Reply to reviewers for manuscript (esurf-2019-20) submission to Earth Surface Dynamics:

Introducing PebbleCounts: A grain-sizing tool for photo surveys of dynamic gravel-bed rivers – Purinton and Bookhagen

Highlighted in **bold** are the reviewer comments followed by our point-by-point replies in regular text. Sentences that will be added or changed from the original manuscript are in *italics*. All changes will be made to the final manuscript submission following completion of the interactive review period.

Reviewer Patrice Carbonneau

This is an excellent contribution which comes at timely point when many technologies are coming together. The new method rests on a genuine innovation in grain size mapping, a clever use of a k-means cluster. The paper is mostly well written and has an excellent level of technical detail. This might be demanding for readers, but a detailed reading of this work is effective in lifting the black-box effect that can arise from advanced image processing workflows with a large number of steps and parameters. One of the major benefits of this work is the fact that it is open-source and written in the popular Python language. This is very timely because Basegrain, the current best option for grain size mapping, is written for windows-7 and no longer in active development. Despite the fact that the creator of Basegrain is very collaborative and willingly shares the Matlab source code, the river sciences community will soon need an updated option with a major preference towards an open-source solution. PebbelCounts seems poised to fill this gap.

We thank the reviewer for positive comments regarding the quality, thoroughness, timeliness, and innovativeness of the submission. The level of detail in the paper is very high, which this reviewer found good, but the second reviewer found obfuscating to our main points. Our goal with the manuscript at this level of detail was to provide sufficient information for a reproducible algorithm and processing chain that can be precisely followed. Previous publications of grain-size estimation algorithms often had a shorter method section that did not allow to reproduce the algorithm. In the spirit of open source and traceable and reproducible algorithms and software, we strongly think that a well-documented algorithm is beneficial to the community for the years to come. We cover points regarding this more in the reply to the second reviewer, but are happy to see that the first reviewer found the highly detailed analysis and writing useful.

I have a few suggestions about corrections but these are not major:

1- I found equations 1,2 and 3 to be laborious and not really necessary. Non-metric RGB cameras destined for the consumer or ‘prosumer’ market have square pixels. As observed by the authors, differences in X and Y resolution are negligible. This whole section could be cut short to a single equation.

We agree that this section can be cut-down without losing too much information, but still feel a description for the interested user warrants a few sentences. Section 5.3.1 will now read (including the suggestion in point 3 below):

We refer to all imagery used as top-down as opposed to the commonly used nadir term, which refers to images taken consistently from a directly downward-pointing vantage, since our images are taken from a variety of near-downward angles. As consumer-grade cameras have square pixels with negligible difference in horizontal and vertical resolution, the image scale can be calculated directly from the camera parameters and camera height with the resolution (R) in mm/pixel given by:

$$R = (S * h) / (f * I) \qquad (1)$$

where S is the sensor height or width in mm, f is the lens focal length in mm, h is the camera height in mm, and I is the image height or width in pixels. S and I should either both be the width, or both be the height of the sensor and image, respectively. This assumes no major distortions within the field of view, which is not valid for oblique imagery, but is negligible for top-down photography at close range using non-fisheye lenses. With h=1.55 m, the resulting image resolutions tested from the Fujifilm were 0.26, 0.35, 0.53, and 1.05 mm/pixel by eq. (1).

2- A better reading of your bibliography. This bibliography is in fact quite complete. But I get a distinct impression that many papers were skimmed, deemed relevant, and cited. I am often struck by points of discussion or relevant findings of other authors which are missing in the text despite the fact that these authors are cited. Another explanation may be that the discussion lacks many key points of a good discussion where elements of other authors will need more consideration. Specifics of this will be seen below.

We have tried to fix and improve these points through some re-reading of the relevant bibliography, and re-writing of discussion points. We have made an effort to avoid spending much time discussing grain-sizing efforts that are based on texture (roughness, variance, entropy) techniques, besides to mention the papers that cover them, which are many in number and highly variable in exact methodology. We have focused our closer readings and the discussion on those studies that also employ image-segmentation techniques to the problem of grain sizing, as these are more relevant to our study and allow to decipher the full grain-size spectrum.

3- The accepted term for ‘top-down’ imagery in the remote sensing community is ‘nadir’. Please use that term.

This point is well taken and we have discussed this among the two authors. Nadir imagery refers specifically to downward-pointing images. The images used in this study are taken from a variety of angles near downward, but hardly ever exactly, since we are using an imperfect camera-on-mast setup. Top-down imagery is a more general term and we feel more appropriate for the type of imagery used in this study, especially because we do not measure the angle of the images taken. We have, however, added a sentence at the beginning of Section 5.3.1 on P10, L5:

We refer to all imagery used as top-down as opposed to the commonly used nadir term, which refers to images taken consistently from a directly downward-pointing vantage, since our images are taken from a variety of near-downward angles.

