Review of Zhang et al.

This work by Zhang et al. presents a “validation for the ITS_LIVE v2.0 dataset” (see my major comment for the reasoning of the quote) using in-situ GNSS data from three glacier regions as a validation set. The authors find good agreement between GNSS data and the filtered ITS_LIVE data, but for small or narrow mountain glaciers, there may be additional bias due to proximity to rock. The authors also show that optical image-derived velocities and SAR-derived velocities have strengths and weaknesses depending on acquisition time and surface conditions, which should be considered in any practical case. Finally, the authors show that (for an optimal subset of the ITS_LIVE dataset) the ITS_LIVE internal error estimates can resemble true errors.

Despite this article containing helpful and valuable information for users to assess the largest open glacier velocity dataset of the world (ITS_LIVE), I find this article hard to read, partly because the article is heavily focused on the comparison of RMSE/NMADs, and partly because some statements and logical links are confusing to me. There were several times when I got lost in the article while trying to find supporting evidence for a certain argument. I have compiled my comments below to explain this in more detail, and I hope they are helpful.

Major comments

- **Why validate a subset?** The title of the study says it is about a “validation of ITS_LIVE v2.0,” but the authors actually tuned a workflow to create a subset of ITS_LIVE v2.0 data using error thresholding, spatial filtering, and a sensor mask (i.e., skipping pre-LS8 data, and perhaps LS9 as well). In this study, the authors validate this subset instead of the full ITS_LIVE v2.0 dataset. I am concerned about whether this is still appropriate to call a validation of ITS_LIVE, since the subsetting workflow significantly alters the original characteristics of the ITS_LIVE data.

  For me, it is more about validating a specific subsetting strategy for ITS_LIVE to achieve good performance in estimating true glacier velocities. This is also important since the full ITS_LIVE data can be noisy, so I would suggest that the authors revise the narrative of this study to better reflect it.

- **Swiss Alps versus French Alps.** The authors state that the validation is conducted over glaciers “*spanning three regions with contrasting dynamic regimes.*” However, I do not see much difference between glaciers in the Swiss Alps and those in the French Alps in terms of their physical characteristics. They are all small, narrow, and slow, and their validation results are similar. In fact, the authors present the results from these two regions together several times in the paper (e.g., Figure 8). Are there contrasting characteristics between the two regions that I missed?

- **Validation should not be limited to comparing RMSE and bias; the presence of observation gaps is also important.** Take the results for the Yukon region, for example. Many high-speed measurements are included in the Sentinel-2 subset, whereas few are included in the Landsat 8 subset. No high-speed measurements are kept in the Sentinel-1 subset. When we compare RMSE/NMAD between these resulting datasets, the Sentinel-2 dataset contains more high-speed measurements and is likely to have higher RMSE/NMAD values, as seen from Figure 5b (the absolute distance to the 1:1 line tends to be longer for high-speed measurements than for low-speed measurements). However, in certain cases we may still prefer the Sentinel-2 dataset since the surge velocities are kept.

  Since this study is about a validation of a subsetting technique (see my first major comment), and since the authors say the thresholding is designed to “*optimally balance data quality and retention of sufficient samples for small-scale mountain glaciers,*” I think at least a discussion about the observation

gaps cannot be skipped. A formal analysis is even preferred, but I will let the authors decide.

- **Positive bias of the Sentinel-1 results.** The Sentinel-1 results in France show a distinct positive bias. In Section 5.2, the authors state, "*The bias may be due to the coarse resolution of Sentinel-1 leading to a larger pixel-level displacements (see section 6.1),*" but I do not see anything in Section 6.1 that further elaborates on it. Also, why do we not see a similar effect in the Swiss and Yukon region?

- **Visualizing the rock proximity effect.** The "rock proximity effect" is highlighted several times in the discussion to explain why the ITS_LIVE results in the Swiss/France regions are biased low. Since each GNSS station is associated with a distance to the closest rock, perhaps the authors could perform a test, plot that distance versus the average bias for each GNSS station, and see whether this effect is evident. My hypothesis is that as the distance increases (i.e., less stable rock area included in the matching chip), the bias decreases. That said, I think this is just my extra thought, and I would respect the authors' final decision on whether they would perform this experiment.

Other specific comments

- **L40:** The ITS_LIVE reported errors do not always agree well with the GNSS-ITS_LIVE deviations, do they? (cf. Fig. 11)

- **L59**: "*when combined with velocity data*": ITS_LIVE is already a velocity dataset, so I do not understand what this means.

- **L64-69**: The sentence "For instance…" has a strange connection to its previous and preceding sentences. It is expected that the unique niche is elaborated after "for instance," rather than the examples of velocity datasets. In addition, GoLIVE jumps out of nowhere at L68. More information is needed for it?

- **L75**: Doesn't ITS_LIVE use Landsat 9 as well?

- **L97-108**: I feel the text here (From "*The GNSS data…*") could be reorganized to form a single paragraph where the introduction about each glacier is better woven with their surge history. Is Kaskawulsh a surge-type glacier?

- **Figure 1**: Consider improving the visualization by (1) zooming in for panel (b) so GNSS points are not clustered together; (2) making several sub-panels (or merging Figure S6) and zooming in for panel (d) so the velocity field and the GNSS points of each glacier are visible; and (3) using a perceptually uniform colormap as the color representing ~300 m/yr and >625 m/yr could be very similar with the current colormap.

- **L128-129**: The citation GLAMOS is missing the year.

- **Table 1**: (1) Should there be points from Silvrettaglescher? (2) For reproducibility purposes, I'd recommend adding coordinates of each GNSS site in "Point location" or as a column.

