Response Letter HESS 07/24

Round 1

Reviewer 1 (Cyril Thébault - Major Revision)

Dear authors,

Please find attached an annotated pdf with comments.

I found the document well presented with neat and informative figures. The quick methodological reminders at the beginning of each section are also useful for the flow of the reading. However, I found that the manuscript could be clarified here and there (see comments in the pdf) and that reading could sometimes be difficult due to the use of numerous acronyms. I also found that the document lacked a few methodological elements and a well-defined common thread to link the different elements presented (e.g. sensitivity analysis made but not used after calibration on the streamflow variable in section 3.6. when it was previously only used for validation).

A main point of improvement I see is that the manuscript can (should?) benefit of a benchmark to compare the SWAT-GL model with a standard approach (e.g. SWAT-GL vs a simple degree-day glacier/snow module for glacier/snow variables and SWAT-GL vs SWAT for streamflow) in order to highlight the benefits and limits of SWAT-GL.

Kind regards,

Cyril Thébault

We thank the reviewer, Cyril Thébault, for his very valuable feedback and hopefully addressed all comments in the following adequately. We agree to most of the comments made and in depth answers are provided and we hopefully managed to improve the clarity of the manuscript.

Comments

That sounds subjective to me. I would suggest you to add a comparative element (good compared to what) to back up this statement.

Answer:

We rephrased it to show that SWAT-GL is capable of representing hydroglaciological characteristics not necessarily based on a reference

This introduction provide a good context to the USGS Benchmark Glacier Project and to SWAT, but I would have liked a little more information on glacier routines in general.

Answer:

We added an additional paragraph for this (and further refer to the methods paper of SWAT-GL for further insights):

“Depending on the research question and data available, several glacier routines with different complexity are available for simulating glacier mass balances and melt contribution in hydrological models. An unlimited ice storage that generates melt water based on a calibrated degree-day factor represents the simplest empirical routine (Naz et al., 2014). However, this routine cannot consider glacier evolution, such as glacier retreat. Conceptual routines, such as the volume-area scaling (VA scaling) (Bahr et al., 2015) or the Δh-parameterization (Huss et al., 2010), simulate the spatial dynamic of a glacier as a simplified function of glacier extent, thickness, and elevation range (Tiel et al., 2020). The largest limitation of these methods is the lack of an actual representation of the ice

flow dynamic, which can be simulated with full physics-based algorithms (Zekollari et al., 2022). Since ice flow modules require several input data and the definition of distinctive boundary conditions, such as bedrock roughness, the application is usually beyond the scope of common water balance simulations in glaciated catchments. In recent years, the coupling of water balance simulation models with global glacier models has proven to be a valuable method for predicting the hydrological response of catchments in mountainous regions under a changing climate (Pesci et al., 2023)."

Terminology questions are often very interesting, but difficult to answer perfectly. Here are a few thoughts that may show that the distinction is not always straightforward.
Increasingly, we're talking about models built from modules to represent different processes. For example, HBV, MORDOR or GR models have their own snow modules. However, they could also be used with other hydrological models. The former case (e.g. HBV snow module used with HBV hydrological model) would be referred to as an integrated module, while the latter case (e.g. HBV module used with GR hydrological model) would be referred to as a coupled model, even though the snow representation will be the same.
In addition, the term "coupled" isn't always appropriate for a snow module, since we often use the results of the snow module directly as input to the hydrological model, without the two exchanging any information, which is more related to a chained approach.
Answer:
We completely agree that terminology is always exciting how they are used between different groups and that not always consensus. We therefore included the chained approach example provided by the reviewer to highlight how we understand the term coupled. However, there is not always a clear distinction.

Avoid nested parentheses, as they're not easy to follow. I suggest using an em-dash instead.
Answer:
Corrected accordingly

Should this description be included in the introduction?
I would suggest to includr these technical detailed in the materials and methods section, with the model description.
I would prefer here to have a broader approach of glaciers routines, whether empirical, conceptual or physically based, and compare the glacier module used in SWAT-GL with what already exists.
Answer:
We added some clarifications on that (see also the answer to comment 2) and moved the technical descriptions to the methodology part.

Why is the watershed boundary shown only for the Wolverine glacier?
Nomenclature between stations is not always consistent. I took the Wolverine Glacier as exemple:
- Wheather stations: WS (Weather Station) - Wo (Wolverine) - 1 (station number)
- Discharge Gauge: W (Wolverine) - 1 (gauge number)
- Catchment name: WG (Wolverine Glacier)
I suggest the following nomenclature:
- Wheather stations: WS - WG (if I'm not mistaken, there's only one station at a time, so we could remove the number here, if I'm wrong keep it.)
- Discharge Gauge: DG - WG - 1
- Catchment name: WG
To be picky, the acronym used for glaciers should also be included on the middle map: for exemple, Wolverine (WG).
Answer:
Fully agreed and corrected accordingly

Same comment here about adding acronyms, which make it easier to follow.

Since elevation variation within the watershed can be high, it would be relevant to add its impact, espacially on temperature (for example, by plotting dotted lines for the temperature on the lowest layer and the temperature on the highest layer).
I'd also like to know the solid precipitation rate for the different months of the regime (this could be represented by darker bars overlapping the initial information).
Answer:
We agree and have done it accordingly, however, as we only have one station per basin of which 2 are far apart we added additional temperature lines only for these two basins assuming a lapse rate of -6.5 °C/km. Of course this assumption involves a high degree of uncertainty but just serves as indication. Besides, as no solid precipitation information at the stations were available we added the fraction of snowfall based on a simple model assuming that precipitation below 1°C is snowfall. Again this just serves as a uncertain proxy but hopefully is what the reviewer intended us to do.