4- The drone/SfM elements of the paper are not well organised. The UAV/SfM paragraph in the introduction could be moved. SfM is now so ubiquitous that it should not shock the reader if you say in the methods that SfM was used for data acquisition. The overall reflection of how this could apply to drones in an SfM workflow needs to be moved to a section in the discussion. This is an area where many elements of the cited literature are not mentioned. Important points are:

The paragraph in the introduction concerning UAV and SfM orthomosaic generation can be shortened, but it is mainly there to provide a segue from previous manual counting methods in the previous paragraph, to photo-sieving methods, and onward to the last introductory paragraph where we introduce our photo-sieving method. We will shorten the paragraph as follows, and will create a new section in the discussion to cover the other concerns of the reviewer (see the next reviewer points below). The new introduction paragraph from P2, L10:

In light of this, measurement from photographs is an attractive option for increasing sample size and decreasing fieldwork, while covering larger areas. Increasingly affordable high-resolution --- 12--24 megapixel (MP) --- cameras, allows the collection of high-quality photo surveys at scales of entire river cross sections or reaches via Structure-from-motion with Multi-View Stereo (Smith et al., 2015; Eltner et al., 2016) at resolutions at or exceeding 1 cm/pixel (e.g., Woodget and Austrums, 2017). Even higher resolution (1 mm/pixel) river surveys can be accomplished with low-flying unmanned aerial vehicles (UAVs) (e.g., Carbonneau et al., 2018), pole-mounted cameras, or using handheld imagery.

i- Acquisition geometry. There is now a large volume of literature on the image geometry that produces the best 3D models from SfM. This remains important here since DEM-distortions will propagate to the orthoimage. So in parallel with the robotic photosieving work, there should be a recommendation of a mixed acquisition with nadir imagery for the actual grain delineation but with oblique views for maximum SfM quality.

ii- Be honest and realistic about scale coverage. The paper states an ambition of covering areas up to 10 000 m2. At the same time the method rests on SfM with surveyed ground control to generate an orthomosaic with a constant resolution. This is in fact an ambitious goal. The acquisition of 80% overlapping imagery with surveyed GCPs and at sub-mm

resolutions over a hectare is a multi-day (or multi-camera) job. This is why the robotic photosieving approach of Carbonneau et al (2018) does not advocate a orthomosaics but uses scaled individual images.

iii- Use of a Mavic as the only reference is perhaps overly pessimistic. Carbonneau et al (2018) mention that a Phantom 4 Pro (20 Mpix imagery) could acquire 0.7mm/pix imagery at 2m altitude. With active collision avoidance, that becomes a workable, if very low, flying altitude. This problem should resolve itself as sensors with more than 20MPix become more available.

Regarding the points i-iii listed here, we have re-written Section 7.3 in the discussion. This section deals with practical considerations for image collection and processing using the proposed algorithms. We now include sub-sections dealing with image resolution and geometric considerations for image collection, using a UAV-SfM robotic photosieving workflow (as covered by Carbonneau at al. (2018)), and addressing the ambitious up-scaling we propose. This new section hopefully addresses the reviewers concerns and adds clarity to the potentials and caveats of PebbleCounts as applied to photogrammetric river surveys, particularly regarding the use of drones. The new Section 7.3 in the discussion will read as follows (beginning from the current location on P25, L29):

7.3 Practical Considerations for Image Collection and Processing

To conclude the discussion, we focus on the collection of imagery by camera-on-mast or handheld setups. This includes geometric acquisition and resolution considerations. We further address the potentials for UAV surveying. Finally, we address the up-scaling potential of the proposed method.

7.3.1 Acquisition Geometry and Resolution of Mast or Handheld Images

Ideally, collecting 9+ top-down images/m^2 (as in our field surveys) or collecting an approximately 1:2 (or greater) ratio of top-down to oblique imagery (as in our experiments with point cloud data dimensions; see supplement Section S1), leads to the highest quality point cloud results in Agisoft. Higher quality point clouds, in turn, lead to less distortion errors during orthorectification and higher quality orthomosaics. Due to the textured nature of gravel images, we were able to get comparable results in reduced time using only 4 top-down images/m^2 in the lab setting. In any case, high overlap of ~80% between images is recommended to ensure the best results. Where a user desires accurate and dense point cloud data in addition to the 2D orthomosaics, it is recommended that (many) more images closer to the surface be collected and from oblique viewing angles (e.g., Verma et al., 2019).

As we find the difference in calculated resolution and subsequent grain-size measurement to be negligible between orthorectified and raw top-down imagery at these scales, the use of orthomosaic imagery is not strictly necessary when using image-segmentation software like

PebbleCounts (e.g., Carbonneau et al., 2018). However, on very rough surfaces with cast-shadows from large grains, generating orthoimagery will overcome distortions present in the raw photos. Furthermore, georeferenced orthomosaics may be preferable for capturing large sites at a constant resolution that can be fed into the algorithm.

