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.

Following both responses is a marked-up version of the newer manuscript compared with the previous. Since a lot of material was also moved to the supplement, the marked-up manuscript is also followed by the new marked-up supplement.

Reviewer Patrice Carbonneau (Second Report)

There remains significant issues with the question of the angle of views. First, top-down is a colloquial term and I remain firm that is should not be used in literature. Nadir is the correct term. Form most of the history of aerial imagery, there has always been error in the angle of acquisition. When aircraft, including drones, try to collect nadir data, there will still be fluctuations around the target angle of view. Having errors in that view is not a good reason to use the wrong nomenclature. If the authors are concerned that the camera mast setup induces error, then their images could be described as 'slightly off-nadir' or 'near-vertical' but nadir would still be fine because it is understood that no airborne camera can be controlled to have perfect orientation. This actually raises a question: To what extent are these images off-nadir? A number is not provided. In the current revision, there is in fact a problem of clarity in this case. When using the camera mast (or in the lab) are the authors deliberately trying to acquire images that are significantly oblique, i.e. actively tilting the camera off-nadir? If not, this is a nadir setup and the authors need to discuss the off-nadir component as a measurement error. This could in fact play in their favour by mitigating the doming effect in SfM-photogrammetry (See the James and Robson in ESPL, 2014). The authors need to:

1- Stop using the term top-down

2- Clarify if they are actively taking oblique imagery.

We understand the concern of the reviewer with our terminology and think we can provide a solution by using the terms near-nadir, off-nadir, and oblique. We have deliberately chosen the term top-down and would like to refrain from using the term nadir, because these images were not always taken in nadir direction. As geologists and geophysicists, we are aware of the clear and precise definition of nadir (and zenith). These are perpendicular to an equipotential surface (for example the geoid) and nadir refers to an observation method that usually points in the direction of the force of gravity. This geometry is achieved by satellites and airplanes and is

often verified by additional measurements taken at the same time (in the case of historical aerial photography, level measurements with water bubbles were taken at the same). While we would have liked to achieve this geometry and we were striving for it, our setup does not guarantee that all of our photos were taken in precise nadir position. They may have been off by a few degrees to any side.

To clarify: We always tried to have the camera pointed directly downwards with the mast setup, but over the course of several hundred to thousand photos collected by hand, there were inevitably minor tilts (< 10°). It is true that these slightly oblique photos likely contributed to improved quality of the point clouds, but it was not our intention to tilt the camera significantly, and instead relied on the ground control targets to limit doming-effects. In addition, for the handheld lab setup, we intentionally collected ~10 images at a ~20° off-nadir angle, to get the sides of the pebbles in addition to their tops. In light of these points, we have changed top-down to near-nadir, off-nadir, or oblique throughout the manuscript (and in the figures), and changed the explanatory text in our new "Section 3.2. Orthomosaic Generation" (formerly part of a subsection "6.3.1. Top-Down Images") P8, L6:

For each test setup, we collected ~10 images from ~20° off-nadir (oblique) and at least 4 overhead near-nadir (tilts < 10°) pictures, for 12-16 photos in total. The collection of oblique images aided in removing doming effects from the resulting point clouds (e.g., James and Robson, 2014) and for capturing the pebble edges and sides (Fig. S1).

In reference to the mast photos, we have clarified at P12, L6 with the sentence:

We refer to the images as near-nadir, rather than nadir, due to the fact that during mast photo collection some unintentional tilting of the camera (< 10°) occurred. These near-nadir photos aided in removing doming effects, but did not allow us to capture the sides of pebbles as in the oblique images taken in the experimental setup (Fig. S1). Capturing oblique images of every patch in the field sites would require infeasible amounts of time and processing power.

The authors have much improved the discussion of their errors in relation to other published work. But their refusal to provide a summary table is not justified nor is it reasonable. The authors are right that different papers report error differently, but that is not a good reason not to summarise information. There is abundant precedent for such tables in other similar works (see Dugdale et al in RRA on 'Aerial Photosieving'). The point is to give a sense, even if qualitative, of relative performance. The textual discussion of this topic is very good. But, I repeat, please provided a summary table, many of your readers will appreciate the condensed information and better absorb your points in this manner.

