HESS GRAINet review answer

Dear Matjaz Mikos,
We thank all the reviewers for their inputs. Please see our detailed answers below.
We hope that the reviewer's concerns could be solved with our clarification.

New changes in the manuscript are highlighted in red. Changes from the first revision are still highlighted in blue.

Editor

After receiving two reviews, there is a clear need for further revisions to take into account the suggestions of the both reviewers. Comments and suggestions of Referee #3 is a bit easier to answer - but please, follow the suggestions to add (American) literature - I am also aware of the Fehr method (Anastasi), but Bunte, Abt, ... should be used and mention.
We already refer to the proposed american literature in the introduction where appropriate (Bunte and Abt (2001), Rice and Church (2010)) . Since our data annotation strategy closely follows Fehr's line-sampling approach, we cannot avoid the reference to Fehr for obvious reasons. However, we now connect this surface sampling procedure to the commonly used terminology in the literature.
Our novel methodology has been evaluated on grain size data for Switzerland. Thus, we adopted Fehr, which is still considered the gold standard in Switzerland and used almost universally in Swiss engineering practice.
Nevertheless, we added an explanation in the paper (footnote line 79) that our methodology is **not** restricted to the Fehr sampling strategy. The proposed CNN can be trained to replicate any kind of grain size data, no matter how it has been sampled. May it be along a transect, a grid, or even a volume sample. (see also our previous answer to RC3).

I would also agree with his remark on mean grain diameter dm-values and its significance for the calibration of numerical sediment transport models. I would rather think into the direction of fraction models using several sediment fractions to be able to incorporate building of an armour layer and the influence of selective transport mechanism. An advanced method for grain size determination should overcome the limitations by a mean grain size d-m models. Why to apply a sophisticated method for d-m determination rather than to get a full GSD?
As we write in line: 91, the novelty of GRAINet is indeed its ability to predict the full grain size distribution at every location on a gravel bar. We agree that this offers a great potential to calibrate numerical models. However, we also show that our data-driven CNN is generic and can also be used to predict characteristic grain sizes. Aggregate statistics such as d_m have their role, too, e.g., it is straight-forward to visualize a characteristic grain size as a spatial map, whereas visualizing the full GSD in a spatially explicit manner is not trivial.

The comments from Referee#2 are more important. Please, give readers more details about the procedure and the equipment (UAV - we have applied Phantom DJI drones for rock fall applications and surface displacements on scree slopes), so that the reader would be aware about the possibilities. This is important, if such field campaigns are performed by

non-experts (we ask geodetical engineers that have pilot certificates). A further clarification is needed in this regard.
All the information needed to reproduce the UAV surveys was already given in the paper in section 3.1.
To cover the reviewer's concerns about the flying time, we added a paragraph in the discussion section 6.4 (lines: 641-647).

Please, follow the suggestion and redo the geographical cross validation. This is a critical issue to be able to generalise the results. Are you sure that grain size distributions at a series of gravel bars of the same river are not inter-related?
The grain sizes of the investigated gravel bars do not follow Sternberg's fluvial abrasion law, but are dominated by discontinuities. The grain size characteristics vary greatly along the same river (see Table C1). As we write on page 31 lines 584-593, the grain sizes may be impacted by several influences: changing river characteristics (slope, bed width), anthropogenic barriers (i.e., dams; these are not free-flowing rivers), tributaries (increased catchment areas), and artificial gravel replenishments.
We also explicitly discuss in line 575 that this experimental setup is valid for our specific dataset and that in different scenarios with slowly varying grain size properties, a river-based cross-validation may be needed to avoid overly optimistic results.
Our available dataset with 25 bars allows us to study the generalization across gravel bars (see title of section 5.4). In line 86 we already clarify the generalization experiment. The paper never claims that our method would generalize to unseen conditions. Rather, we explicitly mention in line 601 that the data-driven approach will only work if the training data and test data follow the same distribution, which goes beyond grain size distribution and also includes appearance (weather, lighting, surface cover of gravel, ...).