In terms of camera and photographic height (and thus resolution) considerations, one first needs to assess the minimum grain size that is desired. Following this, the resolution of the image can be determined using eq. (1) with some knowledge of the camera parameters (focal length, camera height, sensor size, and image size). The smallest grain b-axis needed should be 20-times this resolution. For instance, using a similar camera to the Sony a6000 (24 MP, 15.6 x 23.5 mm CMOS sensor, 16 mm focal length), to measure all grains down to 1 cm one needs a resolution of 0.5 mm/pixel, and thus a maximum camera height of ~2 m. If finer grain sizes are desired, the user can use higher resolution imagery, but must be aware of the longer time needed for processing finer imagery.

7.3.2 On the Use of the UAVs

The > 20 m flight heights typical of UAV surveys lead to cm-scale imagery with currently available 12-24 MP cameras, which is less appropriate for PebbleCounts processing, unless large (> 0.2 m) cobbles and boulders dominate the river site. Carbonneau et al. (2018) build on the work of Carbonneau and Dietrich (2017) to present a workflow for robotic photo sieving on mm to sub-mm UAV imagery without any GCPs. The method uses a number of high and oblique overlapping flights to orthorectify a lower non-overlapping flight with mm-scale acquisition. In their study, the resulting georeferenced single orthoimages are measured using Basegrain, demonstrating the potential of this method to be applied with PebbleCounts instead.

Practical considerations for UAV image acquisition include the use of multiple flight heights for georeferencing, including one low flight to acquire mm-scale imagery, and the collection of both nadir and oblique imagery for improved SfM-MVS results (Carbonneau et al., 2018). Also, the use of a 3-axis camera gimbal is key to reduce blur in the images (Woodget et al., 2018). Imagery at sub-mm resolution is already achievable from newer drone models with high MP cameras flown at low heights. For example, 0.5 mm/pixel imagery from a DJI Mavic drone with a 12 MP camera, wide angle 4.3 mm focal length, and 4.55 x 6.17 mm sensor requires a very low flight height of ~1.4 m, giving a field of view of only ~1.5 x 2 m. This may be somewhat improved using better cameras like on the Mavic 2 Pro (20 MP camera). Regardless, acquiring such imagery with the high overlap (~80%) required for SfM-MVS processing is still difficult (particularly given current ~20-minute flight length limitations from available batteries). Improvements in technology will continue to increase survey sizes from UAVs, but, for the time-being, the single, non-overlapping orthoimage workflow proposed by Carbonneau et al. (2018) has high potential to achieve large-areal results from PebbleCounts using UAV imagery.

7.3.3 Coverage and Processing Limits Using PebbleCounts

Using handheld imagery, a survey site of 1,000 – 5,000 m^2 with ~10 GCPs measured via dGPS can be covered in 2-6 hours by one person (including GCP collection). Using a camera-on-mast setup, this time can be reduced by half, with even greater speed possible using more people and cameras (of the same focal length). The potential to cover even larger survey sites up to or exceeding 100x100 m (10,000 m^2 = 1 hectare) is feasible in a day of work by two people using the proposed method with a 16-20 mm focal length lens and a 3-5 m mast.

Current UAV technology limits mm to sub-mm orthomosaic generation via high-overlap SfM-MVS to relatively small areas, unless carefully applied to single orthoimages as in Carbonneau et al. (2018). However, technology improvements will continue. These include greater battery life, more accurate geo-tags from onboard dGPS, higher MP cameras, and sharper images while in motion. It is thus within reason to expect hectare to multi-hectare SfM-MVS UAV surveys at mm to sub-mm resolution in seamless orthomosaics along entire river reaches in the near future.

One limit of the scalability of the PebbleCounts method is processing time. The KMS PebbleCounts tool is recommended to be applied to maximum 1-2 m^2 patches, depending on the image resolution, as the manual clicking of good grains is time consuming, requiring 5-20 minutes per patch depending on patch size, image resolution, and abundance of finer grains. On the other hand, the AIF PebbleCountsAuto tool can theoretically be applied at larger scales. However, it is also advisable to tile data and feed it to the algorithm in maximum 1-2 m^2 patches for ~1 mm/pixel imagery, since the non-local means denoising can take minutes on very large images (> 2,000 x 2,000 pixels). Again, the use of systems with GPUs or large memory will shorten processing times and allow for larger images to be run.

In practical terms, a workflow to cover a ~2,500 m^2 survey site captured at 1 mm/pixel resolution would be: (1) tiling into 2 m^2 patches, (2) passing each patch to the AIF PebbleCountsAuto tool with quick manual steps of shadow-masking and sand-clicking (if sand is present), where each tile takes 1-2 minutes, (3) selecting a random subset of ~20 tiles to pass to the KMS PebbleCounts tool as validation and uncertainty estimation for the AIF approach. Such a workflow could be accomplished in 1-2 days of work by an experienced user, providing tens- to hundreds-of-thousands of measured grains from the survey site and a robust measurement of the full GSD. To increase processing speed, a gridded subset of tiles could also be extracted from the full survey site, with a 3-5 m step size between patches, to provide complete coverage across heterogeneous gravel-bar features, while avoiding unnecessary over-sampling and processing of every patch in the survey site.

