Response to the comments from Gregoire Guillet, reviewer #3.

The manuscript proposed by Beraud et al. presents a new methodology to filter outliers and estimate glacier surface elevation from dense digital elevation model time series.
Their study relies on the same data as the Hugonnet et al (2021) study.
Instead of Gaussian Process Regression, Beraud et al implement an additional LOWESS filtering step, before feeding the fitlered time series to a localized b-splines scheme in order to interpolate surface elevation measurements.
While the work and the methods shows promise, I strongly recommend that the authors take the necessary steps to strengthen the manuscript.
The paper appears to be in an early stage of development and would benefit from significant revisions before it is considered for publication: sections 4 and 5 are particularly challenging to understand. With the necessary improvements, I am confident the paper has the potential to make a valuable contribution to the field.
I recognize, from first hand experience, the amount of effort that went into this work. This is why I want to restate my strong support for the manuscript as I think it brings an important contribution to the field.
I strongly encourage the authors to address the points I have highlighted, as doing so will enhance the clarity, rigor, and overall impact of the work.

Greg Guillet, University of Oslo

We thank Gregoire Guillet for their careful reading, their comments and help in improving our manuscript.
We considered all the comments. The minor ones have been accepted and do not appear in this response document. Please find below our answer and larger changes in response to other comments. Our answers are in blue.

<u>Answers to general comments</u>

- From this version of the manuscript it is not clear at all to me how uncertainties are calculated, nor what do they actually represent. More details on this throughout the manuscript are severely needed.

  We understand the concern of the reviewer and acknowledge that our manuscript was not hundred percent clear regarding uncertainties. Below we detail a number of uncertainties that we clarified in the revised manuscript.

  - On the figure 6, 3 and 9, we show confidence intervals that are those of the interpolation algorithms (GP Regression and ALPS) and not of the overall workflow. We do not think that the theoretical uncertainty and confidence interval predicted by the ALPS-REML are credible to assess the results of the overall workflow. For instance, Figure A2.a (in appendix) shows well that the filter can remove some critical portions of the surge signal (creating data gaps that could also exist in the original dataset) without the interpolation uncertainty to represent it. A visual verification of the process permits to detect large biases, but does not give any metric value. This is why we develop a lot the sensitivity analysis, the comparison with external reference DEMs or reference study on a number of events. It permits to give some sense of the uncertainty, of the ability and weaknesses of the method.
  - Still, for the revised version of our manuscript, we will give some uncertainties on the surge volumes (e.g. in Table 1). These uncertainties are based on standard methodologies based on the elevation change from stable terrain used as a proxy to estimate the uncertainty on moving terrain (Hugonnet et al., 2022). The first estimations are very conservative and gives uncertainties that, on a median basis, are about 55% of the estimated volume of ice transferred (with volumes of Table 1 of the manuscript; from 19% to 280%). We will try more developed uncertainty calculations and complete the revised version accordingly.

  L360-370:

Be a bit more specific. Please add uncertainty estimates. Mention that differences you show are between median values
L385: “The order of magnitude of the imbalances corresponds to the order of magnitude of the measurement uncertainty”
Can't agree more ! Please add them.
Tables: Please add uncertainty estimates in all tables - as a hunch feeling I would typically assume that this is what drive the reported imbalance (except for Hispar).

Our answer to the general comment covers these three minor comments above. We will give some uncertainty estimates in the revised version, reminding it does not include all kind of bias (e.g. for the wrong processing of the Khurdopin surge).

- Section 4 needs a vocabulary overhaul. The authors use very unspecific language which makes it harder to grasp what is the point they are trying to make. I have addressed some of these in my specific comments.
  As it stands, the writing in Section 5 makes it very hard to understand. I had to re-read some sentences multiple time to make sure understood the statements correctly. I think significant efforts are needed to
  A lot of the sentences start with unspecific pronouns such as "this" or "it" - which required me to backtrack a few times. This severely hampers the readability of the manuscript.
  Acronyms are not always defined on their first use - please check this.
  Finally, there are quite a lot of frenchisms and typically french sentence constructions - it's not a major problem, but the manuscript would gain in readability if the authors addressed this.