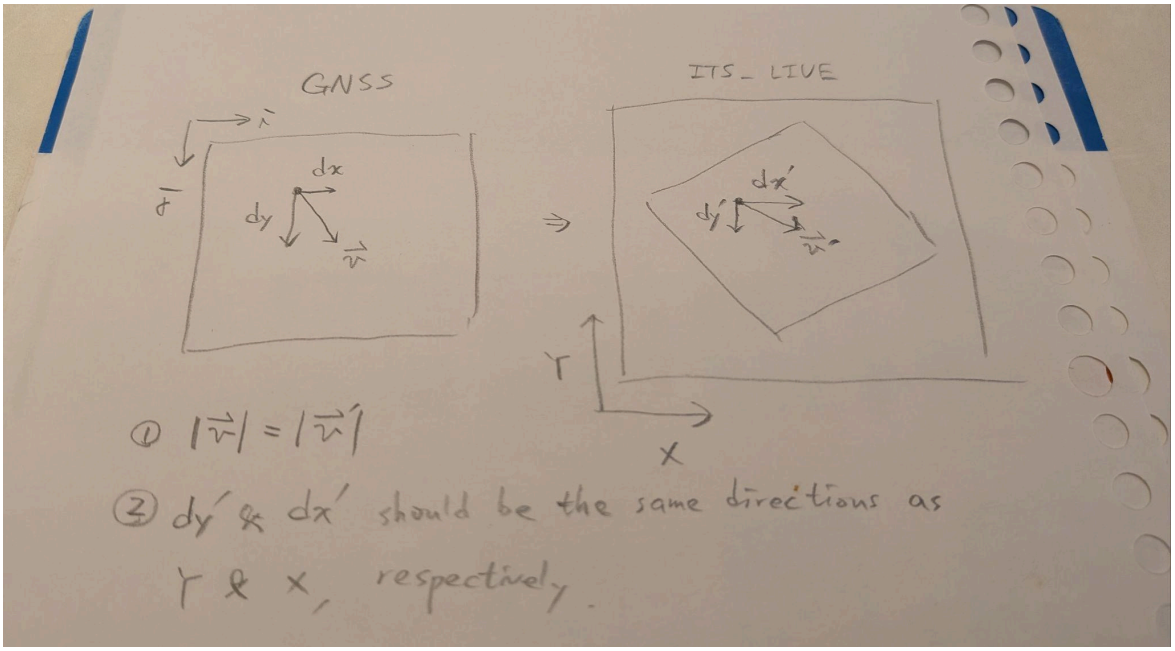
- **L183**: What are "position gradients?"

- **L186**: I am not sure I understand the meaning of "a data validity threshold of 80%."

- **L227-230**: (1) "*Y-axis is oriented toward the central meridian (45°W)*" is a bit confusing to me. The central meridian does not always represent 45°W, and we need the content inside the parentheses to form a meaningful description. Maybe "Y-axis is oriented along the southward direction of the 45°W

meridian”? (2) Why is this conversion to EPSG:3413 needed?

- **Figure 2**: Please see my sketch for two suggestions. I think these should improve accuracy.



- **L257-258**: 24 surrounding values should be U_1 to U_{24}. U0 is the 25th.

- **L263**: The unit of ε is pixels; does this mean that every other quantity (offset values?) is converted to pixels during the spatial filtering? It should be explicit since the authors seem to use displacement in m or velocity in m/yr for the rest of the paper.

- **L266**: Based on Figure 3, I suppose the “*filters described above*” refers to the LOF filter instead of the ITS_LIVE filtering workflow? Since Section 4.2 talks about the ITS_LIVE filtering workflow, I would suggest being more specific here.

- **L284**: Are there any references to suggest that KGE becomes more popular in glaciological research?

- **L318-320**: Why is a bias of 3.7 m/yr described as “substantial?” Why is a bias of -0.1 m/yr described as “negligible?” Both values are smaller than the corresponding RMSE and NMAD of each data set. For both LS8 and Sen-2 results, A simple statistical test cannot reject the case that the true bias equals 0 m/yr.

- **L338-341**: I think a more important reason why RMSE is small is that the velocity measurements are small.

- **L357**: Based on Figure 8, I am not convinced that “*Swiss data points are well aligned with the 1:1 line.*” What is the R^2 value?

- **L371-374**: I think it is very important to point out again here that the Yukon region is an exception: vy goes east-west, and vx goes north-south.

- **L418-425**: The argument about the proximity of bedrock terrain is already explained in Section 6.1. Also see my major comment about this argument.

- **L455-458**: I cannot find any evidence from Figs. 6 and 7 to support the argument regarding the spatial correspondence of the ITS_LIVE-GNSS discrepancies.

- **L463-465**: To support this argument, an RMSE derived from a subset composed of only optical tracking results is necessary. However, Table 3 does not indicate whether this is true.

- **L479-480**: What does "*limited stable terrain*" mean?

- **L491-492**: From where can we see that Donjek Glacier has a relatively stable flow?

- **L499-500**: "*Fig. 9 presents the annual 500 time series of ITS_LIVE v_error compared with the GNSS–ITS_LIVE deviation, with no v_error thresholding applied.*" Based on Section 4.2, I suppose the data presented in the results section have been applied with a v_error thresholding mask. So, are the data used in Fig. 9 different from what is discussed previously?

- **L536-537, 568-570**: I do not fully understand this sentence. It'd be appreciated if the authors could further explain (1) "*seasonal variations of uncertainty*" versus "*seasonal velocity variability*", and (2) the potential reason that GNSS-ITS_LIVE deviation is larger than v_error.

- What is the availability of the data generated from this study (e.g., the subset and point time series of ITS_LIVE at GNSS locations)?

- **Figure S5**: should be "red dots" instead of red lines.