have decided to leave the text as it is.

L533: when comparing your projections to those of Rounce et al. (2023), do you consider only the runs with ERA5 forcing or a combination of all historical climate data sets? I believe Rounce uses ERA5.

R. We used all possible combinations (n = 480 per SSP). We have added a clarification in the caption of Fig. S5 with the climate products used in both studies.

End of 5.4: In 5.2 you concluded that historical sources of uncertainty (ref climate followed by glacier attributes) are more important than future sources of uncertainty, while most other studies focus on only the future sources of uncertainty. This could be emphasized more.

R. Thank you for your comment. This aspect was more strongly emphasised in the original submission, but to reduce redundancy we decided not to repeat this idea in several sections of the discussion. However, we have adjusted the last point of the Conclusions (third point) to emphasise this idea more clearly.

Conclusion

L581: 'six sources of data uncertainty' → 'six sources of uncertainty associated with model choices'.

R. Thanks for the suggestion. Changed as suggested.

Appendix A

L614: Here you refer to Tspinup as the temperature bias. See earlier comment.

R. Thanks for pointing this out. We have corrected the parameter name (Tspinup)

Supplementary Material

Table S2: Longitude – Latitude → Resolution (lat – lon)

R. Thanks for the suggestion. Changed as suggested.

Anonymous referee #1

Review revised manuscript "Unravelling the sources of uncertainty in glacier runoff projections in the Patagonian Andes" by Aguayo et al.

Compared to the previous submission, the manuscript has improved a lot and I think that the discussion and figures have become much clearer. Well done!

I still have a few comments that I think should be addressed before the manuscript can be published. I formulate them here below in order of the manuscript (thus not necessarily in order of importance).

R. Thank you for your constructive feedback and for acknowledging the improvements made to the manuscript. We appreciate your recognition of the clearer discussion and figures. We have addressed all the comments you provided, and we believe these revisions have further strengthened the manuscript. Below, we detail how we have responded to each of your points

1. In the previous round I commented on the use of "catchments" in the manuscript. In the revised version, this has not been removed, but less results are mentioned in terms % of catchment area. While I do see the added value of the results aggregated to catchments in the maps, I still think that the catchment division is not really needed here and, apart from the maps, is also not really used. Why not removing the aggregation to catchment area and rather focus on the hydrological zones and the individual glaciers? If there is a good reason to stay with the catchment aggregation, I would 1) explain what the "catchments" are – a catchment does not have a clear definition without explanation, i.e. the catchments could also have been smaller. How were these derived? What level do they represent? 2) Incorporate throughout the results a description of what came out of this catchment aggregation comparison, to make sure that the reader also understands the "catchment level" results.

R. Thank you for your comments and for raising important points about the use of catchment-scale aggregation in our manuscript. We appreciate the opportunity to clarify our rationale and address your concerns. Here's why we believe it is important to maintain catchment-scale aggregation in our study:

The level of aggregation is a critical aspect of glacio-hydrological studies, and we recognise the challenges associated with selecting the optimal scale. Most previous studies have explored different spatial extents, ranging from gauged mountain catchments (e.g., Huss et al., 2014; Mackay et al., 2019) to major river basins (e.g., Huss and Hock, 2018; Ultee et al., 2022; Wimberly et al., 2024) and even global scales (e.g., GlacierMIP2). Our study differs by focusing on the regional scale and comparing multiple sources of data uncertainty, which required a specific scale of aggregation for effective analysis.

By aggregating results at the catchment scale, we maintain consistency with previous studies assessing future climate impacts on streamflow (see Table A1 in Van Tiel et al., 2020). This consistency is critical for comparison purposes and is consistent with ongoing efforts to integrate global glacier simulations into regional/global hydrological models (e.g., GloGEM in Wiersma et al., 2022; OGGM in Pesci et al., 2023; OGGM in Hanus et al., 2024; PyGEM in Long et al. 2024), which often produce results at the catchment/river scale.