If I've understood correctly, Climaa refers to annual precipitation. It should be stated in the caption. The unit must be mm a-1 therefore.
Answer:
Done

nested parentheses
Answer:
Corrected

You refer to South Cascade Glacier right? You should therefore use SCG instead of SC (that you introduced as abbreviation for Snow Cover)
Answer:
Corrected

The unit must be written as m3 s-1 according to HESS author guidelines
Answer:
Thanks for pointing to this, it was corrected accordingly

where Enorm,i is the normalized elevation of ES i [-], Emax and Emin refer to the maximum and minimum glacier elevation [m], and Ei is the actual elevation of ES i [m].

or

with Enorm,i the normalized elevation of ES i [-], Emax and Emin the maximum and minimum glacier elevation [m], and Ei the actual elevation of ES i [m].
Answer:
Corrected based on Option 1

What do you consider to be a low computational cost? a few hours/days? on a single core/parallel on a computer, on a hpc...? It would be interesting to add this information
Answer:
This is a good point and we provided a short clarification that we mean compared to the number of simulations e.g. of other methods such as Sobol, meaning that regardless of single core or parallel runs the overall simulation demand is reduced:
“A lower computational cost hereby refers to the fact that the total number of simulations required is significantly reduced compared to other methods, such as for the Sobol method.”

Maybe explain why to avoid thinking that this value is totally arbitrary.
In Sarrazin et al., 2016:
"Vanuytrecht et al. (2014) highlight that while a low sample size (n = 25) can be suitable for screening, it can be insufficient for factor ranking. Nossent et al. (2011) find that a base sample size of 12,000 is

needed to ensure the convergence of Variance-Based sensitivity indices in their specific case study, however, a much smaller sample size (n < 2000) is sufficient if one is only interested in ranking the most important input factors."
Answer:
An excellent idea, we added thus:
“A value of r=500 can be considered as sufficient for screening and ranking purposes, for example Vanuytrecht et al. (2014) has shown that a stable ranking was achieved using only 25 trajectories, whereby the relatively small numbers of simulations necessary for ranking was further confirmed by Nossent et al. (2011).”

This may be a little outside the scope of what you want to do here, but do you think the results would have been different if you had also used discharge data in your multi-objective calibration, and to what extent would this have improved the representation of discharge at the outlet over the validation period?
Thanks for the interest in this extended discussion, we indeed think that the results would significantly change in favor of a better discharge representation, especially given the results of subchapter 3.6. When we take the SOO results given in 3.6 as an upper benchmark we assume that at the example of the WG the improved results should (logically) be located between the two CDF curves of figure 9 a). However, what might also be very interesting, though probably outside the scope of this manuscript to avoid overloading, is the trade-off of snow and glacier results, basically where the results of glacier mass balance estimates could be located in figure 9 b).

Table 3 cited before table 2
Answer:
Thanks for the hint, the order was flipped accordingly

Until now, you've been using "Tab. X" rather than "Table X".
Answer:
Homogenized accordingly to: “Table XY”

validation period II?
Answer:
Corrected

I got a bit lost in all the dates. A few questions/clarifications:
- Why isn't the validation I (resp. validation II) period always equal to the Mass balance/hypsometry calibration (resp. validation) period?
- do the "-" mean that the periods are similar to those above or that the variable has not been used? Because for snow I have the impression that you're using the same periods, but for discharge the variable isn't used because it's not available.
- What is the gap rate in the discharge data with the asterisk? This implies that the other discharge records must be considered almost complete. What was your threshold?
Answer:
We see and agree that the dates are rather confusing and not easy to follow. Therefore in the text we added:
“A model run comprised the full available glaciological time series of each catchment, due to differences in the availability of snow cover and the glaciological components. Thus, across glaciers calibration phases can differ.”
“Discrepancies in the first calibration and validation phase across glaciers stem from different starting dates when measurements started. It was aimed to make use of the full glaciological time series to account for transient behavior in the models.”
We also added to the figure caption:
“For snow cover only one calibration and validation phase was used due to the relatively short temporal coverage of the product, while for glacier variables two calibration and validation periods were used.”

In general we wanted to make use of the full available time series of each glacier to evaluate non-stationarities in the best possible way in the models and the different starting dates across glacier arise due to different starting dates of the field campaigns.
The “-” solely indicates that no second calibration or validation phase for a variable was used. We added a statement to the figure caption:

“A minus indicates that a specific second calibration or validation period was not used for this variable.”
We are also happy to remove the minus completely if this leads to more clarity.
For the WG validation period 2 only 1 summer/high flow season was available while for the others always the full information was accessible, except for WG Validation period 1 (71-81) contained 1 missing year. For the SCG no second validation phase was chosen due to not enough enough data available. In general the discharge periods were chosen to best possibly match the mass balance periods, however, a 100% fit is not possible. We hope this clarifies the table appropriately.

an automatic Multi-Objective Optimisation (MOO)
Answer:
Changed accordingly

simulated binary crossover
I think you should use capital letters only if you want to introduce abbreviation, please check author guidelines to confirm.
Answer:
Confirmed and changed

polynomial mutation
Answer:
Changed accordingly

This is not the case for streamflow where the KGE is used.
Answer:
The reviewer is completely right and it was clarified in the text.