  The overall article and these sections particularly will be proofread and improved for the revised version. The specific comments are addressed in this answer.

- There are some problematic statements which show some confusion between GPs and splines. I think these can and should be addressed with minimal changes by removing the statements I have highlighted in my specific comments. In addition, the authors need to reframe comparisons between methods in the manuscript as such : comparisons between the method proposed by the authors and the GP from Hugonnet et al. (2021).

  Regarding the confusion between GPs and splines, we answer directly to the specific comment below in this document.
  About clarifying the different comparisons: we completed the end of the introduction with “We evaluate the performance of the workflow compared to the results of Hugonnet et al. (2021). We also compare the surge characteristics such as volumes transferred to other products and studies *[N.B.: studies of the literature, that worked on the case studies of a few surge events]*.”. We also already insist on the fact that we compare a lot our workflow to the GP regressions <u>as implemented by Hugonnet et al. (2021)</u>. The first paragraph of the discussion subsection 5.4 Methodological Insights and Modifications states that our comparison does not apply to all possible settings of GPR.

<u>**Answers to specific comments**</u>

- “We compare the produced dataset to previous studies […]”.
  If I am not mistaken, you only compared it to the Hugonnet et al. (2021) results. I would hence keep it singular here.

  With this sentence, we also refer to the work done to compare the four surge events we analyse with the dataset produced. In the subsection 5.2 Discussion / Comparison of surge characteristics with the literature, we compare dates, volume transferred etc. against other studies (e.g., Bhambri et al. (2022), Steiner et al. (2018), Gao et al. (2024)...).

- I generally agree with all the statements made in the introduction and they are accurate. However, as it stands, I find the structure of this introductory section quite confusing.
  The authors often switch between very broad statements to nichely precise descriptions of methods/problems without clear guidance of the reader into why this matter, which ultimately leads to significant repetition.

[...]

We reorganised the introduction, with a few minor changes of content.

- Paragraph starting at line 52 - May I suggest adding my own work - Guillet and Bolch, 2023 - where we develop a Bayesian outlier filtering and uncertainty quantification framework to compute thickness changes from DEMs, specifically for surge-type contexts.

  Thanks for this suggestion. We included this work by these few lines : “A recent study has exploited a Bayesian framework by inference applied to elevation change to filter outliers, which requires prior knowledge from diverse sources (Guillet and Bolch, 2023). It has been tested on surge-type glaciers, and it applies equally to dense time series.”.

- Paragraph starting at line 68, I would suggest to bolster this paragraph and be a bit less succinct - you have done a lot of work here and this is a good place to showcase a short summary.

  We completed 3 sentences to add a few details. This now reads as “In this study, we present a workflow designed to filter and interpolate elevation time series of high temporal resolution during surge events. We use established algorithms to filter outliers and interpolate elevations at monthly scale while preserving surge elevation signals. We apply it to an unfiltered ASTER DEM dataset from Hugonnet et al. (2021). We produce a regional dataset in the Karakoram region covering more than 100 surge-type glaciers. We evaluate the performance of the workflow compared to the results of Hugonnet et al. (2021). We also compare the surge characteristics such as volumes transferred to other products and studies.”

- L92 : The temporal sampling is heterogeneous in time and space. Rephrase.

  This now reads as “The temporal sampling is not regular in time, and parts of the mountain range have about twice less DEMs than others”.

- It took me 2 reads of these sentences to understand what was clearly meant. First, "below 50 days apart" should be "less than 50 days apart", but a more general question is whether both sentences are needed. I feel they give redundant information, that is illustrated in Figure 2. I would rephrase to avoid this repetition and clarify the statement.

  We removed the first sentence.

- Is this 400m threshold arbitrary ? It's not a problem if it is - I am wondering if there is a reference to back this up or if its from the author's personal experience with the data.
  Also, the notion of "spatial" filter is a bit confusing as you are filtering pixels that show and absolute difference in elevation (z coordinate) between the ASTER and GLO-90 DEMs, correct ? Or are you operating in the x, y plane ?