In addition, aggregating data at the catchment scale significantly enhances future research by enabling comparisons with other regional hydrological studies. For example, we have made the complete results available at the catchment scale in the Zenodo repository (see Data availability; >60 downloads). This availability supports additional hydrological research and provides crucial insights for future modelers, such as:

- **Identifying the primary sources of uncertainty in specific catchments of the Patagonian Andes.**
- **Determining which modelling decisions should be prioritized to accurately estimate, for example, peak water year.**

As you pointed out, the catchment scale remains in the maps but is also used in the Feature Permutation Importance procedure because the random forest models are based on catchment-wide glacier runoff/melt data. This choice allowed us to assess the impact of factors such as the varying number of glaciers within each catchment due to the use of different glacier inventories. While it would have been possible to use hydrological zones as the aggregation scale, we found this approach to be too broad given the significant spatial variability within hydrological zones.

Finally, we acknowledge that the initial introduction of catchments in our manuscript was inadequate, and we appreciate your feedback highlighting this issue. We recognise that our explanation lacked sufficient detail, leaving room for ambiguity regarding the delineation. To address this, we have revised the study area section to better introduce the catchments.

Wiersma, P., Aerts, J., Zekollari, H., Hrachowitz, M., Drost, N., Huss, M., Sutanudjaja, E. H., and Hut, R.: Coupling a global glacier model to a global hydrological model prevents underestimation of glacier runoff, Hydrol. Earth Syst. Sci., 26, 5971–5986, https://doi.org/10.5194/hess-26-5971-2022, 2022.

Pesci, M. H., Schulte Overberg, P., Bosshard, T., & Förster, K. (2023). From global glacier modeling to catchment hydrology: bridging the gap with the WaSiM-OGGM coupling scheme. Frontiers in Water, 5, 1296344. https://doi.org/10.3389/frwa.2023.1296344

Hanus, S., Schuster, L., Burek, P., Maussion, F., Wada, Y., and Viviroli, D.: Coupling a large-scale glacier and hydrological model (OGGM v1.5.3 and CWatM V1.08) – towards an improved

representation of mountain water resources in global assessments, Geosci. Model Dev., 17, 5123–5144, https://doi.org/10.5194/gmd-17-5123-2024, 2024.

Long, J., Wang, L., Chen, D., Li, N., Zhou, J., Li, X., et al. (2024). Hydrological projections in the Third Pole using artificial intelligence and an observation-constrained cryosphere-hydrology model. Earth's Future, 12, e2023EF004222. https://doi.org/10.1029/2023EF004222

2. The sentences describing "xx% of the glacier area" has reached this or that, came across as a little confusing, and I wonder if they could be formulated as number of glaciers (abs or relative), and then in brackets how much of the glacier area that is?

R. In this study, the way the results are communicated and summarised is inherently complex due to the large number of simulated glaciers (n = ~ 2,000), scenarios (n = 1,920), metrics (n = 10) and variables (glacier runoff and melt) involved. In an earlier version, we considered using the number of glaciers, but this approach was less informative as a few dozen glaciers represent almost the entire glacier area (mainly glaciers in the Patagonian Icefields), and the number of glaciers varies between inventories (RGI6 and 7), adding an additional layer of complexity. Instead, we have decided to present the results in terms of catchment area and glacier area (only glacier area in the revised versions). We think this method better captures the extent of sources of uncertainty across glaciated areas. By focusing on area rather than individual glacier, we aim to provide a more comprehensive understanding of the spatial influence of historical and future uncertainty sources on glacier hydrological dynamics.

3. The sentence "Highlighting the choices we make in the calibration" in the abstract, could some hints be provided on what it has an effect (baseline + model parameters + future starting point(?), as is presented in the discussion)?

R. Thanks for the suggestion. We have adjusted the sentence: "In contrast, the reference climate was the main source in 69% ± 22% of the glacier area, highlighting the impact of calibration choices on baseline conditions, model parameters, and the initial starting point for future projections"

4. P2 L58 (revised manuscript) – should it be relative contribution?

R. Thanks for the suggestion. Yes, it should be relative contribution. Changed as suggested.

5. P2 L63 – I am not sure to understand how limitations in the understanding of glacier processes in the Patagonian Andes can be addressed by downscaling methods?

R. Thanks for the question. We have clarified the first sentences of the paragraph: "Despite advances in glacier research, modelling efforts in the Patagonian Andes remain constrained by limited data for calibration and validation. For example, to circumvent the limited ground-based atmospheric data, many modelling studies have used dynamic and/or statistical downscaling methods based on global climate reanalyses (Table S1)."

