Climate Impacts on Water Resources in a High Mountain Catchment: Application of the Open-Source Modeling Workflow MATILDA in the Northern Tian Shan
(EGUSPHERE-2025-3462)

Final Response to Reviewer Comments RC2

Opening Remarks

Dear reviewer,

Thank you very much for investing the time and effort to review our manuscript. We are confident that the work will benefit significantly from the suggested revisions.

Before we respond to your comments in detail, we would like to make a few general remarks. Our study is designed as a double publication: Part 1 comprises the model description, sensitivity analysis, code and documentation, and has been submitted to Geoscientific Model Development (GMD). Part 2 is a case study with a calibration approach (HESS). We opted for this approach to ensure the model setup and technical implementation including code and documentation were properly reviewed, while also seeking feedback from the hydrological community on the model's application. The editorial support team suggested the procedure to mark both submissions as companion papers and that they would coordinate the review process. However, due to a lack of topic editors, GMD took five months to start the review process. It appears that there was no coordination in terms of aligning the review phases of the two manuscripts. As of the time this response is submitted the search for reviewers is ongoing. For this reason in particular, we acknowledge the importance of both parts to stand for themselves, and we are grateful for your comments. While this is the primary goal, we will do what we can to align the review processes.

Please see our detailed suggestions addressing your concerns below. Your original comments are underlaid in blue.

Overall contribution

1) *The paper is clearly written, but the specific contribution of this "Part 2" manuscript is not fully clear when read independently. While Part 1 introduces the MATILDA model itself, this paper appears to focus mainly on (i) an example application and (ii) the proposed calibration and uncertainty analysis. This distinction should be stated more explicitly in both the abstract and the introduction. At present, the reader has to infer what is new in this contribution.*

That is correct and we will state the objectives more clearly in the abstract and introduction. Additionally, we propose making slight adjustments to the manuscript's general storyline to address Reviewer 1's concerns. Specifically, we will shift the focus away from solely a technical demonstration of a model and toward the actual scientific questions in the study area. Although the workflow, including calibration and uncertainty analysis, is still covered, the main objective will be to answer water-balance-related questions in the study region.

Details on the storyline adjustments will be outlined in the responses to some of your following comments.

Methods clarity

2) *The calibration and validation strategy is difficult to follow in its current form. The workflow involves many steps, datasets, thresholds, and performance metrics, but several of these are not clearly defined when first introduced, or their rationale is explained only later. As a result, the reader* *has to reconstruct the logic of the calibration sequence and understand how individual decisions affect the final parameter sets and projections. Given that the calibration strategy is presented as a key contribution, the manuscript would benefit from a clearer and more structured presentation of the model, variables, and calibration/validation procedure.*

Technical details have been included to enable in-depth discussion of the calibration approach and ensure complete reproducibility. Nevertheless, we agree that the article would benefit from a clearer structure and less detail in the text. Therefore, we will streamline the 'Methods', 'Results' and 'Discussion' sections, moving most of the technical calibration details to Table 2. The following suggestion addresses several of your's and Reviewer 1's comments and is outlined here entirely for better readability. Although the number of subsections will be reduced in the process, the original numbering is referenced. In detail:

Methods:

- A brief model description outlining the modeling chain, its underlying assumptions and data requirements
- A short paragraph titled "Forcing Data" inserted before the Calibration section (3.1) will provide:
    - The data requirements, why ERA5-Land data was used, and how it is preprocessed
    - A brief description of the the climate scenario data and the the bias adjustment
- The separate methods (3.1.2) and result parts (4.2) for every calibration step will be merged into one consistent calibration walkthrough in the Methods section.
- Values already included in Table 2 that are not essential for the workflow description such as filter criteria will be removed from the text.
- Non-essential values not yet included such as sample sizes will be moved to Table 2.
- The section on validation (3.2) will be merged with the calibration procedure.
- The section on uncertainty analysis (3.3) will both be merged with the respective paragraphs in the Results (4.3 and 4.5). , shortened, and moved to the Discussion.

  → The Methods section will therefore consist of a brief model intro followed by the sections "Forcing data", "Calibration data", "Calibration procedure", and "Uncertainty analysis" while all non-essential details will be provided in Table 2.