the Normalized Root Mean Square Error
Answer:
Changed

You are using SC and snow cover in the same sentence. It might seem that you're referring to two different things.
Answer:
Changed

same comment here
Answer:
Changed

Do you have any idea of the typical range (e.g. quantiles) of parameter on glacial catchments in order to know which parameter values to expect?
Answer:
Unfortunately not in our exact catchments, however, if desired it would be possible to trace for example degree-day factors used/estimated in literature reviews such as: https://doi.org/10.1016/S0022-1694(00)00249-3 or https://doi.org/10.1016/S0022-1694(03)00257-9 that show similar characteristics or are located in Northern US. Besides, Hock 1999 https://www.google.com/url?q=https://doi.org/10.3189/S0022143000003087&sa=D&source=docs&ust=1734529152030984&usg=AOvVaw2bPXyJBbDYWXvoowSTLtkn provides a range of 4.5 to 7.5.

Maybe add a sentence to support that red are snow parameters, grey lapse rate parameters and blue glacier parameters.
Or, if you prefer, you could add this information to a column in Table 3.
Answer:
Indeed very helpful and we decided to directly add it to the figure caption.

The non-linearity / strong interaction between parameters is very pronounced for Lemon Creek and South Cascade, do you think it could be due to the fact that the measuring station is outside the basin and the parameters are trying to partially compensate for the lack of information? It would be interesting to test this hypothesis by reproducing the analysis with stations outside of Gulkana and Wolverine to see if this shift to the non-linear part also occurs.

Answer:

This is indeed a very interesting question and analysis. We agree that this might be a great additional study that is framed differently and puts more focus on sensitivities of the individual case studies as the focus is to test the model rather than a deep investigation of these uncertainties. Anyway, we believe that the reviewer is right and that this is definitely one of the reasons (the station being further apart and lower elevated). We also argue that this additional study might be better based on synthetic examples that are not affected by measurement errors. The analysis could also involve time-varying sensitivity analysis for example to indicate how parameter/process importance might have changed already. Besides, we added a sentence that underlines the fact that the stations are located outside and might cause this behavior:
"The strong nonlinearity visible for the SCG & LCG might be caused by the fact that the underlying measurement stations are located outside of the catchments at relatively low elevations."

So for the rest of this work, did you then reduce the number of parameters or keep all the possible ones?

Answer:

This is indeed not perfectly described and maybe contra-intuitive as the parameter space was not reduced further, but the SA was used for exploratory purposes only for the newly developed glacier routine and the optimization should contain all snow and glacier parameters in the optimization for a full comprehensive evaluation approach. Also, given the large (expected) interactions of the parameters we wanted to keep the parameters. However, we agree that this has likely little added value and kicking out for example the refreezing rate and others would have been reasonable. We added a short sentence to make this more clear in the article: "However, given the strong interactions of the parameters, the parameter space was kept constant, including the entire space."

NRMSE

Answer:

Changed

I'm not sure to understand why, given the results of the sensitivity analysis presented in the previous section, GLMMX has the highest $\mu*$ values.

Answer:

We assume exactly what the reviewer observed is the reason for this behavior. Given the high sensitivity of GLMFMX it becomes obvious that values at the lower edge seem preferred by the model to achieve low NRMSE values, however, interestingly a decent share of these low values also seem to produce deteriorated results not guaranteeing model improvements. We think as the plot is a snapshot for only 1 of multiple objectives it might be misleading as the GLMFMX VS NRMSE plot of snow cover for example would indicate a significant superiority of small GLMFMX values with respect to NRMSE compared to higher SMFMX values. Moreover, GLMFMX has also the highest sigma values suggesting its strong interacting behavior can lead to inhomogenous performance criteria distributions. So summarized we assume that low GLMFMX values seem to be associated with a well-spread objective space but do not avoid clustering per se. We also tried to clarify it in the text and added:
"The parameter is among the most interactive ones, as shown before, indicating that parameter clustering does not necessarily result in a distinct objective function distribution. As Fig. \ref{fig:opt_fin_pars} only provides a snapshot for the NRMSE of MB, the pattern of the objective function space for SC for example could look differently."

When we look at the median values of the table, we realize that they are often very close to the limits. Is this the behavior that was expected? Does this suggest problems within the set limits? Interactions between parameters too strong?

Answer:

In fact, this was not expected by the authors and we would argue even that it would be interesting to do the experiment with two major changes, 1) using extended parameter limits (as also asked by the reviewer) and 2) replacing hypsometry by total glacier area as an objective. Optionally, it would be of interest to see how

incorporating discharge as an additional objective (or replacing for example hypsometry) would affect the final solutions. It is definitely remarkable how often boundary solutions are preferred and how often the parameters of the best solution for GLMB & SC are far apart from each other. We agree that the strong interaction between glacier and snow parameters plays a significant role, especially as the snow cover estimates based on MODIS already inherently contain a high degree of uncertainty and subjectivity for example through NDSI threshold selection and cloud cover issues (but of course similar things apply also to the glaciological measurements) results could be strongly affected by potential shortcomings. For example could already a small underestimation of snow cover in the shoulder seasons in the MODIS estimates lead to unrealistically high glacier melt & retreat patterns in the model then. Besides, the general computation of snow cover in SWAT (areal-depletion curve) that takes place on the basin scale (rather than the subbasin scale as for other snow parameters) might cause oversimplification (especially in larger basins), as in the SA (Fig. 5) we could see the importance of the SNOCOVMX parameter across all basins (red star) which is suggested to be the most influential snow parameter. Basically the SNOCOVMX parameter controls the interaction to a high degree and given imperfect measurements and snow cover estimates might give the model a tough time. Future releases of SWAT-GL should definitely include spatially distributed depletion-curves. However, it is difficult without further tests to come up with detailed explanations for the behavior for all different examples without further tests, nevertheless, we think that it shows another very nice example of the difficulty to achieve model consistency among different variables without comprising the individual objectives too much (in such catchments). The example demonstrates clearly how important it is to look into as many processes as possible in our hydrological models to be aware of the tradeoffs our models offer in the end.