  Yes the 400 m threshold is arbitrary. Some surges with up to 200 m elevation change have been observed and we do not know any observation with more than this value, our threshold is thus quite conservative.
  The name is indeed not appropriate, thanks, we now call it “absolute filter”.

- This will probably reveal my total lack of knowledge on DEM generation methods but, is the correlation score a sufficient enough metric to discard data ?
  Or is there value in computing a median DEM from all the available DEMs for this day and using this instead ?

  The reviewer is correct that the correlation score is not always the best metric to exclude some data, we write in the subsection 4.1 Results / Performance of the outlier filtering: “This is often due to scattered elevations and to the correlation score that is not very representative of the actual pixel quality: outliers may have lower uncertainties than more accurate observations (e.g., Fig. A2.e).”. Here below is the figure (Fig. R3.1), showing some points of various elevation error (representative of the correlation scores for a pixel time series) mixed altogether :

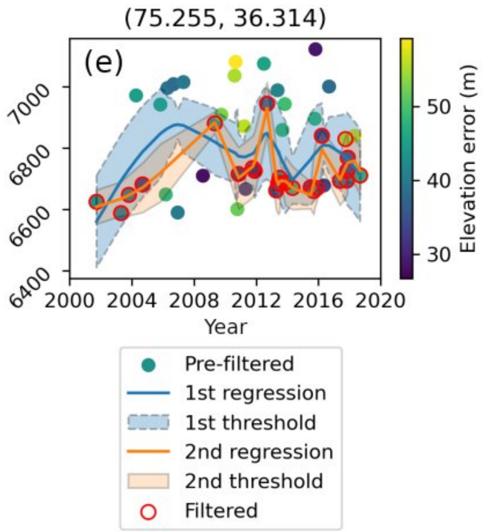

Fig. R3.1. LOWESS workflow over a sample time series.

About merging stripes based on the maximum correlation score versus a median of values, yes both are possible. A few specific cases where in favour of using correlation score. However, overlaps represents a small portion of the data, so the impact should be very limited.

- L111-117 This could be condensed a bit.

  We removed 2 sentences out of these 6.

- L 125-126 : "For our dataset, the output of the regression is to sensitive to noise overall and too smooth over surges to be used directly as an interpolation of the elevation, so we use it for filtering only."
  This is a pretty important point which I feel is a bit brushed over by the authors here, as it is pretty well documented that LOWESS struggles in non-stationary contexts.

  We added the sentence "indeed, regressions struggles in non-stationary contexts (Baumohl and Lyocsa, 2009)."

- L165-170. I have to say I strongly disagree with this statement and I think there is a bit of confusion that needs to be addressed.
  While …

  We thank the reviewer for their explanations, and acknowledge that our text was not correct about these methods. My earlier view on B-splines was simplistic with a "user" point of view and I was lacking the theoretical background. Although we keep a very succinct level compared to reviewer's explanations, the modified paragraph now reads as:

  "We compare Gaussian Process Regression (GP regression) from Hugonnet et al. (2021) and ALPS-REML in our study. GP regression, equivalent to kriging, is a non-parametric method that relies on estimating the data covariance to provide an optimized interpolator. Under certain assumptions, including notably second-order stationarity, GP regression has been shown to be the "best linear unbiased predictor". It is the method used by Hugonnet et al. (2021) on this same dataset, to compute long-term mass balance estimations worldwide. We use a GP covariance with terms estimated in Hugonnet et al. (2021) through a global variogram analysis. This analysis identified several kernel components (periodic, local, linear...), that are not specifically tuned for surges. B-splines and thus ALPS, on the opposite, approximate the data with polynomials under the assumption of a degree of smoothness of the data, with no need for us to inform the behaviour of the data. Although both GP regression and ALPS need domain knowledge to decide the covariance kernel and spline degree/penalty respectively, from a user's perspective using GPs is harder owing to the well studied difficulty of finding the right kernel (Pu, 2024). For ALPS on the other hand, we just need to choose the right degree and penalty order from a small set of choices, and that results in a more stable fit in our case.".

- L188-189: Modifications of this kernel to allow for stronger changes in elevation have not proven to be efficient enough I would be a bit more careful and specific here, as I know of successful attempts at this with different GP kernels.