5- Improve the discussion. The discussion needs a much improved start and overall reorganisation. A good rule on writing a discussion is to start with a sentence or two that distil the major findings that you want the reader to take away from this paper. A discussion needs to go over the substantive elements of the findings and their meaning and contrast to the work of other authors before going into issues. Here, the authors need to start the discussion with a section that tells us what they have achieved and gives the reader a better sense of how PebbelCounts compares to other methods. Without necessarily

running other methods on their data, we at least need a summary table that presents errors reported in literature and compares them to the current work. This is the section where you need to show an enhanced understanding of the literature.

While we agree that the discussion could use some reorganization (see for example our response to point 4 i-iii above) and a better introduction, we would like to avoid tables with exhaustive lists of the errors reported in other studies. In particular, we do not want to make comparisons between errors from the described image segmentation approach and texture based (e.g., roughness, entropy, semivariance) approaches, since these other methods are based on correlative relationships, rather than direct measurements of the grains.

A comparison with other image segmentation studies is made especially complicated by the tendency in different studies to sometimes only report the bias without the error spread, use different metrics of uncertainty reporting, and/or report uncertainties in mm rather than psi units. For example, unfortunately, the only study we found that reports the accuracy of Basegrain versus control data is that of Westoby et al. (2015), however, they only provide percentile bias numbers in mm with no measure of spread.

We feel it is more useful to report a few aggregate numbers from these studies, which together demonstrate the PebbleCounts algorithm to be within the range (and even on the low-end) of previously reported uncertainties. Ultimately, the uncertainty in measurement is highly dependent on the input image quality and complexity (range in grain size, angularity, intra-granular variability) and providing blanket estimates is less useful than end-users applying the KMS tool to a subset of images to validate the results of the AIF approach.

To clarify these points and provide a better intro for the discussion, we propose a modification of Section 7 and 7.1, which begins at P23, L6:

7 Discussion

In this study we developed two new methods for grain-size measurement with low uncertainties and the potential to deliver full GSDs from complex images of high-energy mountain rivers. Our open-source Python-based algorithms perform equally well to other image segmentation tools, but can be applied more quickly over larger areas surveyed by the SfM-MVS workflow we present. Critical to success is the application of a strict lower cutoff, which limits the minimum measurable b-axis grain size to 20-times the pixel resolution. The automated version of the algorithm delivers less accurate measurements, but these can be limited by using low-blur, higher resolution imagery. We focus our discussion on the comparison of our approach with similar work, the effect of the lower truncation on GSD estimates, and practical guidelines for acquiring imagery and applying PebbleCounts, including the application of UAV surveys.

7.1 Performance of KMS and AIF

For comparison of our algorithms to previous work, we do not consider errors reported in studies using texture-based measurements (e.g., Woodget at al., 2018), since these methods are based on correlative relationships rather than physical measurement of each grain. Similar to other image segmentation methods (Butler et al., 2001; Graham et al., 2010), the KMS PebbleCounts approach undercounts grain sizes in each respective size class. This undercounting does not undermine the resulting GSDs and associated percentile estimates, so long as an appropriate lower truncation is defined. This cutoff was found to be 20 pixels (compare to 23 pixels found by Graham et al. (2005a)) in b-axis length (Fig. 13), which explains the degradation in 3--5 mm counting in the reduced resolution lab images (Fig. 8), where the smallest pebbles were only a few pixels in size as resolution was decreased.

As shown in Figure 16, when we apply this cutoff and exclude poorly performing images we find an average m (bias) and e (spread) of 0.03 and 0.09 psi, respectively, for the ~1.16 mm/pixel imagery and 0.07 and 0.05 psi for the 0.32 mm/pixel image. For the AIF approach these values are 0.13 and 0.15 psi for the ~1.16 mm/pixel imagery and -0.06 and 0.05 psi for the 0.32 mm/pixel image. These are averages, which actually increase at higher percentiles in agreement with other image segmentation methods (e.g., Sime and Ferguson, 2003). We thus suggest higher error budgets at higher percentiles.

As demonstrated in Figures 18 and 19, there are significant inaccuracies associated with the AIF approach. The errors associated with the AIF approach can be limited when applied to high-quality (low-blur) ~1 mm/pixel resolution imagery, with better results possible on < 0.5 mm/pixel imagery. Ultimately, the uncertainties are highly dependent on the input image quality and complexity (range in grain size, angularity, intra-granular variability) and providing blanket estimates is less useful than end-users applying the KMS tool to a subset of images to validate the results of the AIF approach.

In spite of this caveat, our bias values of 0.03-0.13 psi are in the range of previously published absolute biases of 0.007--0.33 psi from similar techniques (see Table 2 in Graham et al. (2010)). To our knowledge, the only study to compare Basegrain results to control data by Westoby et al. (2015), makes comparisons in mm rather than psi units. Since the psi scale is logarithmic, in our study the error in mm increases with psi from ~0.8 mm uncertainty at 4.5 psi (23 mm) to ~7 mm uncertainty at 6.5 psi (91 mm) for the ~1.16 mm/pixel imagery in the KMS case. Westoby et al. (2015) report similar bias from Basegrain, again increasing in magnitude at higher percentiles. Regarding the error spread reported in the literature, our range of 0.05-0.13 psi is less than the 0.25 and 0.14 psi values reported by Sime and Ferguson (2003) and Graham et al. (2005b), respectively, for their image segmentation techniques.