A table has been added to the first discussion section (Table 1). There we briefly summarize the results of other authors who used similar segmentation methods. We do not include comparisons with texture-based methods, given the significant differences between segmentation- and texture-based methods, and the many differences between each texture method (semivariance, roughness, wavelets, etc.). We also add an explanation in the discussion P20, L2:

For comparison of our algorithms to previous work, we do not consider errors reported in studies using texture-based measurements (e.g., Woodget et al., 2018), since these are based on correlative relationships rather than physical measurement of each grain. Texture methods work well for homogeneous pebble arrangements in lower-energy settings, but high-energy mountain rivers with heterogeneous pebble arrangements and large ranges in sizes require segmentation approaches.

The discussion of UAV applications is also much improved. But section 8.3.2 should be just 1 paragraph. The second paragraph repeats most of the ideas of the first and it seems like the first was dropped in without a proper reading of the second. This just needs to be merged and have repetitions smoothed out.

We appreciate the close re-reading of our manuscript and have changed this section to read P22, L25:

The > 20 m flight heights typical of UAV surveys lead to cm-scale imagery with currently available 12–24 megapixel cameras, which is less appropriate for PebbleCounts processing, unless large (> 0.2 m) cobbles and boulders dominate the river site. Acquiring 0.5 mm/pixel imagery from a DJI Mavic drone with a 12 megapixel camera requires a very low flight height of ~1.4 m, giving a field of view of only ~1.5x2 m. This may be improved using better cameras like on the Mavic 2 Pro (20 megapixel camera), but gathering 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. Given continual technology improvements (e.g., greater battery life, more accurate geo-tags from onboard dGPS, higher megapixel cameras, and reduced motion blur), it is within reason to expect hectare to multi-hectare SfM-MVS UAV surveys at mm resolution in seamless orthomosaics along entire river reaches in the near future. But, for the time-being, a single, non-overlapping orthoimage workflow proposed by Carbonneau et al. (2018) has high potential to achieve large-areal results. Their workflow, building on Carbonneau and Dietrich (2017), uses a number of high and oblique overlapping flights to orthorectify a lower non-overlapping flight with mm-scale acquisition, with resulting single, scaled images passed to Basegrain, or, alternatively, to PebbleCounts.

Associate Editor Eric Lajeunesse

Thank you for submitting a new version of your manuscript entitled "Introducing PebbleCounts: A grain-sizing tool for photo surveys of dynamic gravel-bed rivers" to ESurf.

I have now received a second report from Patrice Carbonneau. He feels that the manuscript is much improved. Yet he reports a detailed list of small issues which need to be resolved before the paper can be published in ESurf. In particular, I do agree with him that a table summarizing the discussion about the relative performance of your approach would help to synthesize your message.

We have noted this concern and have inserted Table 1 in the discussion section, as mentioned in our above response to Patrice Carbonneau. This summarizes the results of our study in comparison with other similar segmentation-based pebble counting techniques.

I have noted your resistance to take into account several of the comments raised by Pascal Allemand. Yet, I agree with his observation that the manuscript is difficult to follow in many places. Esurf is dedicated to a broad audience of scientists working on Earth surface processes. In its present version, the manuscript is closer to a technical report aiming at a more specialized community. Several points contribute to this feeling:

To preserve the detailed information, some of the original manuscript has now been moved to the supplementary information. For instance, we have placed the detailed Agisoft processing steps in a new supplement Section S4. See the marked-up revised manuscript and supplement at the end of this document for further details. We agree with and support publishing a widely accessible manuscript, but feel that providing details of processing steps are important for generating reproducible science.

- the manuscript contains too many figures (18);

We appreciate this point and have made an effort to reduce the number of figures by moving four to the supplement. There are now 14 figures in the manuscript.

- it makes use of too many acronyms (GSD, SfM-MVS, MP, AIF, KMS, NMAD, ….),

We have changed GSD to grain-size distribution and MP to megapixel throughout the manuscript. We think that the acronyms UAV (unmanned aerial vehicle), SfM-MVS (structure-from-motion with multi-view stereo) and NMAD (normalized median absolute difference) are commonly used within the quantitative geomorphic community, and we feel the broad audience should have exposure to them. Our use of KMS (k-means with manual selection) and AIF (automatic with image filtering) short names to distinguish our two techniques is necessary and widely used through the entire manuscript and supplement. These acronyms make the text more

readable, but we have tried to carefully go through and remove any unnecessary or redundant usage of acronyms.