This now reads as “Reparametrization of the kernel used by Hugonnet et al. (2021) gave slightly worse results than those obtained with the ALPS-REML method.”. Indeed, the reparametrization gave fairly satisfactory results, and it is likely that a better-defined kernel could even outperform the results of the ALPS-REML method.

- L 190-192: It does conserve nearly all known surge events in our study area and period, with one exception being surge events with strong melt before and after the surge When you say study area and period, you mean the 4 surges you look at, correct ? […]

  It also applies to a broad number of surge events in the Karakoram, although we did not check manually each surge-type glacier. We rephrased our statement this way : “It conserves well the surge signal of 3 out of the 4 events we analyse in subsection 4.3, and this observation seems to extend to a number of surge events in Karakoram. One exception is the situation of surge events with strong melt before and after the surge.”. Another case of failure for example is at the front of the Aktash glacier (RGI2000-v7.0-G-14-18524).

- L197-203 This whole section is confusing. […]
  What I get from this section is that your method works better than Hugonnet et al (2021) on-glacier, in parts with that are relatively smooth, but tend to over-filter in areas of low contrast/rough terrain - am I correct ? In any case please make the section easier to understand.

  We applied the minor changes suggested and we simplified some parts of the subsection. We added the following sentence at the beginning of the last paragraph : “To summarize, our filter permits to preserve better the surge signals that were filtered out in the workflow of Hugonnet et al. (2021). However, the new filter is more noise-sensitive over textureless accumulation areas and rough terrain, leading to data gaps or artifacts with large elevation changes.”.

- L205 Again, this is confusing. Please rephrase.

  This sentence and the one at the end of the section 2 “Data” have been thought through quite much during the writing process already, without any simple solution. We could reason by probability, which we did not keep: “Any random date in the time series period have 40% (75%, respectively) probability to fall between unfiltered observations less than a year (two years, respectively)”. We will consider the thoughts of the reviewers on this on a possible second round of review.

  As a reminder, the current version is “After filtering, nearly 40% (75%, respectively) of any date in the time series periods are between unfiltered observations less than a year (two years, respectively)”.

- L209-212; An interesting point is that the GP used in Hugonnet et al (2021) shows an increasing trend in surface elevation, completely omitting the actual data. GPs tend to "fall-back" to the median when there is no data but here, both the median and the uncertainty increase. Can you plot the uncertainty of each measurement ?

  We show on the figure below (Fig. R3.3) the time series coloured by the elevation error (in m), which is estimated with the workflow from Hugonnet et al. (2021). We added the sentence “It is noteworthy to mention that by its design, the original kernel is optimized to preserve a linear trend to extrapolate out of the observation period of each pixel.”.