You did not comment that for GG, LCG and SCG that the the parameter values used to obtain the Best Q are among the worst when you optimize water balance.
This seems a little counter-intuitive for me since I understand previously that the discharge was mainly driven by glaciers.
The same issue appear for WG and SCG with the Best SC.
Answer:
We agree that in general some details were missing and added the following paragraph:
“Looking at the best parameters for different objectives, it is shown that the best SC simulations can deviate significantly from the best MB simulations (in both the difference in the NRMSE of the MB and the parameter values). Large deviations in terms of the NRMSE of the MB can be seen for the SCG and WG (large vertical difference of blue cross and x). With respect to the best final parameters of these two variables, large differences can be observed for GLMLTMP, GLMFMX of the WG (large horizontal distance of blue cross and x), GLMFMN of the SCG and LCG. Moreover, it already becomes apparent that the best discharge simulations do not necessarily coincide with the best simulations of glacier mass balance or snow cover, despite the strong dependency of discharge on snow and glacier melt, which is further discussed in the following chapters”
However, specifically for discharge we kept it short at this stage as we think this will be discussed in detail at later paragraphs in the manuscripts (section 3.5 and 3.6 as well as throughout the discussion) and hope the reviewer agrees that this is the case. If not, we are happy to provide more explanation already at this stage.

The two variables seem inversely correlated, when one is better represented, the other less so, and vice versa. Is this really the case, or is it an artifact? Does it reflect a dependency between some parameters, or some equifinality issues?
Answer:
Indeed this observation is correct and discussed later in more detail. The authors see many different reasons that could contribute to the observation, including the aggregation of the hypsometry objective function, the hypsometry setup (with respect to the ES), equifinality reasons for example caused easily by metrics such as mass balance and snow cover that are based on balances rather than absolute values (infinite solutions can lead to similar MB & SC values and thus to the corresponding OF), bu also the initialization. Also it seems that SWAT-GL might need further tests across different scales in representing mass balance and hypsometry.
The authors later conclude and further explain that another metric, e.g. total glacier area or another setup with a different glacier discretization would be interesting to explore and whether the hypsometry representation in the model could be explored or whether its representation as an objective function might be improved.

The symbol of BestQ is not the same in the legend and on the plot
The fact that the scale is different between basins can be misleading, especially for LCG where, at first glance, the NRMSE value seems to be almost the same between generations, unlike in other basins. I

propose to limit the scale between 0 and 2 (or 1.5) and to indicate the number of points outside this limit to help with the readability, comprehension and analysis of the graph.
Answer:
We agree and changed it to limits of [0.5 to 1.5]

Is "significant" really appropriate? Has a significance test been carried out for all variables? If so, perhaps you should detail further the analysis, and if not, I suggest to use "slight degradations" or "slight variations" since for some catchments and variables the score is better over the validation period.
Answer:
Changed accordingly

To make the table even more easier to read, I suggest adding a sentence saying what is the optimal value for each metric, since they're not the same (NRMSE & PBIAS: 0, KGE: 1).
Answer:
Changed accordingly

This metric should be introduced in section 2.5 "Calibration and Validation Procedure"
Answer:
We introduced the metrics and also tried to modify section 2.5 according to some comments that arise later and also through the other reviewers.

Perhaps you could highlight the range between 10th and 90th quantile of the last generation in darkgrey color to avoid this issue.
Answer:
The idea is very good and we tested it, but had the feeling that the already very full plot got more overloaded and while the added value was visible for the cumulative mass balance, unfortunately for the other two variables in figure 7 it became a little more messy and the percentiles were hard to read. Finally, we decided to stick to the original version unless the reviewer thinks the new version might be beneficial anyway. However, for figure A1 it worked perfectly and we added the percentiles there.

That sounds subjective to me. A "good" RMSE/KGE value for one basin may not be the same for another. This work could benefit greatly from the use of a model as a benchmark to see if improvements are indeed noticeable and if the results are indeed satisfactory or even very good.
Answer:
We fully agree that this was not written objectively enough and also fully agree that without proper benchmarking the numbers can’t be interpreted appropriately. We rephrased the sentences accordingly. Besides, we added the initial performance values to Table 4 in brackets for the calibration period. With initial performance we refer to the starting values of the optimizer that are based on a latin hypercube sampling and could serve somewhat as an indication on what people could expect from just using a random sampling with respect to the quality in representing the different variables. One can now also see that initial discharge results are (although expected) better than after the optimization of the other variables in 3 of 4 cases.

I think this part would benefit from a figure to make it easier to understand and analyze the results
I saw later in the document that there's Table 5 to support your point, so don't forget to refer to it in the text.
Answer:
We thank the reviewer for pointing this out, as we indeed surprisingly did not reference it. This was changed and we think that this also covers the figure comment which is probably not needed anymore.