- the level of sectioning is too high (8.3.1, etc…) and many of the section titles are inappropriate ('previous work on photo sieving', 'Motivations for new methods', `7.4.3. Results: Handlhed Image', etc…).

To reduce the sections, we have combined many subsections and also given the resulting sections more broad and consistent names.

Here are more specific comments:

• Figure 4 and 5 are certainly appropriate for a software manual. But they are not suitable for a journal like Esurf. I strongly advise you to move them into the supplementary information.

These figures are now in the supplementary material.

• Again, many of your section titles are inappropriate for a scientific journal. Here are a few exemples: « 2. Previous work on photo sieving », « Motivation for new methods », « 4. Additional Data dimensions from point clouds », « 5. The algorithms »,… In many cases, you can solve this problem by suppressing these sections and merging them into a larger one with a broader title. As an example, section 2 and (maybe) 3 could be included as part of the introduction. Similarly, there is no need to divide subsection 6.3 (the title of which « images » is again quite clumsy) in 2 subsubsections. These 2 exemples illustrate a more generic problem that you must address.

We have merged a number of subsections into larger sections and also moved some sections (e.g., Point-Cloud Dimensions) to the supplement. See the marked-up revised manuscript and supplement at the end of this document for further details.

• The accumulation of field sites on Figure 14 and 15 is to the detriment of the message. You should restrict these figures to a couple of emblematic results illustrating your message, and use the supplementary information file to present the integrality of your results.

We have moved Figure 14 with the site curves separated to the supplement (Figure S7). Figure 11 shows the aggregated results and the reader is now referred to the new Section S5 in the supplement for the separate results. We prefer to keep all of the data points in Figure 15 (now Figure 12), since this shows the final errors associated with the methods (and referenced in our new Table 1). Although this figure is dense, it is important for understanding the error

distribution between both methods. We refer the reader to the new Section S5 and Figure S7 in the supplement for the detailed curves from each site showing where these data points originated.

• Although the caption of figure 1 starts with « Conceptual », this figure does not provide any conceptual information. It merely illustrates the difference of results obtained using different methods.

We have removed the term conceptual in reference to this figure.

In conclusion, I understand the need for details, but details are sometimes to the detriment of clarity. I therefore encourage you to submit a suitably revised version of your manuscript taking into account the remaining issues. Upon submission, I will need to receive a response file that lists each of the comments and describes how the manuscript has been modified (or not) in response to those comments.

We appreciate the reviewers' and your effort to improve the manuscript quality. These comments are well taken and we have shortened our manuscript accordingly. We would like to point out that the motivation for our detailed descriptions were to generate a reproducible scientific product that can applied elsewhere. We see now that some of our efforts provided too much detail and have subsequently moved material to the supplementary material for the interested reader. See the following marked-up manuscript and supplement for details on our revisions. We hope that the manuscript is now more readable to a general audience while also retaining the vital details to follow along step-by-step.

Sincerely,

From both authors,

Benjamin Purinton

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

Benjamin Purinton[1] and Bodo Bookhagen[1]

[1]Institute of Earth and Environmental Science, Universität Potsdam, Potsdam, Germany

Correspondence: Benjamin Purinton (purinton@uni-potsdam.de)