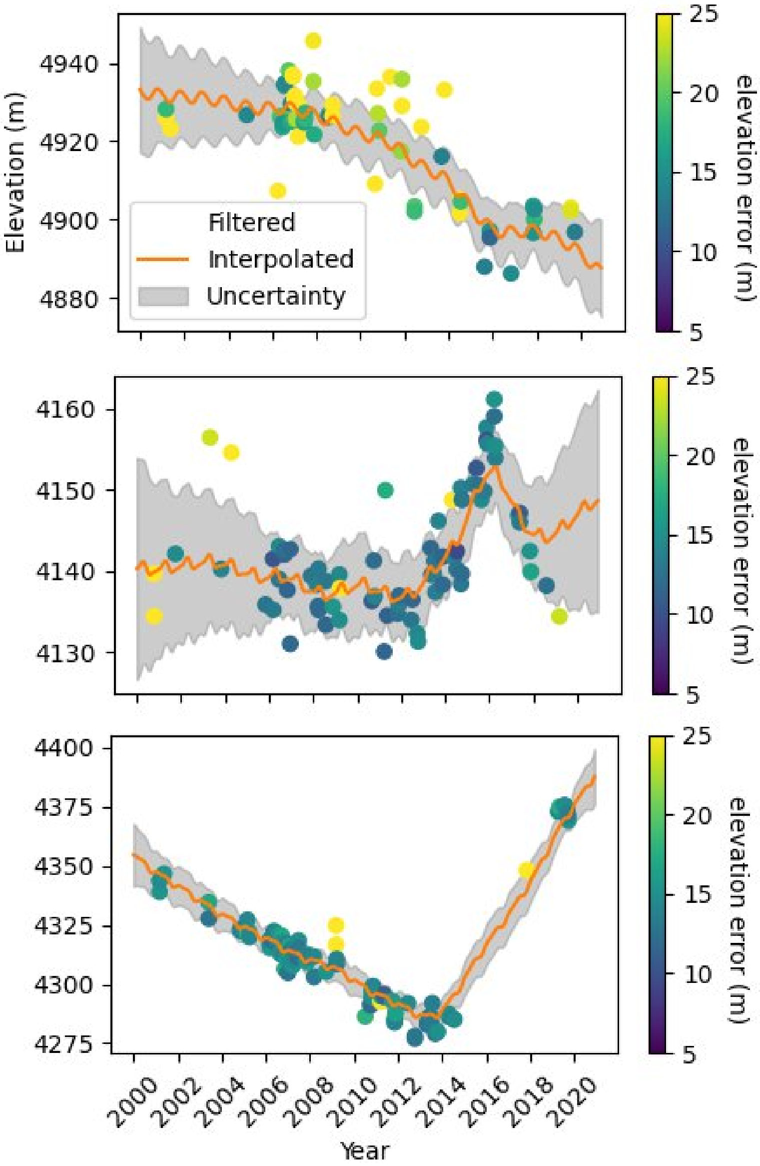

Fig. R3.3. Time series with the interpolation of Hugonnet et al. (2021) and the per-pixel elevation error estimates.

- L215: "[...] creating wavelet artifacts", The term wavelet design something different in signal processing and I would refrain from using it here. I would use "spurious high frequency oscillations" or something similar.

  We thank the reviewer for the appropriate suggestion, we replaced our old term with it.

- L218: have weakest changes, Weaker. But please consider changing to a more specific term.

  We changed to “show smaller elevation changes”

- L220-221: “Some glaciers are more affected by data gaps than others, in agreement with areas with a low number of observations (Fig. 1, e.g. Shisper glacier)”, Shisper is not highlighted on Figure 1 and is not in the studied glaciers.

  Showing Shisper glacier was not intended, it was just an additional piece of information because our dataset did not allow us to add its surge to our panel due to large data gaps, even though it is a well-studied case. For clarity, we removed the mention of Shisper glacier.

- L227: “with a decreasing speed (2009-2012, a1)”, This figure does not show the velocity. Also, the reference to the figure is broken.

We here talk of the surge front propagation, which is visible on the Hovmoller diagram, we do not talk about ice velocity. To avoid further confusion, we replaced “speed” by “rate of propagation”. The reference to the figure has been corrected.

- L246: Add uncertainty estimates - at least some part of the discrepancy is in there. I imagine there is an underestimation in the surface area of the reservoir zone ? Did you account that the surge also simultaneously drains the northern (Yutmaru?) tributary (centerline RGI2000-v7.0-L-14- 27499) ?

  As discussed previously, we do not provide uncertainty on interpolated elevations (and thus on volume transfer estimates) yet, but we will add some in the revised version. A first conservative estimate of the the uncertainties for the transferred volume are : -2411 ±1290 $.10^6$ m^3 (reservoir area) and 3110 ±605 $.10^6$ m^3 (receiving area).
  Indeed some parts of the discrepancy should lies in the area delineations, although the choice of dates could have an even greater impact (we state that “By mid-2018 [end term of the DEM difference] our imbalance is equilibrated”).
  Yes we include the northern/Yutmaru tributary. Here below (Fig. R3.4) we show the elevation map with the distinct reservoir and receiving areas (in black) we used for calculating the surge volume. We have a similar imbalance when we compute it at the scale of the glacier system (the whole red outline, including Kunyang tributary).