Here, it would be really interesting to have a comparison between SWAT and SWAT-GL to highlight the advantages (or limitations) of the glacier module for representing streamflow.
Answer:
We internally discussed the comparability options with SWAT standard for a long time during the preparation of the manuscript and had the impression that the comparability at least at this part of the manuscript would have some limitations, basically summarized: Would the comparison be Q stemming from a MB + SC + hypsometry optimized SWAT-GL model with a Q stemming from a SWAT standard model that was then not optimized, only optimized for

Q, only optimized for SC (as it does not contain glaciers)? Furthermore, we think that the incorporation of SWAT standard (which can't produce glacier melt) would lead to a strong overcompensation of SWAT standard by twisting the precipitation (and temperature) lapse rates to add the water to the system. This could maybe lead to a nice discussion to general structural limitations of hydrological models and how unrealistic parameterizations try to compensate for that. For this, a comparison of SWAT-GL Q optimized versus SWAT standard Q optimized might be adequate.

I propose to reformulate to "The PBIAS is larger in the LCG model" if you did not make significance test.
Answer:
We agree and changed it accordingly.

Please clarify which part of the figure are you describing.
I'm confused by the "mean annual flow" used for Column 1, I was thinking it was the streamflow simulations/observation at the daily time-step.
Also I suggest to use "hydrological regime" for Column 2 if I understand correctly what you have done here.
Answer:
We see the confusion and rephrased the sentence to:
"In addition to the daily performance of Table 4, we further evaluate mean annual flows as illustrated in Fig. 8 together with two separate periods of simulated mean daily discharge (averaged discharge for each day of the year over the indicated periods)."
We just wanted to introduce a new paragraph and move from the daily objective function values (as correctly stated by the reviewer) to the annual and seasonal behavior of discharge to see whether inter-annual variations along with seasonal behaviors are represented well. We also clarified the figure caption. Hopefully this sorts out all points from the reviewer.

You did not comment the variation of GG and LCG.
If I understand the figure correctly, SCG, WG and GG seem rather monotonous. However an increase appear for LCG.
Given you showed that the glaciers are bringing more and more water with the melt, which trend seems the most consistent? It's not necessarily easy to answer this question with the increase of evapotranspiration too.
Answer:
We added a description for the GG and LCG accordingly.
The reviewer is completely right and the driver-response relationship is not completely clear, evapotranspiration that affects discharge but also covers ablation directly from the glacier might only be two points, even though summer mass balance tend to get more negative in the catchments they only represent an aggregated value rather than the full picture. Furthermore, LCG weather station data (Juneau Airport) precipitation shows a slightly increasing tendency over time that might contribute to the LCG behavior. Due to the fact that the paper focuses on the applicability of SWAT-GL, a detailed driver - response analysis might be beyond the scope of the paper and would require further investigations.

How do you explain that the streamflow is better represented in the catchment where the glacier module fails?
Answer:
The reviewer hits the mark with his important question, however, a large part of the good result can be explained exactly through the fact that the results would likely be much worse if discharge data would be available from the 90s (1992) onwards as the period from 1962 to 1976 was actually not that badly represented. The important question to raise is also why the second calibration phase of the SCG is that badly represented.

I'm a little surprised that you're now using discharge in your calibration, given that you said earlier that this variable was used exclusively for validation. I find the idea of using discharge interesting, but I think it's hard to explain why you're only using it now and not before.
Answer:

The reviewer is right and we tried to make the SOO (or discharge part) already clearer in the introduction as well as in the materials and methods section (2.5). The idea is to provide kind of an upper limit/benchmark for discharge estimates based on one example to highlight that SWAT-GL is of course being capable to simulate discharge in these catchments. This has two more ideas: first, it should avoid undervaluing SWAT-GL's capabilities in these catchments as we had the impression that without showing readers might derive flawed conclusions with respect to the SWAT-GL's applicability. Second, we want to critically evaluate and demonstrate (tough theoretically widely known but still practically not always followed) that discharge only based calibrations likely come on the expense of highly reduced model consistency. Hopefully the storyline could be made more accessible and clearer to the reviewer.

Is the MOO calculated over MB, SC and Hypso as previously or is it calculated over MB and Q?
In the first case, the comparison seems difficult, and in the second case it would be necessary to specify it, since up to now the discharge has never been used during calibration.
Answer:
The whole first paragraph of section 2.6 was modified and hopefully adds much clarity.
The MOO values refer to the results shown before (so the MOO results based on SC + MB + hypsometry). Then, two new models were set, one for SOO using Q and for SOO using MB. Both SOO models are compared with the original MOO results. Hopefully we also were able to indicate why we think the MOO VS SOO comparison makes sense regardless of the fact Q was not used in the MOO. Especially, given that the assumption that a proper representation of SC + MB + Hyps lead to meaningful Q representations in these catchments did not hold.

Why? To simplify the analysis because all 4 glaciers behave in the same way, or because the results of the others are difficult to analyze? Either way, I think this choice must be explained.
Answer:
As the idea was to highlight the points aforementioned and given the fact that the manuscript is relatively full combined with temporal demand in terms of simulations (computational demand for all tasks) we decided that the general message could be made without necessarily model all other catchments again, which then would increase again the content (both in terms of figures and written text) significantly. Nevertheless, we fully understand that doing the investigation for all glaciers might be more consistent and also interesting to see.