Abstract. Grain-size distributions are a key geomorphic metric of gravel-bed rivers. Traditional measurement methods include manual counting or photo sieving, but these are achievable only at the 1–10 m^2 scale. With the advent of ~~unmanned aerial vehicles~~ drones and increasingly high-resolution cameras, we can now generate orthoimagery over hectares at ~~sub-cm~~ mm to cm resolution. These scales, along with the complexity of high-mountain rivers, necessitate different approaches for photo sieving. As opposed to other image segmentation methods that use a watershed approach~~to automatically segment entire images~~, our open-source algorithm, *PebbleCounts*, relies on k-means clustering in the spatial and spectral domain and rapid manual selection of well-delineated grains. ~~The result is improved~~ This improves grain-size estimates for complex river-bed imagery, without ~~any~~ post processing. ~~In a second step, we~~ We also develop a fully automated method, *PebbleCountsAuto*, that relies on edge detection and filtering suspect grains, without the k-means clustering or manual selection steps. The algorithms are tested in controlled indoor conditions on three arrays of pebbles and then applied to 12 × 1 m^2 orthomosaic clips of high-energy mountain rivers collected with a camera-on-mast setup (akin to a low-flying drone). A 20-pixel b-axis length lower truncation is necessary for attaining accurate grain-size distributions. For the k-means *PebbleCounts* approach, average percentile bias and precision are 0.03 and 0.09 ψ, respectively, for ∼1.16 mm/pixel images, and 0.07 and 0.05 ψ for one 0.32 mm/pixel image. The automatic approach has higher bias and precision of 0.13 and 0.15 ψ, respectively, for ∼1.16 mm/pixel images, but similar values of −0.06 and 0.05 ψ for one 0.32 mm/pixel image. For the automatic approach, only at best 70% of the grains are correct identifications, and typically around 50%. *PebbleCounts* operates most effectively at the 1 m^2 patch scale, where ~~the algorithm can be rapidly~~ it can be applied in ~~∼5 minutes in many small areas~~ 5–10 minutes on many patches to acquire accurate grain-size data over 10–100 m^2 areas. These data can be used to validate *PebbleCountsAuto* applied at the scale of entire survey sites (10^2–10^4 m^2). We synthesize results and recommend best practices for image collection, orthomosaic generation, and grain-size measurement using both algorithms.

1 Introduction

Gravel-bed rivers transport water, nutrients, and sediment downstream, linking high mountains to populated forelands. The grain-size distributions ~~(GSDs)~~ — and associated percentile diameters, such as the D_{50} and D_{84} — in a river reach are fundamental geomorphic metrics of these systems (e.g., Shields, 1936; Parker et al., 1982; Church et al., 1998). They are used to characterize aquatic habitats (e.g., Kondolf and Wolman, 1993), assess the impacts of human infrastructure like dams (e.g.,

Kondolf, 1997; Grant, 2012), calibrate theoretical models of river transport and erosion (e.g., Sklar et al., 2006; Attal and Lavé, 2006; Attal et al., 2015; Dunne and Jerolmack, 2018), and explore natural phenomena such as downstream fining (e.g., Paola et al., 1992; Ferguson et al., 1996; Rice and Church, 1998; Gomez et al., 2001; Chatanantavet et al., 2010; Lamb and Venditti, 2016), which is essential for nutrient transport and ecological diversity.

Accurate grain-size measurement is elusive in nature given the heterogeneity of gravel-bed rivers, particularly in steep mountain catchments where the range of grain sizes is large. Traditionally, ~~GSDs~~ grain-size distributions have been gathered via physical clast measurement and counting along grids (Wolman, 1954), lines (Wohl et al., 1996), or in $\sim$1 m^2 patches (Bunte and Abt, 2001), all truncated at some lower observable limit (e.g., Rice and Church, 1998). Not only are these techniques time consuming, prone to operator bias, and disruptive to the environment, but they also require large (hundreds of pebbles) sample sizes to accurately estimate the characteristic nature of the grains in each location (Wolcott and Church, 1991).

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 via Structure from Motion with Multi-View Stereo (SfM-MVS) (Smith et al., 2015; Eltner et al., 2016) at scales of entire river cross sections or reaches ~~at resolutions ator exceeding~~ with 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.

We build on previous work and introduce the addition of color-space clustering techniques to present efficient new semi-automated (*PebbleCounts*) and fully automated (*PebbleCountsAuto*) algorithms for grain ~~identification and~~ sizing from imagery in high-energy mountain rivers. Our algorithms are built on Python with a few popular libraries and are open source. The instructions and code can be accessed at: https://github.com/UP-RS-ESP/PebbleCounts (Purinton and Bookhagen, 2019). In this study, we present previous work on grain-size measurement from rivers and our motivation for new developments. The processing chains of *PebbleCounts* and *PebbleCountsAuto* are then discussed. We test the algorithms in controlled conditions and then in a more challenging field setting in the northwestern Argentine Andes. The limits and caveats of the method are discussed using imagery of varying resolution, and suggestions for photo collection and processing are provided.