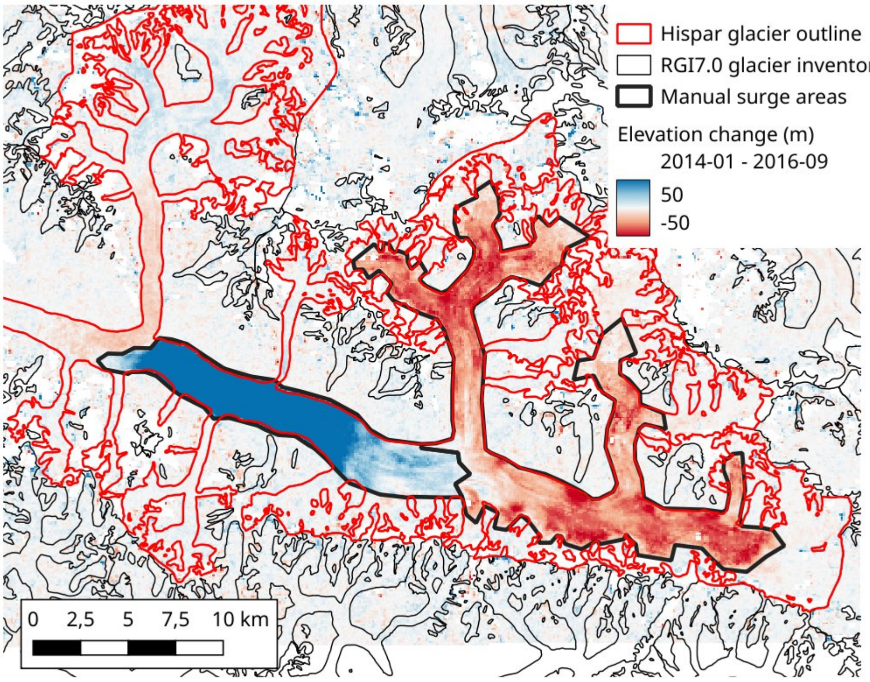

  Fig. R3.4. Map of the elevation change over the Hispar 2014-2016 surge and boundaries used for our calculation of transferred volume.

- L253 : “The "build-up front" or kinematic wave”; To stay consistent with current terminology I would use surge front, not build-up front or kinematic wave.
  L337: “pre-surge thickening front or kinematic wave”; Same as before. The propagation of a thickening front is one of the definitions of a surge. In your case it is still the early stage and has not reached the dramatic proportions it will eventually attain. Stop using kinematic wave.

  Thanks for rising this point. It is indeed different from a kinematic wave, although for example Turrin et al. (2013) use both terms ambiguously in a similar situation. We modified extensively several paragraphs of this section for the sake of clarification and revised the terminology. Here is the corrected version:
  <u>Results</u>: “Khurdopin glacier has a strong build-up signal until the full surge onset. The lower limit of its reservoir area, or dynamic balance line, propagates downward as a surge front during the build-up phase. It is visible as an area of positive elevation change trend slowly propagating down-glacier (Fig. 8.b, area b1). The surge front extends from about 27 km of the glacier head in 2002 to about 33 km in 2015, representing a regular

advance of about 460 m per year, which is approximately 6 times faster than the surface velocity after the front passage, according to the NASA MEaSUREs ITS_LIVE project repository (Gardner et al., 2022). During this period, we do not observe a clear mass transfer from an upper reservoir area , which seems thus different from a slow surge onset”

Discussion: “We now discuss the recent surge of Khurdopin glacier. The propagation of the surge front during build-up has not been observed on this glacier, to our knowledge. The existence of kinematic waves or surge fronts that propagate the surge instability have regularly been observed on other surges (e.g., Cuffey et al., 2010; Kotlyakov et al., 2018; Turrin et al., 2013), with unclear definition of the phenomena. For the Khurdopin glacier, the mechanism seems different from both a kinematic wave or a slow surge onset, with constant thickening after the passage of the surge front during build-up showing the extension of the reservoir area and no upper reservoir area drained. Turrin et al. (2013) observed with velocity data the propagation of a surge front several years before the Bering glacier surge, triggered consecutively to the passage of the activation front down the reservoir area. The activation front for Khurdopin also propagated faster than the surface velocity.”