Results:

- As described above:
    - 4.1 "Reanalysis data" → Methods
    - 4.2 "Calibration" → Methods
    - 4.3 "Validation" → merged with "Calibration" (Methods)
    - 4.4.1 "Bias Adjustment" → Methods
- 4.5 "Uncertainty" will be moved to the Discussion
- 4.4.2 and 4.4.3 will be slightly extended to cover the recent climate trends as well as more climate impact indicators, especially regarding the flow regime and droughts
- The first paragraph of the result comparison to Chevallier et.al. (5.2.3) will be added here to directly contrast the outcomes

  → The Results will focus solely on the recent and projected climate trends, and impacts on the study site's water balance. All intermediate results will be moved to other sections.

Discussion:

- The general structure will be adjusted to better link uncertainties with modeled results and focus on the robustness of the assessment:
    1. Deficiencies in ERA5-Land, the implications for the calibration, and how the shortcomings in the precipitation product can be addressed in local studies (e.g. using high-pass filters, orographic precipitation models, and station data if available)
    2. Deficiencies in the NEX-GDDP-CMIP6 dataset and the implications for the projections
    3. Uncertainties from the calibration data and their implications for century-long projections
    4. Parameter uncertainty and how it is addressed in the calibration strategy
    5. The use of a conceptual model for century-long projections especially regarding parameter stationarity
    6. Robustness of the results and use cases in the light of the discussed uncertainties

    → Uncertainty estimates (4.5), validation procedures (5.2.1), and study comparison (5.2.3) will be integrated where they add values rather than in a separate section

Dual publication and model description

3) *Even though this study is part of a dual-publication framework, the model should still be briefly described before introducing the data and experimental setup. In particular, the main assumptions, structure, and intended strengths of MATILDA should be summarized and justified on their own merits, rather than relying on the companion paper for context. This would make the manuscript more self-contained and accessible to readers who encounter it independently.*

We will add a paragraph briefly summarizing the modeling chain, its underlying assumptions, and data requirements at the beginning of the Methods section (see comment 2). Most of the components are well established and all required sources referenced in Part 1 will be added here as well. A direct reference to Fig. 1 in Part 1 that illustrates the models core routines will be added. Abbreviations including parameter names will be introduced carefully. Long names of all model parameters will be added to Table 2.

Suitability for long-term projections

4) *The manuscript acknowledges that MATILDA is a lumped conceptual model with static parameters, limited spatial representation, and simplified glacier dynamics. However, the implications of these limitations for 80–100 year projections are not discussed in sufficient depth. Given the strong non-stationarity expected in glacier cover, snow regimes, and land-surface processes, this issue deserves a more critical examination.*

We agree that stationarity in the model parameters may limit the ability to capture long-term changes in snow regimes, vegetation and evapotranspiration, and other land-surface feedbacks over long timescales. In response, we will revise the discussion to link model limitations and associated uncertainties more directly to the interpretation of the projected climate change impacts (see comment 2). We will also clarify that, in its current setup, MATILDA is primarily intended as a scientific and educational tool to analyze and illustrate potential climate change impacts, such as shifts in seasonality, drought trends, and changes in runoff composition. It is not designed for operational forecasting or hazard-related applications. We will also place this issue in the context of previous studies on the robustness of conceptual hydrological models under climate change in the light of parameter instability and non-stationary conditions (e.g. Merz et al., 2011; Li et al., 2012; Duethmann et al., 2020, van Tiel et al., 2020). In the revised discussion, we will briefly refer to these findings to better frame the long-term uncertainties of MATILDA and to relate them to its intended use.

We will furthermore briefly discuss possible ways to address these challenges. Data-related aspects already mentioned in the manuscript, such as vegetation changes including mountain greening, will be moved to this section and discussed in the proper context. In addition, model-related options, including those discussed by Duethmann et al. (2020), will be briefly considered in light of the intended scope and application of MATILDA.