2 ~~Previous Work on Photo Sieving~~

~~Manual digitization of each pebble was previously necessary for grain sizing from pictures (e.g., Kellerhals and Bray, 1971; Ibbeken and Se~~

~~.~~

1.1 Prior Studies

Modern digital grain sizing is divided into texture- and segmentation-based image-processing methods~~. Texture~~ , as opposed to previous manual digitization (e.g., Kellerhals and Bray, 1971; Ibbeken and Schleyer, 1986). Many texture methods rely on the relationship between grains and their shadowed interstices to derive size estimates over image windows. Examples include semivariance (Verdú et al., 2005; Carbonneau et al., 2003, 2004; Carbonneau, 2005), entropy or inertia calculated from gray

level co-occurrence matrices ~~(GLCM)~~ (Haralick et al., 1973; Carbonneau et al., 2004; Carbonneau, 2005; Dugdale et al., 2010; de Haas et al., 2014; Woodget and Austrums, 2017; Woodget et al., 2018), and autocorrelation (Rubin, 2004; Warrick et al., 2009; Buscombe et al., 2010). These methods only provide one estimate of grain size (e.g., D_{50}), which often requires site-specific calibration.

Buscombe (2013) achieved full ~~GSD~~ grain-size distribution measurements using wavelet decomposition~~on gray-scaled sand and pebble imagery, and also published their technique as~~ , and published an open-source Python tool, *pyDGS*. This is ~~another~~ a texture method that ~~does not measure each grain individually, and it is more apt for~~ has been designed for the analysis of thin sections or beach sands ~~, since it requires that each grain~~ and requires each grain to be fully resolvable and ~~that the distributions be relatively~~ the distributions to be fairly homogeneous in size and shape. ~~An additional texture method relies~~ Additional texture methods rely on the 3D texture (or roughness) of point clouds to relate the variance of bed-scale topography to average grain size (Brasington et al., 2012; Rychkov et al., 2012; Westoby et al., 2015; Woodget and Austrums, 2017; Bertin and Friedrich, 2016), however, ~~this technique~~ these techniques also requires site calibration and the relationships have been found to vary widely ~~depending on, among other things, grain sorting and packing~~ (Pearson et al., 2017).

In contrast to texture methods, the focus of segmentation is the full delineation and measurement of every visible grain. Segmentation is error prone in images that contain overlapping grains, a large range of grain sizes including sand patches, changes in landcover (e.g., vegetation), pebbles that are highly irregular in shape (non-ellipsoid), pebbles with intra-granular color variations or texture such as veins or fractures, and in which shadowing is irregular. Herein, we refer to these factors collectively as image complexity. Furthermore, segmentation-based methods also require high-spatial resolution point clouds or images that resolve the specific grain geometries. The benefits are that segmentation does not require any site calibration besides knowledge of the image scale and it provides a full ~~GSD~~ grain-size distribution and all the commonly used percentiles ($D_{5,16,25,50,75,84,95}$). Published methods ~~include the work of~~ by Butler et al. (2001), Sime and Ferguson (2003), and Graham et al. (2005a, b) ~~, all of which~~ all rely on edge detection followed by watershed segmentation and ellipse fitting to each separate grain ~~region~~ to get the long (a) and intermediate (b) ~~grain~~ axes. Detert and Weitbrecht (2012) added some sophistication to the ~~edge detection and watershed steps~~ algorithm of Graham et al. (2005a, b) and provide a free — though closed source — application called *Basegrain* for ~~the commercial software package~~ *Matlab*^{TM}, which has become a standard tool (e.g., Bertin and Friedrich, 2016; Bertin et al., 2017; Langhammer et al., 2017; Carbonneau et al., 2018).

2 ~~Motivation for New Methods~~

1.1 Motivation

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~~ their application to areas < 10 m^2.

In the interest of attaining GSDs from these settings and in images with a mix of clasts and sand patchesThus, we are motivated to develop a new semi-automated technique that uses k-means clustering of pixels and rapid manual selection of well-defined grains, herein referred to as the K-means with Manual Selection (KMS) or *PebbleCounts* approach, and a fully automated version that uses filtering of suspect grains, herein referred to as the Automatic with Image Filtering (AIF) or *PebbleCountsAuto* approach (Fig. 1). By avoiding over-segmentation and misidentificationassociated with the watershed approach, we are able to select fewer grains per image, but be sure that those selected are correctly delineated, thus improving the resulting GSD distribution (Fig. 2), with the intention of up-scaling to include many thousand grain measurements over large areas. Despite the selection of fewer grains, Figure 2 demonstrates that these grains do represent the entire distribution represent the true grain size through the close match in GSD between distribution with hand-clicked and KMS results.