You should elaborate more on this issue (using classic papers from fluvial geomorphology on sediment sources and sediment links in a fluvial system). Please, compare results of the geographical cross validation using all data (this is already done) and only data from different rivers (skipping potentially cross-related data from neighbouring gravel bars on the same river). We should clarify this issue before going on with the publication process.
On the request of the reviewer, we have identified five bars that are per definition independent, as they belong to separated river reaches: *Aare km 040.8, Kl. Emme km 030.3, Emme -, Rhone km 083.3, Rhone km 114.0*.
These bars are separated from the other bars with the same river name by new inputs of sediment. Thus, for these bars, there are no neighbouring bars in the training data.
In our per-bar cross-validation experiment, the average error (MAE) for these five bars is 0.3cm higher than for random cross-validation, whereas the average error over **all** bars (including the supposedly not geographically separated ones) is 0.4cm higher. See the attached Figure 1 in this document. We conclude that, within our dataset, seeing bars from the same river during training does not lead to over-optimistic results.
Generalization is much more affected by unique local environmental factors (e.g. wet stones, algae covering) that were not seen during training. The dominant error case in the generalization experiment was that all samples with wet grains are from the same bar (i.e., *Aare km 156.7*; see also line 528 in the paper). We clearly see that *Aare km 040.8*, although on a separated river reach, performs better in our per-bar cross-validation than the majority of gravel bars along the river Aare. In general, we observe more pronounced drops for bars

with coarse grains (Gr. Entle, KL. Emme, Rhone), which we assume is due to sampling bias, as samples with coarse grain sizes are rare (see line 598, 636).
After these additional analyses we are firmly convinced that our study is correct (for the available dataset), and consistent with our careful discussion on the generalization performance in section 6.2. We would thus prefer to keep section “5.4 Generalization across gravel bars” and not replace it with a “per river name” cross-validation, with similar quantitative outcomes, but a smaller data basis. If the editor insists that the replacement is necessary, we would of course oblige. Nevertheless, we include this analysis in the paper section 6.4 (lines 577-583) with the new Figure in the Appendix Fig. H.

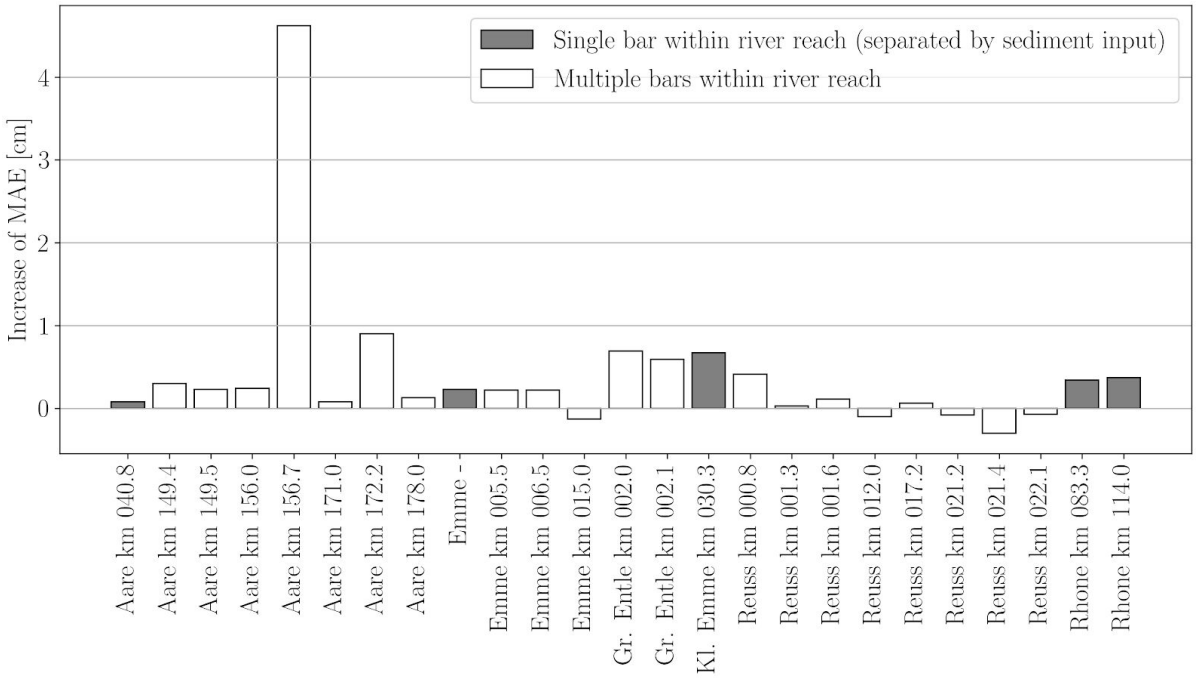

Figure 1: Increase of mean absolute error (MAE) from the random cross-validation to the cross-bar generalization experiment. A positive value means that the generalization experiment has a higher error.