Relevance for practitioners

5) *One of the main strengths of this work is the accessibility of the MATILDA framework, which has clear potential to support practitioners who lack the resources to assemble extensive datasets or implement complex glacio-hydrological models. The manuscript would be more convincing if the contribution were framed more explicitly as a transparent assessment of what can—and what cannot—be robustly inferred with current data and modeling approaches. This would better align the conclusions with the results and provide more realistic guidance for practical applications.*

We agree to use less assertive language, without suggesting practical applications that are not supported by the current study. Shifting the general storyline more towards the practical application (see comment 1) will help to evaluate the performance of the model and the suitability of the approach more realistically and in light of local challenges. The conclusion section will be adapted to align accordingly, focusing more on the impact of climate on the local water balance within the study catchment. The manuscript will conclude with a rigorous assessment of what MATILDA can and cannot reliably provide.

Appendix figures

6) *There are ten appendix figures related to climate projections. I suggest condensing these into two to three figures. In particular, showing ERA5-Land together with uncorrected projections in the same plots seems unnecessary, as these biases are expected and are addressed during the bias-correction process.*

Agreed. We suggest showing only the ensembles with confidence intervals while the performance metrics of the bias adjustments remain in the text.

Calibration and validation

7) *Validation is a weaker point of the study, and the calibration does not explicitly evaluate interannual variability. The validation period (2018–2020) is short and affected by data gaps, which the authors acknowledge, but the implications are not fully discussed. In particular:*

1) *Winter performance is poor, yet this is barely addressed, despite its importance for low-flow and drought-related applications.*

We agree that the lower winter performance is relevant for the interpretation of low-flow conditions and should be addressed. However, we do not consider this to be sufficient to infer a generally reduced ability to simulate drought-related dynamics. The following will be added to the Discussion:

- The seasonal performance metrics are calculated for 6 months periods. This leads to two opposing effects:
    - The summer score (Apr-Sep) benefits from the strong seasonality signal in runoff.

  - The winter performance (Oct-Mar) is more strongly affected by errors in the timing of snow accumulation and melt due to the lumped approach (as discussed in 5.2.1) and a generally weaker seasonal contrast.
- Low-flow conditions are more sensitive to processes that are only partly represented in the current lumped setup, including spatial variability in snow storage and melt, subsurface storage and release, and local controls on freezing conditions (e.g slope, bed morphology).
- Winter low-flow dynamics are therefore more uncertain in the current model configuration. Their representation could be improved by a more spatially explicit model calibration (e.g. additional sub-catchments or HRUs).

2) The discrepancy between RGI6 glacier area and the random forest estimate around 2002 (about 23%) is large and concerning. It is mentioned but not explored further, even though it directly affects confidence in the glacier evolution results. More details on the random forest model and its training would be helpful.

The difference is indeed large and needs to be addressed. We will add more details on the random forest workflow, including a link to the GEE Code Editor to reproduce the procedure.

However, we consider it likely that the discrepancy is mainly methodological. In particular, the automated classification tends to include glacier-adjacent transition zones such as debris-covered ice, dead-ice bodies, shaded firn or snow patches, and other spectrally similar features. In comparison, the outlines included in the RGI6 involved considerable manual clean-up and other refining processes that differ between regions and source datasets and are difficult to reproduce based on the literature. Our aim in using the same peer-reviewed automated workflow for both dates was to ensure a consistent and reproducible basis for estimating the relative glacier change.

As a compromise, we suggest implementing a clumping and connected-component filter after classification. These filters merge adjacent glacier pixels into coherent glacier bodies and remove small isolated patches that are likely classification noise. This is a standard refinement step in glacier mapping and improves comparability with the reference.

The filter threshold will be defined using the 2002 comparison with RGI6 and then applied to the 2022 classification to preserve methodological consistency. We will describe this procedure and its limitations more explicitly in the revised manuscript.

Title and Abstract

8) *Consider quantifying key results directly in the abstract.*

Will be done.

9) *The graphical abstract is excellent. However, showing glacier area as a “local trend” may be misleading, as it is primarily a climate-driven impact. Splitting temperature and precipitation might better preserve the two-panel layout.*

Thank you. We will implement this accordingly.

Introduction

10) *The regional context is strong but could be slightly condensed.*

We will condense this section accordingly. However, due to the shift in focus more towards a regional study (see comment 1), we will revise the introduction to align with the new storyline.