Furthermore, faced with diverse camera models and the rise of SfM-MVS for the generation of georeferenced orthophotos, we wish to explore reasonable and appropriate combinations for covering hectare-sized acre to hectare areas while maintaining accurate measurement of characteristic GSDsgrain-size measurement. Fundamentally, our aim for the KMS approach is not in the delineation of a single high-resolution image from a $\sim$1 m^2 patch as in previous segmentation work, but rather a method that can cover areas of 10–100 m^2 containing complex grain arrangements, despite missing many grains at the patch scale. These semi-automated photo-sieving results can then be used to validate the AIF method at much greater spatial scales (10^2–10^4 m^2). This work serves as both a presentation of a new algorithm and a guide for the successful collection of GSDs in complex mountainous settings over large survey areas, where physical grain sizing is not feasible and previously reported image processing counting is infeasible and previous methods are unreliable or time consuming.

2 Additional Data Dimensions from Point CloudsAlgorithm Description

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 Our methods are similar to previous work by Graham et al. (2005a) and Detert and Weitbrecht (2012), with some key differences. A flow chart of both methods is shown in Figure 3 and the processing is presented briefly. We direct the interested user to the manual (https://github.com/UP-RS-ESP/PebbleCounts) for a full description of the steps. Our algorithms use 2D image processing in the spatial and spectral domains, which ignores 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 . The reader is directed to the supplement Section S1 for a full description. our efforts in this regard.

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 $\sim$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 and Bourke, 2019)

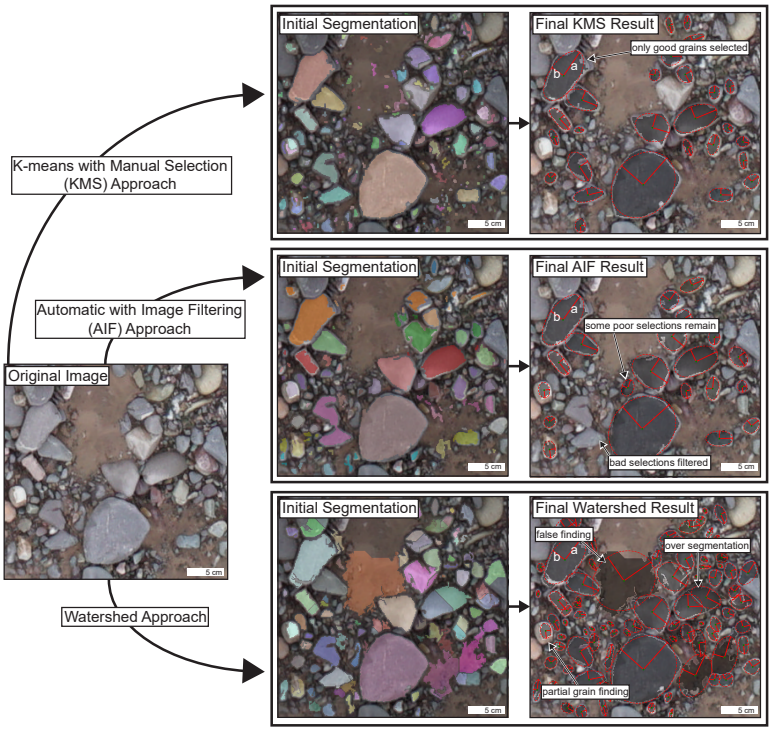

Figure 1. ~~The conceptual difference~~ Difference between our K-means with Manual Selection (KMS) and Automatic with Image Filtering (AIF) approaches versus a fully automated watershed segmentation approach on a gravel image from a high-mountain river. The a- and b-axes of each grain mask are found via an ellipse fit to the same area. Fewer grains are found in the KMS and AIF results, and there is still some misidentification in the case of AIF, but less than in the watershed result.