L383: “bulge front”
Just use bulge. Also, as mentioned before, it's not pre-surge.

The sentence have been changed to “or the propagation of a surge front during the build-up phase preceding the Khurdopin surge” after the main comment of the reviewer.

- L254-255: “representing a regular advance of about 460 m per year, which is approximately 6 times faster than the surface velocity, according to the NASA MEaSUREs ITS_LIVE project repository” I am not sure I get what you mean here - the surface velocity data at Khurdopin clearly shows seasonal behavior with velocities reaching around 400-450 $m * yr^{-1}$, starting in 2013 with a quasi-linear increase in velocity up to 2017. Although I might be wrong, i would expect you to be able to see that the surge front advance rate is slower between 2000-2012 than 2012-2017 when the glacier slowly starts to shift to a velocity weakening regime.

  We used annual velocities (much slower) for this estimation, and we did not wish our analysis to go deeper on this topic, although indeed our hovmoller diagram also seems to point toward an acceleration after 2011 at first glance. For the revised version of our manuscript, we will revise this analysis, add details or clarify it.

- L275-76: “The buildup and emptying of the first surge seems weaker than the second one, and extends less up-glacier of the junction, compared to the second surge” Again, refrain from using weaker as it gives the false idea that the surge did not dissipate as much energy - something we have no idea on. The peak velocities of both of Yazghil surges are actually pretty similar and both are visible up to the glacier front in the surface velocity record.
  L277-278: “This may be related to the effect of the tributary surge, that stopped at the junction but could have yet increased mass input by a blocking effect.”, I really don't get what you mean by that, please explain.

  We will revise our analysis and statements in the revised version of the manuscript. We may remove these sentences if not supported enough.

- This whole section is very confusing. I do not understand what the first sentence is supposed to mean, how can an uncertainty estimate over a quantity reflect the filtering capabilities of a filter ?

  This should be more clear after rephrasing : “The uncertainty estimate of the ALPS-REML algorithm, which is represented in the figures, does not represent the uncertainty of the whole workflow. The performance of the filter is not taken into account in such uncertainty, although it has a major impact on the result”

What does it mean that the surge of Khurdopin shows that "that a discrepancy of a hundred meters is credible on exceptional events." ?

It means that it is possible to reach an elevation estimate error of 100 m on some dates of a surge, as it is the case of the Khurdopin surge. We give an order of magnitude because it can occur on other glaciers with other values. We clarified the sentence : “The case study of Khurdopin glacier surge shows that a wrong estimate of a hundred meters of our workflow is credible on exceptional events and at precise dates during the surge (Fig. S2.a in Supplement, in 2017).”

In addition, to further test the outlier filtering side of your methodology, you could generate false erroneous measurements and further quantify how well your method performs at filtering simulated outliers.

We thank the reviewer for their suggestion. It would indeed be interesting, however we do not wish to implement this at this stage, as this may require several days of work to get it done right, and would not change the main results of our study. The way false erroneous measurements is critical to obtain meaningful results, and it would be a whole study in itself.

- L281: “keep true elevations”
  Be careful with the use of "true". All measurements are imperfect representations of the "true" elevation, which is by definition, unattainable.
  This now reads as “to keep accurate elevations observations”, and similar replacements have been done in two other sentences.

- L309, Add a full stop before "To test"
  We divided this part in two paragraph, the second one starting at this sentence.

- L376: “one of the shortest surge cycles in HMA.”
  Is this from Bhambri et al. (2017) ? Make sure to add proper reference
  It is a generic information which can be inferred from various inventories, which find shortest cycle durations between 5 to 8 years.The sentence is now:“ among the shortest surge cycles in HMA (Bhambri et al., 2017; Sun et al., 2022; Yao et al., 2023; Vale et al., 2021).”.

- L378: “Our data suggest it started 1-2 years later, implying a longer quiescence phase of 11-13 years”
  Do not make this a general statement on the dynamics of Yazghil glacier - ≈ 8 years of quiescence is not different from ≈ 11 when the number of considered events is 2.
  It was not intended as a general statement. We precised it to avoid further confusion, “implying a quiescence phase of 11-13 years for this cycle.”. Furthermore, Bhambri et al. (2017) are indeed speaking of a cycle duration of 8 years, while we do speak of quiescence duration.