11) *Clearly distinguish what Part 1 covers versus what this paper contributes (see major comments).*

See comments 1 - 3.

12) *Explicitly state research questions and objectives rather than implying them (e.g., "evaluate the proposed data products in the light of scarce observations" is rather vague).*

The motivation for this part of the study is to demonstrate the tool's capabilities to assess actual climate impacts at the local scale, describe the calibration approach currently used in Matilda-Online and discuss the uncertainties. However, following the reviewer's comments we will slightly adjust the story line (see comment 1) and state the following explicitly:

The primary goal of the revised manuscript is to quantify the impact of projected climate change on the water balance and runoff composition of the glacierized Kyzylsuu catchment through the 21st century, providing a scientific basis for freshwater management in this vulnerable region. To this end, we used an accessible, open-source modelling toolkit designed to enable a wider range of users in the region to carry out these kinds of assessments.

Research Questions

1. How will projected warming affect the total catchment water balance, specifically regarding glacier mass loss, seasonal snow storage, and increasing evapotranspiration?
2. To what extent will the composition of runoff (the relative contributions of glacier melt, snowmelt, and rainfall) be altered by 2100, and how will these changes shift the discharge seasonality?
3. How will dry periods and a changing cryospheric buffering capacity affect the risk of seasonal water scarcity?
4. How significantly do biases in atmospheric forcing and inconsistencies in calibration data impact the reliability of these hydrological projections?
5. What level of hydrological detail can be reliably provided by an accessible, open-source modeling toolkit in a data-scarce region - and what can't?

Study Site

13) *I suggest removing Table 1, as many elements could be incorporated into the text or are already shown in Figure 1.*

Will be done.

14) *Clarify which decades are meant by "recent decades" when referring to peak runoff.*

We will remove this term entirely since the finding refers to the full available timeseries.

15) *Briefly discuss the representativeness of the Kyzylsuu catchment for the wider Tian Shan region and explain why this specific catchment was chosen.*

The following information will be added to Section 2 (or pointed out more clearly):

- MATILDA was developed through a Kyrgyz-German research collaboration, which is why a site in Kyrgyzstan was chosen.

- Long-term data is sparse. Few monitoring stations survived the 1990s, and water-related data is a politically sensitive issue in the region.

- Kyzylsuu is a mesoscale catchment that supplies a small community that relies primarily on agriculture and tourism. Therefore, it represents MATILDA's target areas while offering the rare combination of long hydrological and glaciological records, and a weather station. The gauging station and mass balance record both provide the minimum required data coverage after 2000, a critical period in glaciological terms. Although MATILDA is designed to address data scarcity, its effectiveness can only be evaluated if some in-situ data is available.

- At least one glacio-hydrological projection study has been carried out in the valley, enabling comparison.

Methods

16) *See general comment on model description and framework. It is difficult to review the paper while switching between manuscripts, and the same issue is likely to affect readers.*

We will restructure the manuscript to be self-contained as described in comment 2 and 3.

17) *It is not clear which variables are used as model inputs.*

All inputs will be clarified in the model description (see comments 2 and 3).

18) *Consider adding a column to Table A1 listing the corresponding variables used from each dataset and the associated time period.*

Will be done.

19) *"bias adjustement" → "bias adjustment"*

Corrected.

20) *The justification for selecting Barandun et al. (2021) as the reference SMB dataset should be strengthened.*

We extend on the following more clearly:

- For the calibration period there are merely 7 years of annual surface mass balance observations for only a single glacier in the catchment. As discussed in 5.1.3, the observed glacier is likely to have higher melt rates than the catchment average and the available period is short relative to typical glacier response times.
- Barandun et al. provide the most regionally consistent dataset for Central Asia, whereas all the other datasets are either global or cover the entire HMA. Furthermore, they use transient snow line observations to constrain annual estimates, which they then validate using selected in-situ observations. However, Barandun et al. report a conservative uncertainty of ±0.37 m w.e. yr^{-1} for the absolute annual mass balance values, dominated by temporally correlated (systematic) errors. At the annual scale, this uncertainty is large relative to inter-annual differences, and individual extreme years (e.g. 2012) should therefore be interpreted with caution. The uncertainty range will be added to Figure 2 to reflect this. Given the combined observational uncertainty and the simplified process representation of degree-day-based glacier models (discussed in Part 1), annual-scale evaluation is uncommon and not recommended in comparable glaciological studies. Thus, while we consider the dataset to be the most qualitative available, we calibrated MATILDA to reproduce only the long-term climatic mass balance.