~~, 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, orthoimagery alone provides satisfying results.~~

3 ~~The Algorithms~~

~~The methods developed here hold similarities to previous work by Graham et al. (2005a) and Detert and Weitbrecht (2012), with some key differences. Processing is presented briefly, and we direct the interested user to the manual for a full description of the steps: (Purinton and Bookhagen, 2019).~~

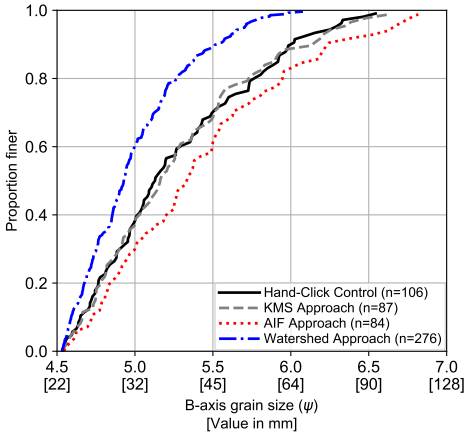

Figure 2. Watershed segmentation (blue, dashed and dotted line) versus KMS (gray, dashed line) and AIF (red, dotted line) approaches compared with a hand-clicked b-axis ~~GSD~~ grain-size distribution (black line) for a $\sim$1 m^2 river patch (S09 in Figure 6b). Watershed approach leads to over-segmentation of grains, giving an unreasonable number of clasts (276 versus 106 in the control) and an overly fine ~~GSD~~grain-size distribution.

2.1 *PebbleCounts*: K-means with Manual Selection (KMS)

~~The general outline of *PebbleCounts* is shown in Figure 3.~~ We employ the additional color spaces HSV (hue, saturation, value) and CIELab (Russ, 2002), aside from traditional RGB (red, green, blue) and gray-scale, to enhance differences in the spectral domain separate from lighting. First, the RGB image undergoes strong non-local means denoising (Buades et al., 2011) to smooth intra-granular color difference, interactive gray-scale shadow masking (Otsu, 1979) to separate obvious interstices, and HSV color selection for sand-patch masking (whereby sand is filtered by a narrow, user-selected color mask). The image and shadow/sand edge mask are then windowed for further processing.

At each window, the RGB image undergoes another weaker non-local means denoising, is then converted to CIELab, and the chromaticity bands from this color space undergo bilateral filtering (Tomasi and Manduchi, 1998) to preserve inter-granular edges while further smoothing color. Following this, edge detection on the smoothed, gray-scaled image occurs via a combination of top-hat, Sobel, and Canny methods with feature-AND selections (Russ, 2002), in which an edge is added to the full mask only if it overlaps with a found edge in the ~~shadow-~~previous edge mask, ~~sand-, or previous edge-mask,~~ thus piece-wise building an edge map while avoiding lone (i.e., intra-granular) edges (Detert and Weitbrecht, 2012).

After edge detection, our algorithm uses k-means clustering (Lloyd, 1982; Sculley, 2010) to further segment the pebbles. First, the matrix of non-masked pixels is converted into a vector that includes the spectral information at each location. This $N \times 4$ dimensional vector (N being the number of non-masked pixels) includes two spectral observables: the green-red and blue-yellow smoothed chromaticity bands from CIELab; and the two spatial observables: the x and y coordinates of the pixel