6- 3D data The section at the end of the discussion on the integration of 3D information does not sit well. Move it to the introduction with the intention of stating that you will not be using 3D clouds or DEMs. It would be worth citing the work of James Brassington and Damia Vericat who have developed particle sizing based on TLS. But also you could mention that Woodget et al (2018), already cited, found that 3D information did not

improve particle size estimates. This section could be the place in the intro where you discreetly place a few SfM/drone citations but just to further justify that this will be an image-based method.

We have significantly shortened this section on point clouds to a few key points and moved the majority of the information, including the current Figure 20, to the supplementary material. The remaining section is now in the introduction. We have incorporated the suggested citations. We have created a new section in the introduction preceding the current Section 4 where the PebbleCounts algorithms are presented. This will be the new Section 4 and reads as follows (we also include the new Section S1 below):

4 Additional Data Dimensions from Point Clouds

As mentioned in Section 2, previous authors have attempted to incorporate roughness from point-cloud data into measurements of average grain size (e.g., Brasington et al., 2012), which has potential if the range in sizes is large enough to be expressed in 3D in the point cloud (e.g., Woodget et al., 2018). Such work highlights the potential to exploit third height dimensions from irregularly spaced point clouds generated via lidar or SfM-MVS, but stops short of object detection and segmentation. We briefly summarize key points we found in this regard and direct the reader to the supplementary material Section S1 for a full description.

Our efforts to incorporate height information were complicated by vertical noise (scattering around a mean value) inherent to the SfM-MVS technique derived from a limited set of overlapping photos. Vertical standard deviations from flat target surfaces in our field data were ~1.7 mm, and likely much higher on steeper grain surfaces. It is possible to get lower values of 0.2 mm with many more oblique images taken under ideal conditions at close range (e.g., Cullen et al., 2018; Verma et al., 2019), however, for field surveys this is not feasible while also covering large areas. As the point cloud actually has a lower resolution (since it is based only on matched points) and more vertical noise than the orthomosaic (which exploits the full camera resolution), the imagery alone provided more detail. This is particularly important around grain edges needed for segmentation, which are not captured in top-down imagery alone (Figure S1). To conclude, the potential for additional data dimension integration into pebble counting may be possible using higher dimensional object detection schemes, but, for the time-being, the orthoimagery alone provides satisfying results.

Supplement Section S1

The results presented here are similar to other studies segmenting grains from 2D imagery. This ignores the potential to exploit the third height dimension of the data from irregularly spaced SfM-MVS point clouds and associated DEMs. Many authors have already begun to look at patch-scale variance or roughness (e.g., Rychkov et al., 2012) from point clouds on gravel-bed rivers to determine bulk characteristics, but this stops short of object detection and *segmentation. Here, we briefly describe some of our own efforts to incorporate this additional information into PebbleCounts.*

Our simplest approach was including the gridded DEM information, resampled to the same resolution as the orthomosaic. We inverted the elevation raster and flood-filled from the lowest points (tallest grains) using watershed approaches, conceptionally similar to lidar tree-detection algorithms (e.g., Chen et al., 2006; Alonzo et al., 2015). For large, prominent grains with semi-spherical shapes, the flooded area was found to linearly increase until reaching the grain boundary, at which point the rate of area change jumped. We explored this break point as a potential segmentation tool for larger grains, but found that in the complex natural setting the shape of most grains is far from spherical, and furthermore, overlapping grains led to inconsistent behavior in the area breaks.

In an additional approach, we calculated both roughness and curvature at a variety of scales (5, 10, 50, 100 mm) directly from the point cloud using the open-source CloudCompare software (CloudCompare, 2018). This information was then gridded into a raster of the same resolution of the orthomosaic. While roughness could at times identify the smoother sand patches, it was difficult to discern between a sand patch and flat rock, and a color threshold on the orthoimagery was more successful. Curvature showed some spikes at grain boundaries, with the potential to aid in edge detection, however, we found that curvature was also high on intra-granular features.

In general, this analysis was complicated by vertical noise (scattering around a mean value) inherent to the SfM-MVS technique in the generation of dense point cloud data. In the field, for ~9 photos taken from a height of ~5 m, the vertical standard deviation of points on a detrended flat surface (one of our coded targets) was found to be 1.7 mm for 13,014 points. On the other hand, in the perfect lab setting with 16 photos from ~1.5 m, the detrended flat carpet around the pebbles achieved a standard deviation of 0.2 mm (33,371 points), similar to other SfM-MVS studies using large numbers of carefully collected images (e.g., Cullen et al., 2018; Verma et al., 2019). These standard deviations from detrended flat surfaces represent a best-case scenario, whereas, in our field setting, the vertical uncertainty on the complex, overlapping pebbles is likely higher. Such vertical noise is absent from the orthomosaics and limits the applicability of point clouds at these scales.