What exactly these parameters do?
In other word, what routing scheme is used (simple lag, lag and rout, muskingum, kinematic wave, ...)? Why did you consider these additional parameters for SOO and not MOO? The comparison between SOO and MOO is therefore no longer fair.
Answer:
We hopefully were able to clarify the questions with the rewritten paragraphs and answers before. However, the authors had in fact several discussion on exactly the point mentioned by the reviewer and finally came to the conclusion that the introduction of these parameters can be justified given the idea to provide a best possible discharge representation (upper benchmark) using a SOO, especially as Q was also not included in the MOO. The introduced parameters should only affect streamflow estimates and would therefore have no or a negligible impact on the MOO which was why they were not included in the MOO at all.

focus on
Answer:
Changed

not easy to read with multiple parentheses, is it possible to use "a and b" instead of "a) and b)"?
Answer:
Changed accordingly

on
Answer:
Changed

If I understand Figure 9 correctly, when calibrating with MB SOO we get similar performance to MOO for NRMSEmb and better performance for KGEq. This shows that it's better to use MB SOO rather than

MOO if we are focusing on MB and Q only. However, this analysis is highly dependent on the variables you used on MOO.
If MOO corresponds to MB + SC + Hypso, why did you not shown SC and Hypso to support the fact that the benefits of MOO appear on the representation of these variables.
If MOO corresponds to MB + Q, I can't understand why MB SOO perform better over the two variables.
Answer:
Exactly, the reviewer is completely right, the best SOO and MOO MB results are comparable (based on NRMSE) in their best values as well as an overall better CDF of the SOO MB. It is also correct that the SOO MB produces better results for Q as the MOO. However, MOO corresponds to SC + MB + Hypsometry as in the previous chapters and we agree that this was a missing element, so we added an appendix figure where we show the large MOO superiority for snow cover over both SOO approaches.

Why the last generation was not used for SOO?
Answer:
The last generation of the SOO would create a vertical line at the optimum value, however, we thought it might be interesting to show how early the optimal values are already found in the optimization process and to show how the distribution at that temporal snapshot looks like.

What is the calibration period used for streamflow? Is it 2002 - 2015 as for MB calibration for WG?
Answer:
It is the combination of the calibration periods (2002 - 2015 and 1971-1981)

Is this totally true in a context where glaciers are melting faster and faster?
Answer:
The reviewer raised a good point and we adapted the sentence to:
“While SWAT applications generally often focus on larger scales compared to the relatively small SCG, more evaluation is necessary on the reasons for the bad performance and whether it might be linked to the small spatial scale. Something which is especially important to investigate given the overall trend of shrinking glaciers across the world.”

I agree, but this needs to be specified in section 3.6.
Answer:
Done

To what extent?
Because I have the impression that you have shown in this document that the representation of MB, SC and Hypso are not enough on their own to simulate the discharge (especially in part 3.6., KGE of 0.64 vs. 0.9).
This raises 2 questions for me:
- the KGE values that could be expected if we had chosen MOO composed of MB, SC, Hypso and Q
- the differences between MOO and Q SOO in terms of dynamics, and particularly in terms of seasonal dynamics (are flows during the melt period well represented, for example?).
Answer:
We completely agree and also want to highlight that also the good results for Q in its SOO came along with a huge negative MB (not shown in the manuscript) which leads to a unrealistically high and fast glacier retreat and this underlines the danger of SOO approaches in these environments. This overcompensation in water somewhat indicates that missing water for discharge was a result of the MOO, as MB is only a balance of course similar MB estimates could likely be reached while increasing precipitation lapse rates and therefore more water. In general it is great to demonstrate the complexity of these systems and how things should be handled and interpreted with caution. Absolute values-based variables such as SWE would be very interesting to explore as using snow cover in the calibration is far from guaranteeing reasonable solid precipitation or SWE estimates.
These are indeed interesting questions which we are currently evaluating further, also incorporate climate projections to demonstrate how large and far from each other hydrological projections can be based on these strategies. Besides, this was absolutely the intention and one of the key messages that representing the chosen snow and glacier variables are neither a sufficient nor a necessary condition.

I agree, meteorological forcings are particularly impacting on mountain basins and generally very complex to estimate. See e.g.:
Evin, G., Le Lay, M., Fouchier, C., Penot, D., Colleoni, F., Mas, A., Garambois, P.-A., & Laurantin, O. (2024). Evaluation of hydrological models on small mountainous catchments: Impact of the meteorological forcings. Hydrology and Earth System Sciences, 28(1), 261–281. https://doi.org/10.5194/hess-28-261-2024

Answer:

We thank the reviewer for that very interesting publication and we are working on a publication in Central Asia where we demonstrate the effects of all these various precipitation products in the universe on the water balance of 60 to 70 large scale basins (also considering ET products). The study provided was incorporated in the manuscript.

Appendices must be reordered in the order in which they appear in the document

Answer:

Done

Are the dotted lines for GC and WG confused with the solid lines, or has the deviation only been applied to LCG and SCG?

Answer:

Indeed it could only be done for the SCG and LCG where the measurements started earlier than those of the GG and WG as the intention was to show both the cumulative MB from the same starting point (earliest intersection date of observations) and the full individual cumulative MB lines of each glacier.

Reviewer 2 (Anonymous - Major Revision)

General comments

This paper presents a thorough sensitivity analysis and calibration for the newly developed SWAT-GL model, evaluated across four glaciated catchments. The findings are valuable for the broader scientific community, including those studying glaciers and using SWAT models. While the methodology and case study contents are clear, the results analysis are not well structured and disjoint. The manuscript needs improvements in writing. It is lengthy and difficult to read. I suggest the authors shorten it and consider using an English editing service. Below are detailed comments.