21) *"See 2 → See Figure 2".*

Corrected.

22) *The sentence "The simulated and observed mean annual SMB are compared…" would fit better in the calibration procedure section.*

Will be done.

23) *The calibration procedure is hard to follow without fully reading the companion paper. Several acronyms are not defined.*

See comments 2 and 3. We will define all abbreviations at first occurrence, and restructure the calibration walkthrough for clearer step-by-step logic.

24) *The validation section should be merged with the calibration section, clearly stating from the outset that a split-sample approach is used.*

Done. See comment 2.

25) *Using only 10% of the data (two years) for validation seems very limited; values closer to 30–50% are more common, especially given the non-stationarity in the catchment.*

We agree that the validation period is relatively short and will clarify this limitation in the revised manuscript. However, the nominal 2000–2020 observation period is substantially reduced by data gaps to just less than 17 years, so the effective proportion of data used for validation is higher than the quoted 10%. More importantly, the available record is temporally fragmented, with a major gap between 2015 and spring 2017. Extending the validation period would therefore not simply increase the validation fraction, but would result in a discontinuous validation sample and a substantially shorter, less coherent calibration period.

Given these constraints, we chose 2018–2020 as an independent evaluation period to preserve a sufficiently long calibration period for the multi-year to decadal simulations. We will make this rationale more explicit in the manuscript and note that, particularly under sparse and non-stationary conditions, the suitability of conventional split-sample validation has been questioned in recent hydrological literature (e.g. Arsenault et al., 2018; Shen et al., 2022).

26) *"The hydrograph ranges for (a–d) illustrate parameter uncertainty." Parameter uncertainty of what, exactly?*

Extended to:
"The hydrograph ranges for (a–d) illustrate the uncertainty in simulated total runoff resulting from the joint uncertainty of the unconstrained parameter sets at each stage of the hierarchical calibration (ranging from 21 open parameters in 'a' to 11 in 'd')."

Results

27) *Section 4.1 (Reanalysis data) does not include figures or tables to support the comparison with the local weather station. I suggest computing KGE (or its components) for precipitation and temperature to assess whether ERA5-Land captures seasonal variability but fails mainly in bias.*

We will add the KGE in the text and a figure to the appendix comparing both datasets in the observation period.

28) *Section 4.2 (Calibration) could be shortened or removed if space is needed; Table 2 could be moved elsewhere.*

Will be merged with 3.1.2 (see comment 2).

29) *Performance metrics are very good, likely due to strong seasonality. Is the model able to reproduce interannual variability?*

We agree that strong seasonality can inflate performance metrics. To evaluate this, we will add:

… the KGE computed for monthly discharge anomalies,

… and correlation metrics for observed and simulated annual discharge totals.

These additional metrics will be reported on and discussed in the performance assessment.

30) *The consistency checks applied to climate projections are not clearly described. It is unclear whether the climatology of the historical period in the bias-adjusted projections is expected to match ERA5-Land, and how deviations are evaluated.*

A brief description of the bias adjustment procedure will be added to the Methods section (see comment 2). The following will be stated there:

- The currently applied checks refer to annual aggregates of each ensemble member.
- They exclude members that:
  - deviate more than 3 standard deviations from the ensemble mean in temperature or precipitation.
  - exhibit inter-annual temperature changes of more than 5 K within a single year

Discussion

31) *The discussion covers the main result patterns well but often reiterates results rather than critically interpreting their implications.*

In the revised discussion, we will link the main findings more explicitly to the corresponding sources of uncertainty and clarify for which purposes the results are informative and where caution is required. We will restructure this part of the manuscript so that the interpretation of the projections is directly connected to the model limitations discussed above (see comment 2).