- Figure 1: Readers familiar with surges and HMA will know where the glaciers you mention are, I am not sure this is the case for the broader audience - maybe you could zoom in on a bounding box around the selected glaciers.
  I am not sure the whole Kararokam region needs to be displayed since you focus on specific glaciers.
  The interest of plotting the map of the whole Karakoram region is to show the disparity of DEM observations along stripes, although indeed the area is too large. We will zoom in for the revised version.

- Figure 3
  This is a bit of a nitpick here but I would refrain from using two similar colors for the lines in "Interpolation it. 1" and "Interpolation it. 2" - being colorblind, I can't see the difference between them.
  In addition, it would be beneficial if you showed the uncertainty associated to each measurement on the plots to the right.

Finally, I see no mention to any Student-T distribution in the paper (because the methods you rely on make no explicit assumption on the distribution of the data). Rename the "t-interval" into "Confidence Interval".

Figure 6
It's really hard for me to see the individual points between raw elevation and filtered measurements.
I think the symbols and the figure in general are just too small.
Again, just remove t-confidence interval and use confidence interval. i think it's too specific for most readers - if they want to know more, they will read Shekhar et al. (2021).

Figure 7
Increase the size of the figure and individual panels.

We will go though a second round of colour blindness test. We will improve these three figures accordingly in a revised version of the manuscript.

- Figure 12
  Please add the red A, B and C regions in the captions. It's a shame to have to go into the text to grasp what the figure shows.
  We completed the caption with "The areas and trends designated in red are discussed in subsection 5.3. They highlight areas of large surge smoothing or removal (zone A) or overall smoothing of elevation changes (trend B) by the original method (Hugonnet et al., 2021), and artefacts created by the presented workflow (zone C)."

- Table 2
  This table is pretty confusing.
  I would suggest replacing Table 2 with a figure showing the distributions for each glacier.
  This would avoid having 2 columns as the 90th percentile and show the full distribution.
  We replaced the table with the following figure :

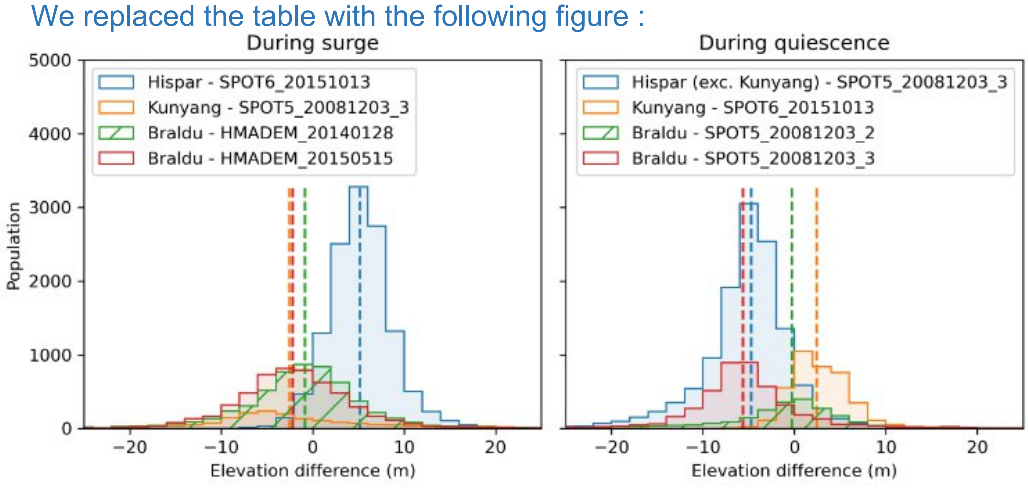

Figure 11. Histograms of the elevation difference between the references DEMs and the DEMs of the workflow interpolated at the same dates. We consider only surge-affected areas. Vertical dotted lines are the median of each histogram. The largest median is 5.18 m (resp. -5.63 m) during surge (resp. during quiescence).