Report 1

This revised manuscript has many improvements on issues like model architecture and training. Class activation maps also provide a very useful insight into how the model works. Overall, this method is very innovative and it produces results that are high quality and very difficult to achieve with other methods. However, the authors still do not clearly acknowledge the limitations of their method and this rests on 2 points: an unclear understanding of the logistical costs of acquiring sufficient data for GrainNet with a UAV and a false result for their so-called geographic cross validation.

First, the authors begin their response letter by stating that they do not see an issue with the acquisition of drone data at 0.25 cm of spatial resolution and characterise it as a ‘minor technical detail’. I will therefore clarify my comment with a worked example. Start with a 1 hectare (100x100 metre) bar a a unit sampling area. The project uses a DJI P4 pro, let’s simplify the problem by assuming a 90 degree FOV meaning that the image footprint is twice the flying altitude. The images were acquired at 16:9 aspect ratio with 5472x3648 pixels. From this we can derive that the drone was approximately 6.8 m above ground. Given that the method needs an orthomosaic, I will assume that the images are flown at 80% forward overlap with a 50% sidelap. The image height is 9.1 m. Between images the drone must move 20% of the image to get the 80% overlap. This is 1.82 meters. On the P4 pro and with the fastest SD card on the market, you need to leave abut 2s for the mage to write to disk, anything less and the drone will start missing images during the mission. So the optimal flight method is to get a slow continuous motion of the drone. 1.82 meters in 2s is 0.9 m/s. It will therefore take roughly 0.9 minutes to complete 1 line of 100m. With the images being 12m wide and a 50% sidelap, We need about 17 flight lines to cover 1 hectare. For a total flight time of about 15 minutes/hectare.

Now consider an alternative setup that is used to get grain size data for alternative texture mapping methods. In this case imagery acquired at 2-3 cm of spatial resolution is suitable. In this case, flying a P4 pro at 50 m altitude will deliver suitable imagery at about 2cm. At 50m, the footprint of 1 image is 100m x 56m. At 80 overlap and the same 2s interval between images, the drone flies at 5.6 m/s. Given the image width, we only need 2 lines to cover the hectare. Meaning that the total operation needs only 34 seconds/hectare.

Therefore, data acquisition for GrainNet requires drone operations that are 30x longer than for older methods. That is not trivial and readers deserve to know this fact.

The authors’ suggestion that magnification can solve the problem is incorrect. When you magnify you increase the focal length and thus reduce the image footprint, flight velocity for SfM acquisition remains the same. Whilst it is true that a higher resolution camera could indeed improve things, that trend is very slow. The current UAV market for science is now dominated by consumer, non-scientific, drones made by DJI. The simple reason is cost. The P4 pro resolution of 20 Mpix is already on the high side. The only way to improve the performance would be to use top of the line cameras that have high speed writing buses. For example, mounting a Canon EOS on a big drone like a Matrice 600 would indeed be much faster, but then you are talking of a 1 order magnitude increase in cost for drone equipment. Either way, the acquisition of appropriate data for GrainNet is a significant barrier to access.

To cover the reviewer's concerns about the flying time, we added a paragraph in the discussion section 6.4 to make the reader aware of this fact. We note that all drone data acquisition was performed by a hydrological consulting SME in the course of their commercial operations, which shows that the approach, even in its current, unoptimised form, is fairly practical and accessible.

The second issue is the geographic cross validation. My view is still that the authors' approach is mistaken and unjustified in geomorphology literature. The authors state on line 564 that there is no strong correlation between grain sizes on the same rivers. This statement is not evidenced and it flies in the face of decades of fluvial geomorphology. It has long been known that grain size decreases exponentially with distance downstream with periodic discontinuities (Rice, 1999; Rice and Church, 1998, 2001). This was again observed in recent remote sensing studies(Carbonneau et al., 2005). So barring the incidence of a source of coarse grained material, two successive bars on the same river can be expected to have a similar grain size composed of similar material of the same source. So unless the authors can show that between each and every one of their sampling bars there is a new input of sediment, then we must expect that the majority of neighbouring bars in the dataset are similar and LOOCV is not an appropriate method. I again make the request that the authors revise this process to hold-out entire rivers.
Please see the answer to the editor above. Our dataset includes five cases where only a single bar is located on the same river reach and thus is independent of all others bars with the same river name. We looked at the cross-validation results for these bars, which are no worse than for other bars that are not formally independent (in the sense of "separated by significant sediment inputs"). The analysis confirms our previous argument that grain sizes along the investigated Swiss rivers are dominated by anthropogenic influences (dams, channels, replenishments) as well as by frequent tributaries. Correlations due to continuous abrasion are minimal.