Specifically, we will emphasise that the results are most reliable for identifying long-term, large-scale trends, such as changes in the timing of seasonal runoff, a decline in glacier contribution and a transition towards a more nival-pluvial runoff regime. These results are therefore ideal for evaluating the direction of hydrological change, comparing scenarios and informing long-term water resource planning. However, we will also clarify that the model is less robust with respect to exact timing and absolute magnitudes. In particular:

- The general tendency toward earlier runoff and peak flow is more robust than the exact timing of these shifts, which is affected by the lumped conceptual setup and its limited representation of spatial variability.
- Projected increases in dry spells are more robust as relative trends than as absolute numbers because of forcing-data biases, particularly the sensitivity to precipitation correction.
- Glacier area and mass-loss projections are more suitable for long-term estimates of the impacts declining glacial runoff contribution has on the amount and seasonality of total discharge. However, given the simplified glacier routine and sensitivity to calibration data, the intensity of glacial decline is highly uncertain, even on decadal timescales.
- Changes in runoff composition can be interpreted as robust evidence for a shift in dominant processes, while the exact quantities of rain, snowmelt, and ice melt remain uncertain.

- Low-flow and winter runoff behavior require particular caution because they are more strongly affected by limitations in the representation of groundwater, soil moisture, and winter conditions.

Therefore, we will state more clearly that these results can be used to identify tendencies and vulnerabilities rather than precise year-to-year interpretation, operational forecasting or detailed local management decisions.

32) *The claim that SWE calibration can compensate for precipitation biases appears overstated. SWE constraints help but cannot fully resolve structural or seasonal precipitation biases.*

We will rephrase this part more cautiously, clarifying that SWE constraints can reduce but not fully resolve precipitation biases.

33) *The discussion of observational uncertainty is strong but could be more forward-looking by identifying which specific measurements would most effectively reduce projection uncertainty.*

We strongly agree and are happy to provide suggestions regarding field observations and qualitative studies as well as data sharing policies and other political impediments.

34) *While model limitations are listed, their implications for century-scale projections are not fully explored. A clearer statement on when and for what purposes the results should (and should not) be used would strengthen the discussion.*

See comment 31.

Figures and Tables

35) *Figure 1: The elevation color bar appears truncated; there is no reference to the yellow shading.*

Fixed.

36) *Figure 2: Consider using a different color for “Simulation,” as it is very similar to Miles et al.*

Will be done.

37) *Figure 3: Consider using a consistent color (e.g., black) for observations.*

Will be done.

38) *Table 2 could be integrated into the text, as it is essentially a one-column list.*

Agreed. We will distribute the information in Fig. 1 and the text.

39) *Appendix figures are not always referenced in order.*

Will be done.

References:

Ali, A., Dunlop, P., Coleman, S., Kerr, D., McNabb, R., and Noormets, R. (2023): Glacier area changes in Novaya Zemlya from 1986–89 to 2019–21 using object-based image analysis in Google Earth Engine, Journal of Glaciology, 69, 1–12, https://doi.org/10.1017/jog.2023.18

Arsenault, R., Brissette, F., and Martel, J.-L. (2018): The hazards of split-sample validation in hydrological model calibration, Journal of Hydrology, 566, 346–362, https://doi.org/https://doi.org/10.1016/j.jhydrol.2018.09.027.

Duethmann, D., Blöschl, G., & Parajka, J. (2020). Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change? Hydrology and Earth System Sciences, 24(7), 3493–3511. https://doi.org/10.5194/hess-24-3493-2020

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., & Yan, D. H. (2012). The transferability of hydrological models under nonstationary climatic conditions. Hydrology and Earth System Sciences, 16(4), 1239–1254. https://doi.org/10.5194/hess-16-1239-2012

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. Water Resources Research, 47(2), W02531. https://doi.org/10.1029/2010WR009505

Shen, H., Tolson, B. A., and Mai, J. (2022): Time to Update the Split-Sample Approach in Hydrological Model Calibration, Water Resources Research, 58, e2021WR031 523, https://doi.org/https://doi.org/10.1029/2021WR031523, e2021WR031523 2021WR031523.

van Tiel, M., Stahl, K., Freudiger, D., & Seibert, J. (2020). Glacio-hydrological model calibration and evaluation. WIREs Water, 7(6), e1483. https://doi.org/10.1002/wat2.1483