6. P2 L65 "have overestimated" – all of the studies?

R. According to the detailed analysis of Sauter (2020), previous studies have overestimated the icefield-wide precipitation with values greater than 7-8 m w.e. yr^{-1} (Table 1 in Sauter 2020). According to the same study, the icefield-wide precipitation averages (period 2010–2016) are likely to be within 5.38 ± 0.59 and 6.09 ± 0.64 m w.e. yr^{-1} on the NPI and 5.06 ± 0.51 and 5.99 ± 0.59 m w.e. yr^{-1} on the SPI

7. P3 L70 – SPI and NPI are not yet defined

R. Thanks for the suggestion. We have defined SPI and NPI in the first mention.

8. L62-L75 – I think that this paragraph is hard to follow, what message does it want to convey? Could it be reformulated?

R. Following your previous comment (5. P2 L63), we have adjusted the first sentence to clarify the message of the paragraph. The revised first sentence now better emphasises that despite progress in glacier research, significant challenges remain due to gaps in data for calibration and validation. This sets the stage for a discussion of the specific problems (e.g. climate discrepancies) and limitations of glacier modelling efforts in the Patagonian Andes.

9. P3 L83 – Sentence starting with "Huss et al. (2014)" – suggest to connect with previous sentence and add something along the lines of "Such studies have shown that xxx"

R. Thanks for the comment. We have modified the sentence to improve the connection with the following examples (Huss et al. 2014; Mackay et al. 2019), highlighting the spatial scale of the previous studies. "However, few studies have compared the influence of multiple components of the modelling chain on projected glacio-hydrological changes, and those that have been conducted are typically local (basin-specific), limiting the broader applicability of their conclusions. For instance, Huss et al. (2014)…"

10. P4 L108 "that showed a strong capacity to reproduce recent glacier changes" – what is meant here?

R. Thanks for the question. Caro et al. (2021; already cited in the main text) used machine learning models (LASSO) to explore variables that explain the spatial variance of glacier surface area in the Andes. They found that the spatial variability of climatic variables had a higher explanatory power than morphometric variables. This means that the spatial patterns of precipitation and temperature within these zones were found to be accurate indicators or predictors of how glaciers have changed over time. A clarification has been added to the specific sentence.

11. P5 L124 "to model the evolution of all the glaciers" – does that contradict with the catchment requirement of having a glacier area larger than 0.1%?

R. Thank you for pointing out this inconsistency. We recognise that the statement "to model the evolution of all glaciers" contradicts the requirement that only glaciers with an area greater than 1 km² are considered, as well as the catchment requirement of having a glacier area greater than 0.1% (although all glaciers meeting this requirement are in catchments with a glacier area greater than 0.1%). To resolve this, we have removed the specific part of the sentence to accurately reflect the scope of our study ("all glaciers ").

12. P6 L134 "obtained from the nearest grid point"- possibly explain grid point, as it could also have been a station (the use of gridded climate data is not yet clear?)

R. Thanks for the comment. We have clarified this sentence: "The climate variables are obtained from the nearest grid point of the climate gridded product (see Section 3.2.2)."

13. P6 L 135 "a value commonly used in the study area" – not only there, but everywhere, since it is the global average tropospheric lapse rate?

R. Thanks for the comment. We have clarified the sentence: "…a value commonly used (see local examples in Table S1)."

14. P7 L 162 "mass balance" – as it is written, it could be read as being zero? Should it be mass balance profile?

R. Thanks for pointing this out. It should read that the apparent mass balance integrated over the whole glacier must be zero. See adapted paragraph.

15. P7 L173 "using the reference climate" – would "perturbed reference climate" work here to avoid confusion

R. Thank you for your suggestion, we adapted this as suggested for clarity.

16. P12 L 305 – I commented on this before, but unfortunately it has not yet become clear; the change in model performance is calculated as the change in RMSE. But what is taken here is "Sim" and what as "obs"? In that sense, there is no model performance, as there are no measurements? Or what is referred to here?