We thank the reviewer for the helpful suggestions and feedback and agree that the results needed to be better presented and were hopefully able to provide a much improved and polished version than before that satisfies the reviewers expectations.

Major comments

1. Section 2.4 (Line 240): It would be beneficial to explain why the analysis exhibits a monotonous pattern. Additionally, in Section 3.1, clarify why values of 0, 0.25, and 0.5 correspond to linear, monotonous, and nearly monotonous behavior, respectively. Please include references if applicable.

   Answer:

   The reviewer is completely right and we added the references to both chapters in which the details can be accessed:

   Merchán-Rivera, P., Geist, A., Disse, M., Huang, J., & Chiogna, G. (2022). A Bayesian framework to assess and create risk maps of groundwater flooding. In Journal of Hydrology (Vol. 610, p. 127797). Elsevier BV. https://doi.org/10.1016/j.jhydrol.2022.127797

   and

   Garcia Sanchez, D., Lacarrière, B., Musy, M., & Bourges, B. (2014). Application of sensitivity analysis in building energy simulations: Combining first- and second-order elementary effects methods. In Energy and Buildings (Vol. 68, pp. 741–750). Elsevier BV. https://doi.org/10.1016/j.enbuild.2012.08.048

   Especially in Garcia et al. 2014 it is shown that if all EE of an input factor have the same sign, the factor has a monotonic effect on the model response (consistently increasing or decreasing depending on the sign). Using the sigma/mu* ratio it is shown that ratios smaller than 0.1 factors have mostly linear effect (assuming normality - 95%). Besides, it is shown that if the ratio is smaller 0.5 most EE have similar signs and are thus monotonic. Nonlinearity stems from the relation between sigma/mu* to sigma/abs(mu) where nonlinear effects are clearly visible for sigma/abs(mu) > 1. Indeed there was a typo and the slopes are logically 1, 0.5, 0.1 which makes it hopefully much more logical.

2. Section 3: Numerous aspects of optimization are analyzed in this section. I recommend starting Section 3 with an overview of the different optimization aspects being investigated. Clearly state the goals of each aspect and explain how they complement each other in the results analysis. Without this context, the structure of the subsections appears disjointed.

   Answer:

   We thank the reviewer for this suggestion and acknowledge that the different sections might have appeared disjoint/confusing. As suggested, we therefore tried to put more clarity with an overview paragraph at the beginning of section 3. In addition, we hope that the consideration of the comments of all reviewers leads to further clarity and structure throughout the manuscript.

Minor Comments

1. Figure 1: Could you add state/province names to the middle map? This would help readers identify regions, such as US states or Canadian provinces.

   Answer:

   We updated the figure accordingly.

2. Figures 2 and Onward: Please add abbreviations next to the glacier names in all figures and tables to help readers link the glacier names in the figures/tables with their abbreviations in the text.

   Answer:

   We added the abbreviations for the first figures.

3. Figure 6: Clarify how the three blue symbols correspond to the best solutions for mass balance, snow cover, and discharge.

   Answer:

   We clarified this point in the figure caption

4. Table 4: Please specify how KGE and PBIAS are calculated, or add references for these metrics.

   Answer:

   We introduced both now with more appropriate explanations why and when they are used and how they are calculated in 2.5 and hopefully could solve the reviewer comment

5. Line 370: Add the parameter symbol after its name to help readers link the parameter to Figure 6.

   Answer:

   We thank the reviewer for this good idea and added the symbol to the first usage of the parameter in the text.

6. Line 433: Use consistent formatting—should this be "Table 2" instead of "Tab. 2"?

   Answer:

   We corrected it throughout the document and thank the reviewer for the hint

7. Line 437: Include references to support the statement "results often considered satisfactory in hydrological studies."

   Answer:

   We clarified it and added Moriasi 2007 and 2015 as reference

8. Line 445: Is the "most significant" conclusion based on a statistical significance test? Please clarify.

   Answer:

   In line with other reviewer comments we changed the wording accordingly as no significance test was conducted.

9. Lines 473-475: Please check the brackets in these lines for errors.

   Answer:

   Was changed in order to avoid brackets within brackets, we stick just the variables “a” to “d” without a bracket

   Some sentences require grammar correction or rephrasing for clarity:

- Lines 360-361

Answer:

Done

- Lines 466-467

Answer:

Done

- Lines 500-501

Answer:

Done

- Lines 543-544

Answer:

Done

10. Lines 299-322: Consider shortening or breaking this paragraph into smaller sections to improve readability and focus

    Answer:

    We completely agree and substantially revised section 2.5 also including multiple subsections to make the individual components more clear. Hopefully, the methodology has much improved in readability and therefore clarity.

Reviewer 3 (Dipti Tiwari - Minor Revision)

Comments

The authors have invested significant effort into this work, making it an interesting read. The recently released SWAT-GL model effectively addresses key limitations of the traditional SWAT model in glaciated mountainous catchments. Efforts to expand its applicability in such environments are significant. The inclusion of a glacier mass balance module, based on the degree-day approach, and the delta-h (Δh) parameterization for dynamic glacier retreat represents a valuable improvement. Benchmarking the model against USGS Benchmark Glacier Project glaciers is a decent approach.