This is critical because as it stands, this method does provide unprecedented data over a gravel surface, but as I show above, the logistic costs of data acquisition are an order of magnitude more in time or cost when compared to older methods. If it turns out that the method does not generalise to new rivers, then local calibration will be needed at each acquisition thus increasing the total cost of the method. I do not doubt that in certain applications, such a large field effort will be justified in order to produce such high quality outputs, but the reader deserves to get a clear indication of these costs upfront.

Patrice Carbonneau
December 2020

We hope that the reviewer's concerns could be solved with our clarification.
The paper never claims that the presented method would generalize to unseen conditions. Rather, we explicitly mention in line 601 that such a data-driven approach will only work if the training data and test data follow the same data distribution, which goes beyond grain size distribution, but also includes appearance (weather, lighting, environmental factors).

Report 2

Overall, I can see several improvements in this revised version of the manuscript. That said, I still have some concerns about this manuscript:

1. Introduction. I see small improvements in this Section. I think that the main points (suggests) in my previous review were not well addressed. For simplicity such points are reported below (lines refer to the previous version of the manuscript):

L 15-28. This part is not very useful. It would be more useful to focus on why grain size data are crucial (e.g. process understanding, modelling).
Done. We now have included the motivation to collect grain size data to advance the understanding of stream processes with references to Bunte and Abt (2001).
L 38-42. Reference to traditional approaches is very poor. I would avoid the reference to Fehr (1987), maybe a good reference in the German-speaking countries but not worldwide (and in an international journal). I would suggest to look and refer to classical works by Church, Bunte, and many others. For instance, a look to Bunte and Abt (2001, USDA) would be very useful to put this work in the general context of sediment sampling in gravel-bed rivers.
Done
L 56. "….is more efficient than traditional field measurements…": I would say that automatic grain size is much less time consuming but it is also, commonly, less accurate. This should be pointed out since it is probably not obvious for readers who are not familiar with sediment sampling.
Done
2. Testing of the approach. L 606-607. "…Our CNN-based approach makes it possible to robustly estimate grain size distributions and characteristic mean diameters from raw images…". Comparison with field data (real data) is weak in this work.
We did our best with the available field data by comparing at the bar level. Precise geolocation was not available for the field data, which was assembled from multiple past projects and campaigns (in operational practice). We agree that with today's technology (e.g., real-time GNSS) it may be useful to routinely record precise geolocation, and have mentioned this possibility in the paper (line 374). But the data that was available to us, collected with today's field sampling practices, did not include precise georeference.

Overall, the authors do not fully recognize that field sampling is more accurate that sampling from images. It seems to me that the final message of this work is as follows: sediment characterization from images is more reliable and accurate than characterization by field measurements. I do not think this is the case: it would be useful to clarify better advantages and limitations of both approaches (this would be very useful in the "Introduction").
We clarified advantages and limitations in line 50. However, our message is certainly not that sampling from images is more accurate and reliable, and we never say this in the paper. If we claim any advantage over the gold standard of field measurements, then it is the ability to scale up the data acquisition to advance the calibration of automatic data-driven methods that ultimately lead to holistic analyses of gravel bars (e.g. dense maps Sec. 5.3.5, Fig. 17).

Finally, I think that a sound test to assess the performance of the model was not carried out in this work: it should be relevant to point out this for future research development.
We agree that a more extended comparison of the manual digital line sampling and field samples should be done in the future. (see line 374). This would require precisely geolocated line samples in the field that could then be compared against manual digital line samples. Nevertheless, in our view the extensive evaluation of the model against the digital line samples clearly demonstrates the advantages and limitations of this approach.

SPECIFIC COMMENTS

L 480-481. "…In order to calibrate numerical bedload transport models, a single representative dm-value of a gravel bar or a cross-section is essential…": I am not so sure about this, could you better support and justify this statement?
We removed this sentence as it was misleading. We agree with the opinion that the potential for numerical model calibration lies in the ability of GRAINet to estimate the full grain size distributions at a high spatial resolution. Nevertheless, robust estimates of characteristic grain sizes at the bar level are important to support large scale analysis along gravel-bed rivers.