Ultimately, as the point cloud actually has a lower resolution (since it is based only on matched points) and more vertical noise than the orthomosaic (which exploits the full camera resolution), the imagery alone provided more detail. This is particularly important around grain edges needed for segmentation, which are not captured in top-down imagery alone, as shown in Figure S1. The lab setting resulted in point clouds with sufficient density and precision to identify individual grains with point-cloud processing tools. Thus, achieving higher quality SfM-MVS point clouds is possible, but only through more intense data collection during fieldwork (Fig. S1).

Alternatively, lidar point clouds with distance measurements based on phase shifts have a lower standard deviation of ~1 mm in multiple settings and distances (up to ~300 m) and could allow *more precise delineation using roughness and curvature calculations directly on the point cloud, however, such devices remain costly. Additionally, the development of affordable hyperspectral cameras with additional wavelengths will help in image segmentation in the spectral domain. To conclude, the potential for additional data dimension integration into pebble counting may be possible using higher dimensional object detection schemes, but, for the time-being, the orthoimagery alone provides satisfying results.*

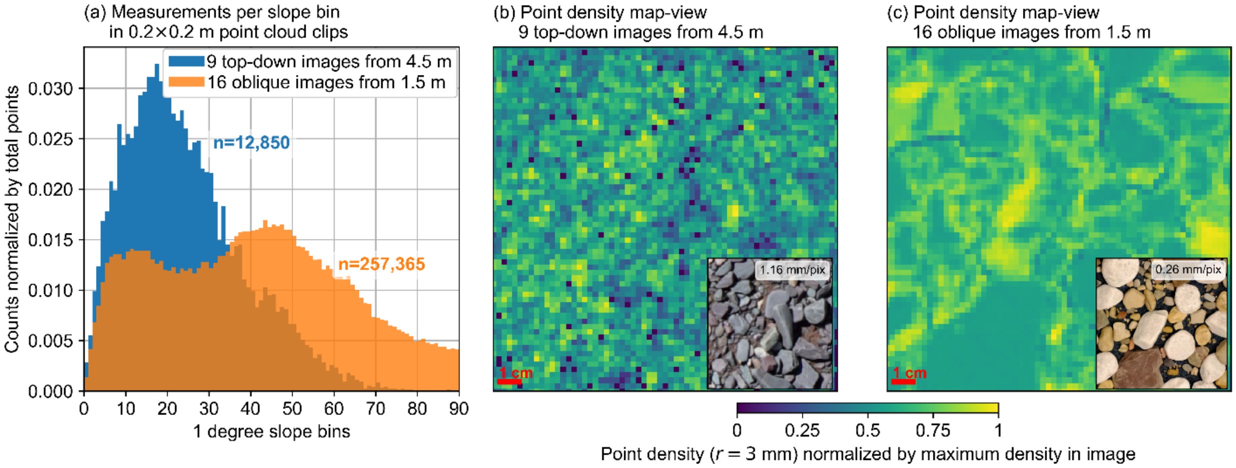

Figure S1: (a) Slope distribution in field (top-down) and experimental (oblique) point cloud clips. The point cloud slope was calculated in \textit{CloudCompare} \citep{cloudcompare} by first calculating the normals at each point using the 6 nearest neighbors and then extracting the dip of each normal. (b) Map-view of point density normalized by the maximum for the 9 top-down field images and (c) the same for the 16 oblique experimental images. Point density was calculated as the number of points in a radius of 3 mm. The clips were from a 0.2\$\times\$0.2 m area, visually selected to have similar grain sizes and numbers of grains, shown in the inset images in (b) and (c). The average point density for the 16 oblique photo setting was 59 points/cm\$^2\$, whereas, in the field using 9 top-down photos the density was 17 points/cm\$^2\$. Note the higher point density on grain edges in (c) compared to (b), which are important for segmenting grains directly on the point cloud.

Reviewer Pascal Allemand

This paper describes an interesting open source software for automatic measurement of grain size distribution from images. Compared to existing systems, this open source software is able to work on ortho-images obtained by phogrammetric methods on image collections covering wide areas. The algorithm seems very efficient but the results are deserved by the text which is too long and a discussion where key problems are downed out among less important elements. I suggest to the authors to re-write the paper in a more concise and linear way.

We appreciate that the reviewer finds the work relevant and sees its potential for grain-size mapping. However, we disagree that the paper needs major rewriting. As mentioned in our first reply, our goal in having such a high level of detail is in reporting every step in a very complex study. Interested users will be able to follow every step we have taken from data collection to algorithm development and assessment via a close reading. Hopefully some of the confusion was eliminated in the large amount of discussion rewriting and reorganization that we accomplished in reply to the first reviewer. Nevertheless, we have tried to accommodate the comments of the second reviewer in some of the points below.

P4 line 2: How to be sure that the detected grains are representative of the whole grains? It is a point to discuss.