R. Thanks for the question, this point was also raised by the other reviewer. We recognize that the use of RMSE might be misleading considering that RMSE commonly refers to the error between predicted and observed values. In our study, the "Permutation Feature Importance" uses RMSE to measure the difference between predictions from the original model and those obtained after permuting individual features, rather than comparing predictions to actual observed values. The complete procedure is as follows:

1. **Baseline Performance: First, the model's performance is measured using the chosen metric (i.e., RMSE) with all features intact. This provides a baseline performance score.**

2. **Feature Permutation: Next, the values of a single feature are shuffled, breaking the relationship between that feature and the target variable. The model is then run again using this modified dataset.**

3. **Performance Comparison: The performance of the model with the permuted feature is compared to the baseline performance. If the performance metric worsens significantly after**

shuffling the feature, it indicates that the feature is important to the model. Conversely, if the performance remains relatively unchanged, the feature is less important.

4. **Repetition for All Features: This process is repeated for each feature in the dataset, allowing for the assessment of the relative importance of all features based on how much their permutation impacts model performance.**

We have clarified this procedure in the text.

17. P19 L 409 – suggest to first mention precipitation projections (physical reason) and then the low ice volume (statistical reason)

R. Thank you for your suggestion. We have rearranged the sentence to first address the precipitation projections, followed by the mention of low ice volume.

18. P19 L412 – I was a bit confused as the sentence starts with "At the hydrological zone scale"… but this was also discussed in the sentences before?

R. Thank you for pointing that out. To improve clarity and avoid redundancy, we have revised the sentence by removing the phrase "At the hydrological zone scale".

19. Figure 10 – what is the aggregation scale used here? Glaciers, catchments, zones?

R. Thanks for the question. The aggregation scale corresponds to catchments. From Section 3.5: "For this analysis, we selected 329 catchments with at least one glacier (area > 1 km2) in both inventories". We have clarified this in Figure 10 ("Each boxplot aggregates the results obtained from Permutation Feature Importance using the 329 catchments").

20. P23 L 476-477 – Here I am not sure to completely follow, so it may be good to write the reasoning more explicitly. If the ice melt volume does not show a dependance on emission scenario, then that suggests that regardless of emission scenario, we get approximately the same amount of melt volume? It means that the increased melt is offset by the decreased glacier area? If yes, that would be good to describe as in terms of hydrology, the absolute volume is of importance.

R. Thanks for the question. Although the specific mass balance shows a clear dependence on the emission scenarios (Fig. S1), the ice melt component of the runoff does not show a similar dependence (Fig. 9). This is because although higher emission scenarios lead to increased melt rates, the glacier area shrinks significantly throughout the century (Fig. S1). As a result, the total ice melt volume remains relatively constant across different emission scenarios. In other words, the increased melt is largely offset by the reduced glacier area. This balance explains why the ice melt volume does not vary significantly with emission scenarios, highlighting the importance of considering both melt rates and glacier area when assessing hydrological impacts. This clarification has been included in the revised version.

21. P24 L 500-503 – does it therefore also affect the "starting conditions" of the future runs?

R. Yes, the reference climate does affect the "starting conditions" of future runs. Since it establishes the baseline conditions against which future changes are assessed, it directly influences the initial setup and parameters of the models, impacting how future climate scenarios are simulated and interpreted.

