Response to the comments made by Referee #2

Dear Dan Cadol,

We appreciate your valuable comments and interesting ideas regarding some key elements of the manuscript, which have significantly helped to improve it in our opinion. We agreed with most of your suggestions, and have made the modifications accordingly. Below, the comments are reported in italics, and our responses in normal font (blue color). The indicated line numbers refer to the tracked-changes version of the revised manuscript.

Comment 1: *This paper presents a valuable synthesis of efforts to clean SPG signals and make comparisons across field sites. While complex, I found the 'apparent' impact cleaning method to be clear in the end. But key to my understanding was Figure 7. I was a bit lost regarding the thresholds until I came to that valuable figure. Even after cleaning, the bedload flux prediction errors up to 5-fold suggest that there is still work to be done. But this is a major step forward. And it's difficult to know how much of the discrepancy is due to direct bedload measurment uncertainty or the SPG signal features and processing.*

My main question/suggestion is also related to Figure 7. Using fixed thresholds is reasonable due to its simplicity. But Fig 7b suggests that thresholds that are functions of both MaxAmp and MaxAmp/f would exclude apparent pulses and retain real pulses more reliably, without the step-like effect of the fixed value thresholds mentioned by the authors. Given the computational and field effort required to obtain the MaxAmp and f_centroid data, this extra bit of analysis seems trivial by comparison. But perhaps the need to automate the signal reduction makes such an approach untenable? Or the exact slope and intercept of the dividing lines in Fig 7b that I am suggesting varies from site to site? A brief exploration of the differences in the MaxAmp vs. MaxAmp/f plots for different sites and flume setups, or just a comment on the differences, might help.

Another possible advantage of sloped thresholds is that you could make them nonoverlapping, eliminating the double counting of packets. I wasn't entirely convinced by the statement that double counting impacts is a non-issue.

Response: We kindly thank you for your positive comments.

We have considerably reworked Section 2.5.2 (Lines 252-303) to make the threshold definition clearer.

Your idea of thresholds defined by a smooth function rather than a step function is interesting. In a recent study, we showed that the signal response of the SPG system was similar at all four sites (Figure 7, Nicollier et al. 2022). Using a unique function to split real from apparent packets would thus be possible, even though we expect some small variations between sites. However, what is certain is that defining the exact slope and intercept of the site-specific dividing lines between real and apparent packets is impossible because the true origin of a packet in field data cannot be determined. The only dataset that enables us to do so is the one resulting from the flume experiments at Obernach conducted with the partition wall. In the study mentioned earlier (Nicollier et al. 2022), we did not select real packets using class thresholds but simply on the basis of two criteria, one of them similar to Eq. 3. This procedure is therefore comparable to your suggestion of a continuous function. Before developing the concept of lower and upper thresholds, we tested whether combining such a simple criterion as the amplitude threshold defined by Wyss et al. (2016a) would result in satisfying transport and grain-size estimates over all four stations (Table 3). Even though the results were close to the estimates obtained with the AF method, they were never as good.

In our opinion the overlapping class boundaries are convincing for the following reasons:

(i) We used the 75th percentile values of the ranges visible in Fig 5 to define the thresholds. Most of the impacts will therefore generate packets that are being classified in classes that are lower than the class *j* corresponding to the true grain size. Overlapping class boundaries therefore enables a less strict classification of the few packets that are on the edges of the classes. In Figure 7b out of 2256 packets recorded by G2 (blue), 144 packets have been counted twice. But interestingly, not a single of the 153 packets recorded by G1 (red) encompassed by the class boundaries has been counted twice. We could therefore state, even if this might be somehow exaggerated, that the overlapping class boundaries also contribute to an increase of the signal to noise ratio.

(ii) Counting packets twice should not introduce any type of bias because in the second step of the presented signal conversion procedure, we have calibrated the SPG system by dividing the number of packets recorded in a given class by the bedload mass of the corresponding grain-size class. The packets counted twice are therefore included in the calibration coefficients $k_{b,j}$, resulting in will slightly larger values.