Certain aspects of this article would benefit from clarification and adjustment. More clarity in the methodology and case study explanations is needed. The results section lacks coherence, affecting the storytelling. The article also reads as overly lengthy, which complicates readability. The article would benefit from being more concise by removing sections that do not directly contribute to the objective. The work is fantastic! Incorporating the suggestions, where appropriate, will enhance the article and improve the clarity of its contributions. Detailed comments are provided below for further guidance.
Answer:

We thank the reviewer, Dipti Tiwari, for the great feedback completely agree that the paper was lacking clarity and might be hard to read and thank the reviewer for pointing this out and also for valuing the work. We tried to work on the suggestions and think that the revised version has significantly improved in readability and structure. Given the comments from the other reviewers, it was however hard to reduce the overall length of the paper.

General Suggestions:
The article uses certain words excessively or without appropriate context, leading to disjointed sentences. Terms such as "however," "in addition to," "in summary," "in contrast," "in detail," "in general," "as mentioned," "moreover," "furthermore," "although," "is not beaten by," and "in greater detail" often serve as sentence starters that do not connect with the preceding sentences. This disrupts the flow of the text and create confusion, as these phrases appear without sufficient context or relevance to the main text. It is advisable to remove these terms or use them where appropriate.
Answer:

We agree that there was space for improvements and hopefully the completely revised version has improved in readability.

Figures:
It is suggested that the figures be adjusted to use the same x-axis and y-axis extents for comparison. Having different scales on the axes makes it difficult to compare the graphs effectively. Utilizing consistent x-axis and y-axis ranges will enhance clarity and facilitate a more accurate interpretation of the data. For example, Figure 5
Answer:

We agree that shared axes are generally favorable and tried to align them where possible, however, given the different scales of the variables between the different catchments for example in Fig. 7, matching axes limits reduced the clarity and readability unfortunately. Thus, it was not always possible to incorporate shared axes. However, figures such as Fig. 5 were updated accordingly.

Section 3.4:
This section discusses the MK-Test, WRS and Pettitt test; however, the purpose of conducting these statistical tests on observed and simulated values is unclear. It would be helpful to explain their necessity and whether they provide meaningful insights or are merely an application of statistical tools. Table 5 contains many 0 and 1 values for all glaciers for the Pettitt test, MK test, and WRS. Clarifying what these values represent—such as p-values, h-values, or K statistics—and why they are presented as integers would be beneficial. Much information seems to be missing from the table, making it difficult to derive any logical conclusions.

Answer:

We agree and added further formulations already to the methodology section. The idea is to test whether SWAT-GL works well under nonstationary conditions which are present through the long simulation runs and increasing temperatures in that period to which all catchments are sensitive to. As the manuscript was completely revised we assume that the general readability has improved including the mentioned section.

Section 3.6:

Discharge is now being incorporated in the calibration, despite earlier statements indicating that this variable was used exclusively for validation. What is the rationale behind this? While using discharge is an interesting idea, it raises questions about why it wasn't included from the beginning.

Answer:

We made the statements clearer or removed them when they were misleading throughout the manuscript. Besides, we added a short overview paragraph at the beginning of section 3 that hopefully provides guidance to the reader and reduces unclarity.

Minor corrections:

Line 76: sentence incomplete. “A trend which is …….

Answer:

Corrected

Table 1: The row “All glaciological Basin mean ……” is not clear.

Answer:

Corrected (it was meant all catchments)

Line 97: “Homogeneous data processing methods ……..” maybe mention method names and reference

Answer:

We fear that this would incorporate a longer description here, whereby we refer to the studies of O’Neel.

Figure 1: For me personally the figure legends are not up to the mark. Use star for watershed legend instead of square black box. Maybe place them horizontally instead of fitting them inside the figure.

Answer:

Figure 1 was modified in a way to match all reviewers expectations

Line 171: The author mentioned “years from 1972-1992 had to be chosen”. Please explain the reason.

Answer:

Done

Line 172: The author mentioned “we can see a tendency” mention where? Refer to table or figure.

Answer:

Done

Line 252 – Line 263: This paragraph needs to be rewritten properly.

Answer:

Rephrased based on further comments of other reviewers

Line 274: This line is not clear.

Answer:

Done

Line 324: Check if the sentence is correct.

Answer:

We slightly rephrased.

Line 338: Why GLMFMX is substituted by TLAPS, any justification?
Answer:

Substituted with respect that the order of importance has changed

Figure 5: Different X and Y axes make comparison difficult. It is suggested to keep the axes consistent for all four figures and divide them into four quadrants based on any specific μ* and σ values. This will simplify the identification of which quadrant represents the more sensitive parameters.
Answer:

We agree that generally shared axes limits are preferred and updated the figure accordingly.

Line 343 – Line 347: Needs to be rephrased. Mention names of four most important factor's name
Answer:

Done

Line 349: The statement is not, right? Please verify.
Answer:

We checked the statement again and found that in fact the SNOCOVMX parameter is in two out of four cases among the 4 most sensitive parameters, in the two other cases it is ranked 5th and in the GG it is even the 2nd most influential parameter, which also aligns with its strong interaction with the glacier routine. However, we rephrased it slightly.

Line 369: "An exception exhibits here ……" Not clear
Answer:

We added that we mean clusters are created of parameter values

Line 410- Line 411: These lines can be removed as they are already covered in the previous section.
Answer:

Rephrased to make a different point

Line 426- Line 429: These lines can be removed as they are already covered in the previous section.
Answer:

Rephrased rather than deleted to make a different point

Line 507- Line 514: These lines need to be rewritten.
Answer:

We have rewritten the paragraph