We have hand-clicked every visible grain in the control images, thus representing the true distribution of the grains. We can be sure that the grains detected using the image segmentation approaches are representative of the whole distribution because these grain-size distributions match very well to the hand-clicked control data as shown in, for example, Figure 14. We add a sentence at P4, L2:

Despite the selection of fewer grains, Figure 2 demonstrates that these grains do represent the entire distribution through the close match in GSD between hand-clicked and KMS results.

P4 line 6: The fact that current methods are limited to some m2 is a real limitation that should be indicated in the part concerning the current methods.

We clarify this point at the start of Section 3 by modifying the first paragraph beginning at P3, L24:

Watershed segmentation is effective for interlocking, uniformly colored, oblate grains, however, energetic gravel-bed rivers in mountains often have more complex grain compositions with intra-granular variation, irregular shadowing, and a large range of sizes. The automated watershed methods proposed suffer from over-segmentation, grain misidentification, and the need for significant, time-consuming post-processing (e.g., in Basegrain with the split, merge, and delete tools) when applied to complex images. These issues limit the application of previous methods at areas > 10 m^2.

Figure 1: the differences in results between AIF KMF and water shed methods should be discussed in the discussion

We feel that this discussion belongs in the introduction. Our algorithm does not use the watershed method and this section and Figure 1 and 2 are intended to highlight the reason why we avoid this technique before we go on to explain the steps our algorithm does take, which is certainly introduction material. We have however, added a sentence to the end of Section 7.1 in the discussion:

Importantly, we emphasize that the previous image segmentation techniques discussed here all rely on the watershed segmentation step, whereas, neither of our algorithms use this step for the reasons demonstrated in Figures 1 and 2.

Figure1: Concerning the watershed method (basegrain?), you show, I thing, the gross results. The results can be filtered with basegrain by post processing.

Yes, the results can be post-processed, but the point here is that the post-processing is time-consuming, subjective, and limits the applicability of Basegrain to larger areas. We have highlighted this point with the modified paragraph mentioned in the P4, L6 point above.

Page 5 line 8: What type of denoising method do you use? Does it preserve edges?

This is covered at P5, L12-13:

…chromaticity bands from this color space undergo bilateral filtering (Tomasi and Manduchi, 1998) to preserve inter-granular edges while further smoothing color.

Page 5 line 10: how do you filter the sand patches? on color? on texture?

This is stated at P5, L10:

…HSV color selection for sand-patch masking.

We clarify by adding:

…HSV color selection for sand-patch masking (whereby sand is filtered by a narrow, user-selected color mask).

Figure 3: I suggest to merge figure 3 and figure 6 and to shorten the text referring to the manual user of your software.

We feel that the text is sufficient and should not be cut for the sake of clarity to the user. We will merge Figure 3 and 6 as shown here:

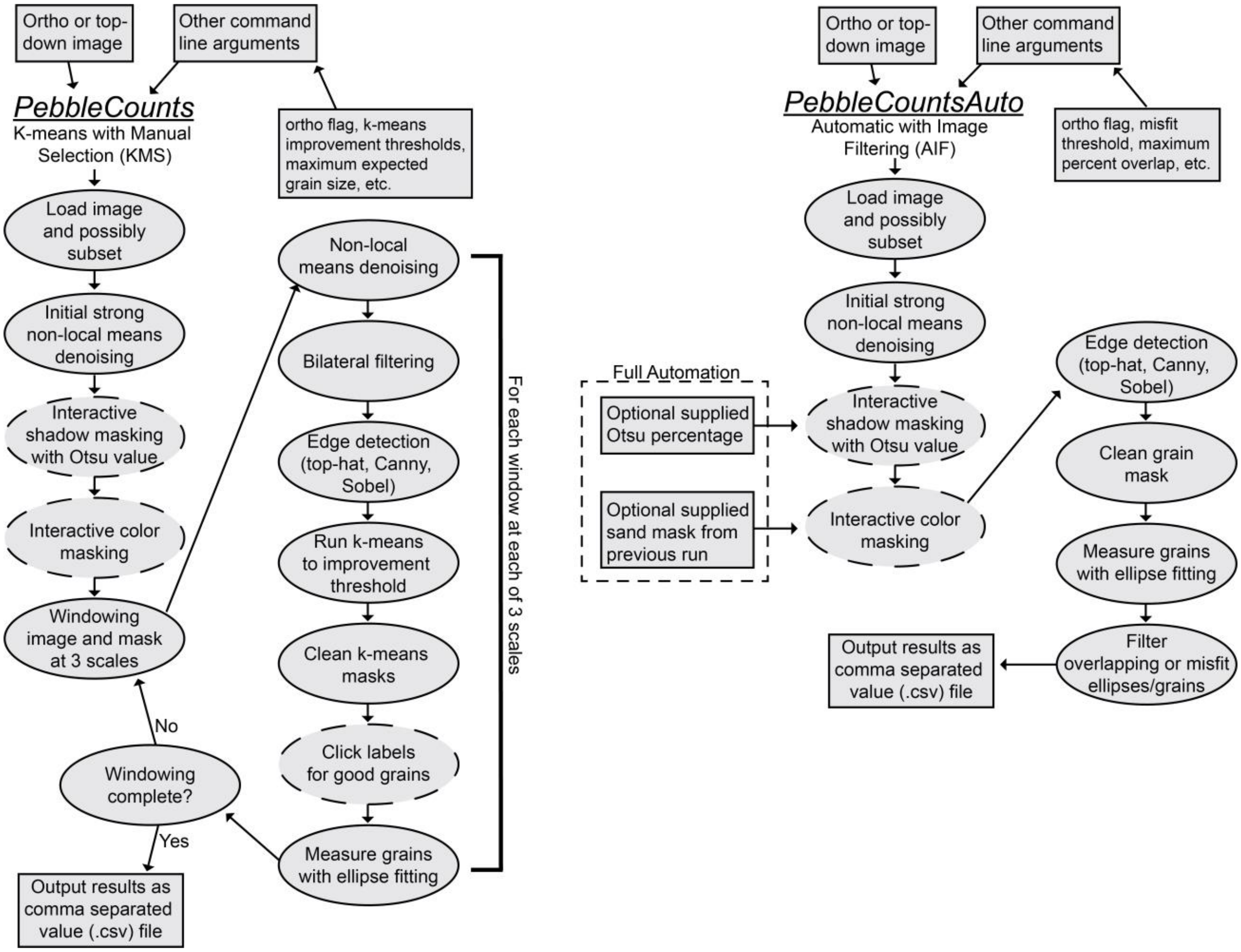
Ortho or top-down image
Other command line arguments
PebbleCounts
K-means with Manual Selection (KMS)
ortho flag, k-means improvement thresholds, maximum expected grain size, etc.
Load image and possibly subset
Initial strong non-local means denoising
Interactive shadow masking with Otsu value
Interactive color masking
Windowing image and mask at 3 scales
Non-local means denoising
Bilateral filtering
Edge detection (top-hat, Canny, Sobel)
Run k-means to improvement threshold
Clean k-means masks
Click labels for good grains
Measure grains with ellipse fitting
For each window at each of 3 scales
No
Windowing complete?
Yes
Output results as comma separated value (.csv) file
Ortho or top-down image
Other command line arguments
PebbleCountsAuto
Automatic with Image Filtering (AIF)
ortho flag, misfit threshold, maximum percent overlap, etc.
Load image and possibly subset
Initial strong non-local means denoising
Full Automation
Optional supplied Otsu percentage
Optional supplied sand mask from previous run
Interactive shadow masking with Otsu value
Interactive color masking
Edge detection (top-hat, Canny, Sobel)
Clean grain mask
Measure grains with ellipse fitting
Filter overlapping or misfit ellipses/grains
Output results as comma separated value (.csv) file

Figure 3. Flowchart of PebbleCounts (left) and PebbleCountsAuto (right). The boxes are user supplied input or output from the algorithm. Dashed lines indicate a user input step during processing, either entering and checking values or clicking.

Figures 4 and 5: difficult to read and not necessary. I suggest to remove or to rework in a more concise and readable way

We feel that these figures are useful and provide the user with an idea of how running the algorithm actually looks. This goes into our point about high level of detail in the manuscript so the interested user can follow along very well upon close reading. In a digital version of the study, these images can be zoomed into, which yields high quality vector and raster graphics (tested at 300% zoom in Adobe Acrobat Reader).

Concerning "5 Calibration and Validation Test I: Controlled Experiment": shorten and get to the point. The part concerning the cameras is not useful. What is important is the result (description of the Photoscan parameters is useless for example). Size of pixels do not

matter. What is important is the ratio between the resolution of the image to the size of the smallest grain detected.

We disagree with the reviewer here, and feel that the discussion of camera types and our experimental setup is very useful to users that will want to apply the method directly and repeat the processing for their own field sites. This again goes towards the thoroughness of our study, where we have left none of our processing steps out.

Same for “6 Calibration and Validation Test II: Field Surveys”. I suggest to remove the useless details and to go to the point. You could show only the better and the worth examples and discuss why the “best” example give good results and why the “worst” example give no such good results (but good anyway ïA¿Ł)

We feel that a close reading of the section demonstrates the good and bad results and the reasons for them. One example we point to here on P19, L7-9:

Importantly, S24 is the only site not from a major river stem, but rather from a debris-flow fan draining a small tributary catchment in the Quebrada del Toro. S34 also had a high Adiff=−2.11. In this case, poor performance is due to significant blurriness of this image, and again a small sample size (n=47).

Figure 19 (they are too many figures): for me, what is important is to discuss why its work or not, in what case and How I can use your software and what error can I expect, by adding some advices on the acquisition procedure I should follow. These points are discussed in the current version but are not enough highlighted.

We feel that this figure is demonstrative of the difference in the AIF and KMS routines and very instructive to the end user concerned about image quality and how it will affect the results from each technique. Regarding advice for acquisition, we have rewritten a large part of the discussion in response to the first review (see above). The point cloud integration has been removed which should add to the discussion clarity, and we end the discussion with a clearly outlined section of the “Practical Considerations for Image Collection and Processing”.