We have added these considerations on Lines 519-522.

- Wyss, C. R., Rickenmann, D., Fritschi, B., Turowski, J., Weitbrecht, V., and Boes, R.: Measuring bed load transport rates by grain-size fraction using the Swiss plate geophone signal at the Erlenbach, *J. Hydraul. Eng.*, *142*(5), https://doi.org/10.1061/(ASCE)HY.1943-7900.0001090,04016003, 2016a.

- Nicollier, T., Antoniazza, G., Rickenmann, D., Hartlieb, A., and Kirchner, J.W.: Improving the calibration of impact plate bedload monitoring systems by filtering out acoustic signals from extraneous particle impacts. *Earth Space Sci.*, *9*, e2021EA001962, https://doi.org/10.1029/2021EA001962, 2022.

Comment 2: *Fig 8b. There are fewer pulses/kg of the two smallest particle classes relative to the third smallest. Is this decline due to saltation? It's an interesting result, which was masked by the counting of apparent packets in the amplitude-only thresholding method. I would appreciate some thoughts about it in the discussion.*

Response**:** Thank you for this good observation. A similar decrease towards the two smallest classes has already been described by Wyss et al. (2016b). On the basis of flume experiments they could show that the number of impulses detected per unit weight strongly decreases towards zero after reaching a peak value at around 25-37.5 mm. Because of the steel plate's mass, the low energy released by impacts of such small particles but certainly also because of the longer saltation lengths, the detectability of particles decreases with decreasing particle size for the two lowest size classes. We have followed your suggestion and have added some comments on this interesting characteristic of the calibration coefficients on Lines 522-525.

- Wyss, C. R., Rickenmann, D., Fritschi, B., Turowski, J., Weitbrecht, V., and Boes, R.: Laboratory flume experiments with the Swiss plate geophone bed load monitoring system: 1. Impulse counts and particle size identification, *Water Resour. Res.*, *52*, 7744–7759, https://doi.org/10.1002/2015WR018555, 2016b.

Comment 3: *Line 545: Why are there fewer impacts/kg when the particle (or flow) velocity is higher? I would think greater particle velocity would produce more readable impacts, and thus more impacts/kg? I think the text is suggesting that it's because of more saltation, and thus skipping over the plate. Is this correct? Just a little more clarification of your hypotheses for this feature in the data would be appreciated.*

Response: Yes, this is correct. Wyss et al. (2016b) had already proposed that this is due to fast moving particles being less likely to collide against the Swiss plate geophone than slower moving ones, which are more frequently in contact with the bed. A secondary effect concerns is the increased energy released by an impact with higher flow velocities. Wyss et al. (2016b) hypothesizes that smaller particles that are transported faster collide with more energy against the Swiss plate geophone than slower particles, probably due to increasing saltation height or increasing turbulence with increasing V_f. Still, the better detectability of particles that arises from the higher impact energy seems to be insufficient to compensate the strong reduction of the number of impacts on a plate with increasing flow velocities. This possibly arises from the fact that larger flow velocities (without increased turbulence) may also lead to flatter saltation trajectories, thus decreasing the vertical component of the impact force. We have tried to clarify these effects of the flow velocity on Lines 593-594 and Lines 598-601.

Comment 4: *Line 590: It's good to make clear that uncertainty in the direct measurements used for calibration is very real, and that this may contribute to weaknesses or biases in the predictive power of the modeled estimate. You do mention this, I just think it can be easily forgotten in general, and perhaps merits emphasis.*

Response: We agree that this point is essential and can only be stressed more. We have added a sentence that underlines this problematic on lines 555-556.

Further minor changes

We have also made some further minor changes to the original manuscript. These mainly concern typos, update of recently published references, and general reformulations of terms or sentences. All changes can be found in the "tracked-changes" version of the manuscript.