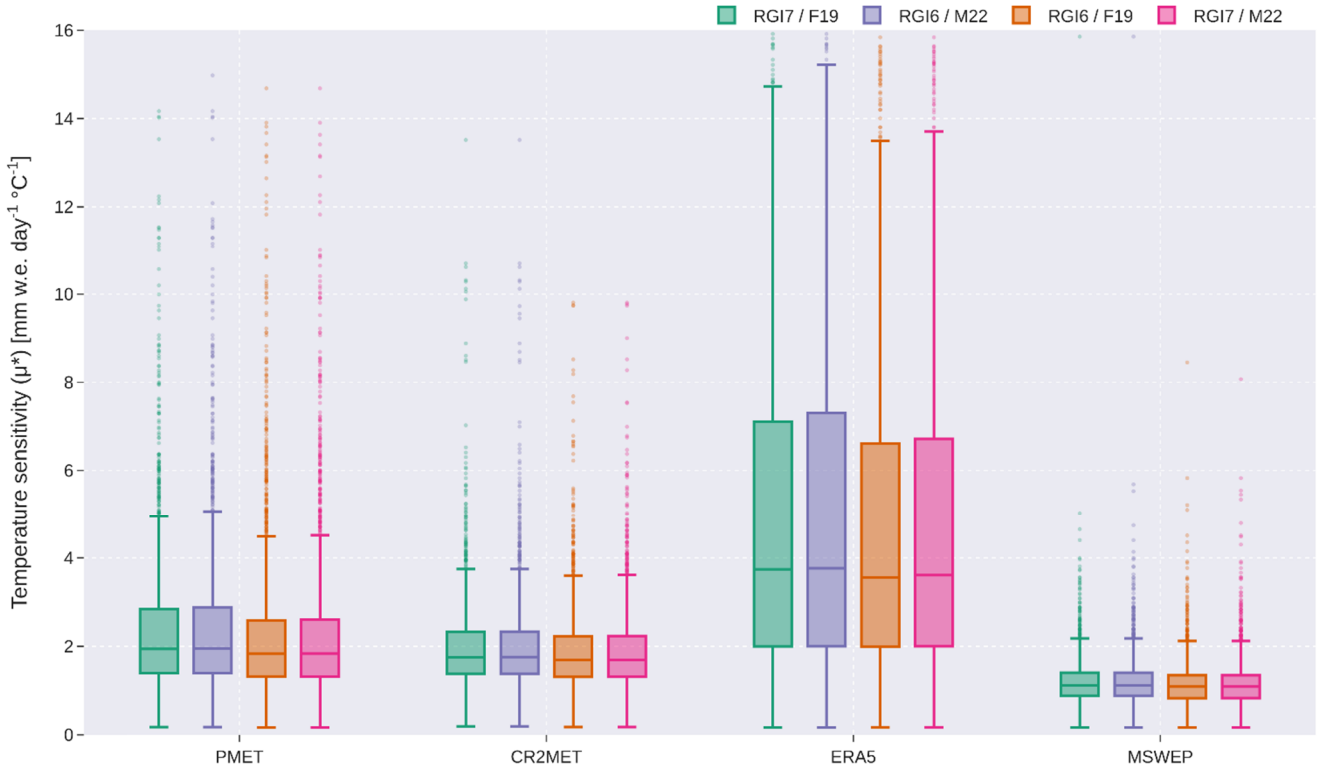
22. P24 L504 "low sensitivity" – should this be "high sensitivity" to fit with the following sentence?

R. Thanks for the question. No, "low sensitivity" is correct in this context. The sentence emphasises that despite the overall variability in conditions, only a few regions showed low sensitivity to the reference climate. The following sentence explains that this low sensitivity is unusual because most climate products have significant differences in solid precipitation, which generally leads to high sensitivity in glacier runoff and melt evolution.

23. P24 section 5.3 Here it would be good to have an actual discussion on how the model parameters differ with different reference climates. This would allow some discussion on how certain parameters + reference climate combinations lead to certain change in glacier runoff + glacier melt

R. Thank you for your comment. In response, we have added a new figure to the supplementary material to illustrate how model parameters differ with various reference climates. Additionally, we have expanded the discussion in Section 5.2: "The reference climate influences temperature and precipitation patterns, which directly shape the seasonal response of glaciers by affecting both melt and accumulation processes. Moreover, the choice of reference climate plays a critical role in parameter calibration, subsequently impacting the model's sensitivity to climate change. For example, Fig. S5 demonstrates that the temperature sensitivity parameter varies significantly with

the reference climate, more so than other factors like glacier geometry and thickness. This highlights how specific combinations of model parameters and reference climates can lead to different outcomes in terms of glacier runoff and melt responses".



Temperature sensitivity for each historical scenario (n = 16). The historical conditions involved in the calibration process considered the geometry obtained from the glacier inventories (RGI6 and 7), the volume obtained from ice thickness datasets (F19 and M22), and the reference climate dataset (PMET, CR2MET, MSWEP and ERA5). More details on the historical conditions can be found in Section 3.2.1. Each boxplot aggregates all simulated glaciers (glacier area > 1 km2), corresponding to 2,034 and 1,837 glaciers for RGI6 and RGI7, respectively.

24. P27 L607 "Downstream hydrology" – please specify which hydrological processes you mean

R. Thanks for the comment. We have specified some hydrological processes in parentheses.