Response to the comments made by Referee #3

Dear Roger Kuhnle,

We appreciate your valuable comments regarding some key elements of the manuscript and thank you for having raised the interesting issue of real packets being potentially eliminated. We believe that your comments, questions and suggestions significantly helped to improve the manuscript. 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: *The manuscript reports on a new method to calibrate the Swiss plate geophone (SPG) which uses a combination of data collected in a laboratory flume and at four different field sites. The SPG has been shown in previous studies to be an excellent indirect method to measure the rate and size of bed load transport in gravel bedded streams and rivers. This study further develops the science of turning impact data into quantitative data of the transport of bedload. The study of bed load transport using impact plates at several field sites has indicated that a general calibration of the SPG has been difficult to develop for a variety of reasons. This study uses a combination of amplitude and frequency to calibrate the SPG in the quest for a general calibration relation to turn impact data into quantitative values of mass and grain size for gravel bed load transport. This manuscript contains much valuable information and should be published, however, suggestions for improvement of the presentation are given below.*

Response: Thank you for these positive comments on our work.

Comment 2: *Lines 246-251: In these lines it is related how all packets were filtered using equation (3) and packets which do not meet this criterion are ignored in further analysis. It is clear to me how and why this was done, however, in Figure 7 there appears to be substantial overlap between real (blue) and apparent (red) peaks measured in the flume experiments. How many real peaks were rejected using this criterion in the flume experiment data? Also, could the authors estimate how many real peaks were rejected from the 4 field data sets considered in the study? I believe text should be added to the manuscript discussing this issue.*

Response**:** Thanks for this interesting question that is not trivial to answer. We have performed single-grain size experiments in two different ways. The first type, which is also the most relevant one for this study, was conducted without the partition wall. Eq. 3 has been applied to the data collected during these experiments only. A comparison of the signal responses before and after the application of the criterion described in Eq. 3 is shown in Figure 5. After the application of the criterion about 39% of all packets are remaining. This

information has been added at the end of section 2.5.2, on Line 290. These are being used to derive the lower and upper threshold values. In absence of the partition wall, it is difficult to give any indication about the number of real packets that were filtered out by Eq. 3, because the true location of an impact at the origin of a packet cannot be verified. This is also true for the field measurements.

The following short quantitative analysis of Figure 7 could help the reader to better understand the effect of the application of the two-dimensional thresholds on the number of real and apparent packets. As mentioned on Lines 167-169, the single grain-size experiments performed in presence of the partition wall are used to illustrate the performance of the two methods (see Figure 7). Eq. 3 has therefore not been directly applied to the dataset resulting from these experiments, only the class threshold values mentioned earlier have been applied. Through the application of the AF thresholds (Fig. 7b and d), an important part of the packets detected by G1 and G2 is being eliminated. "The AH thresholds encompass in total 1945 packets for the shielded geophone G1, and 4823 packets for the unshielded geophone plate G2. In comparison, the AF thresholds encompass in total 159 packets for the shielded geophone G1, and 2202 packets for the unshielded geophone plate G2 (counting the packets in the overlapping class boundaries only once)." (inserted in Lines 356-359). An important point is that the blue dots (packets detected by the unshielded sensor G2) that are in the overlapping area with red dots (packets detected by the unshielded sensor G1), are most probably not real packets. We have added a short explanation to clarify the reason for this certainly confusing but very important point on Lines 345-347. The overlapping area arises from the fact that a seismic wave generated by an impact on the concrete bed follows a similar path towards both sensors, resulting in the recording of two apparent packets with comparable characteristics. We expect that also packets originating from "unclean" impacts, for example close to the edge of a steel plate, or with a large horizontal component (low amplitude), could be found in this overlapping area, but unfortunately we cannot indicate how many. The fact that there is, most probably, not a sharp line splitting real from apparent packets, might be a good explanation for why the most optimal class thresholds include a part of the packets located in this overlapping area.

Comment 3: *Lines 239-240, and 257-260: In these sentences the lower and upper thresholds for the amplitude-frequency method are described. Is it correct that the lower threshold (V) was based on the minimum grain size of the size fraction and the upper threshold (V Hz) was based on the maximum grain size of the size fraction being considered? The clarity of the text should be improved to make it easier for the reader to interpret the details of how this technique was implemented.*

Response: Yes, this is correct. We have reformulated several sentences in Section 2.5.2 (Lines 252-303) and hope that it clarifies the entire process leading to the lower and upper thresholds.

Comment 4: *Figure 9: This Figure is too small and has too much information contained in it. This renders this Figure very difficult to interpret other than for a general impression of the data trends. Consider simplifying this Figure or possibly presenting this information on two Figures.*

Response: We have modified the format of Figure 9 as well as the corresponding Figure S1 in the Supporting Information in order to improve their readability. However, the main goal of this figure is really to give a qualitative impression of the accuracy of the transport rate estimates and is not meant as standalone element. All the values required to evaluate the performance of the two methods are listed just below the Figure in Table 5.

Comment 5: *Lines 394-400, Table 5: The comparison of the two methods for arriving at quantitative rates and sizes of bed load is interesting. Was the criterion in eq (3) applied to the data before the amplitude histogram (AH) method was implemented? It is clear that the criterion it eq (3) was used as part of the technique for the amplitude-frequency (AF) method. Some text should be added to make it clear as to whether eq (3) was applied in relation to the data before applying the AH method.*

Response: Thank you for this important remark. In order to compare the new AF method with the original AH method developed by Wyss et al. (2016a), we did not apply any filtering of packets using Eq. 3 before implementing the AH method, since this was not part of the original procedure. We have added this information on Lines 273-274.

- 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.

Comment 6: *Lines 571-598: It is clear that the AF method performed better than the AH method in some cases such as the Erlenbach, however, it is also clear that for the other 3 field sites the AH method yielded results for bed load that were quite close to that obtained from the physical samples. It is also not clear whether the general calibration calculated in this study would give.*

Response: We agree that the results obtained with the AH method for the three "natural" sites were already good and of similar accuracy. Indeed, all three sites had already very similar site-specific calibration relationships before filtering out apparent packets (see Fig. 8a). The main motivation behind the AF method was not necessarily to improve the accuracy of site-specific calibration relationships but rather to better understand and possibly reduce large

differences among (all) sites. Considering Fig. 8b, one can notice that through applying the AF method, the differences between the calibration relationships of the three natural sites and the ones of the Erlenbach site have been reduced. Most results in the present study seem to suggest that the improvements arising from the use of the AF method mainly concern the comparison of the Erlenbach and the other datasets. However, as one can see in Fig. 8, also the $k_{b,j}$ coefficients of the three other sites underwent important changes (see also Table S4). In our opinion, the fact that the accuracy of the estimates at the three natural sites has not been reduced with regard to the AH method, even after the removal of an important part of the detected packets, supports a use of the AF method. This considerations has been added on Lines 493-494.

While the lack of accurate flow velocity measurements is certainly one of the critical points of the study, one could argue that another one is the low variability between the site-specific calibration relationships already before implementing the AF method. Indeed, it would have been extremely interesting to test the method on a larger number (and variety) of sites. Unfortunately, these are the only sites at which a full geophone signal has been recorded during calibration measurements. These considerations have been added on Lines 504-508.

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.