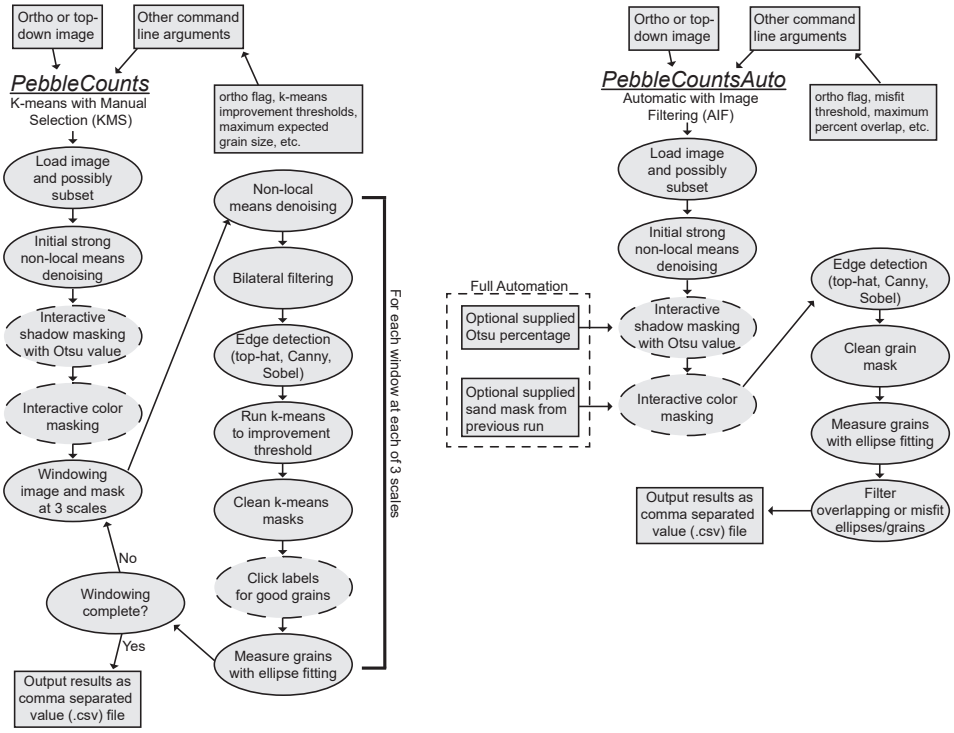

Figure 3. Flowchart of *PebbleCounts* (left) and *PebbleCountsAuto* (right). The boxes are user supplied input or output from the algorithm. Dashed lines indicate a user input step during processing, either entering and checking values or clicking.

in image space. To avoid over-segmentation by anisotropic or image-spanning grains, the x, y coordinates are rescaled to 50% of the color, which is also rescaled from 0 to 1.

We attempted using agglomerative Ward hierarchical clustering (Ward, 1963) to further improve results on anisotropic and/or large grains, however, this approach is prohibitively slow on large images, and test results did not show significant improvement. K-means clustering requires a user-supplied number of clusters. Here, we add clusters beginning at 1 and recalculate ~~the k-means clustering~~ up to an inertia improvement threshold of 1–10% ~~. The resulting k-means~~ (user supplied). Resulting labeled masks are cleaned via binary operations and the user is prompted to select the labeled regions that contain full, single grains within a simple pop-up window ~~.~~ (Fig. S3).

After selection, the orientation and a- and b-axes of an ellipse fit to the labeled region, shown to accurately approximate grain size (Graham et al., 2005a), are recorded and the grain is added to the final list and the masked region. This processing takes place over three separate scales representing a "burrowing" of the algorithm through the image (from largest to smallest window/grain size). Scales are set by the user supplied longest expected a-axis and image resolution. In contrast to the 46 variables employed by *Basegrain*, *PebbleCounts* has 20 command-line variable flags — of which 15 exert influence on the

results — with most requiring little to no modification (Table S1). Examples of the command-line interface and ~~manual~~ clicking steps are ~~shown in Figure ?? and Figure ??, respectively~~in the manual and supplement Section S2.

~~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.~~

~~Example of command-line and pop-up interface for *PebbleCounts*. (a) Interactive Otsu thresholding using percentage of Otsu value and yes ('y') or no ('n') confirmation. (b) Interactive color masking by yes ('y') or no ('n') and resulting color mask after selection. (c) K-means clustering and pop-up window for pebble selection by left clicking, with black arrows measured in final output and red arrows ignored after right-click removal (see Fig. ??).~~

~~Clicking tutorial continued from Figure ??c. Following k-means clustering at each scale a mask overlaid on the original image is presented (a), and grains are selected by a left click anywhere in the segmented area, resulting in a black circle at the click location. When clicking is finished the mask is closed by pressing 'q'. To view the original unmasked image the user may press 'r' (b). Using this switching the user can see which grains are poorly delineated and remove the last click with a right click on the mouse (c). The original black circle selection turns to red to signify this grain is off and will not be measured in the final output (d).~~

2.2 *PebbleCountsAuto*: Automatic with Image Filtering (AIF)

~~The general outline of *PebbleCountsAuto* is shown in Figure 3.~~ This method applies the same initial non-local means denoising and interactive shadow/sand masking, with the option to input user supplied values for full automation. From here, we diverge from the windowing and k-means approach and move directly to edge detection on the entire image using the same top-hat, Canny, and Sobel combination with feature-AND selections.

The resulting mask is then cleaned via binary morphological operations ~~(e.g., erosion and dilation) and each disconnected label in the resulting mask is measured as a grain~~ and each label is measured via ellipse fitting. To reduce the misidentified grains, the ellipses are filtered in a three-step chain: (A) Does the centroid fall within another ellipse?; (B) Does the ellipse overlap with any neighboring ellipses above some threshold?; and (C) Is the percent misfit (ellipse area vs. grain-mask area) above some threshold? At each step, an answer of yes leads to the elimination of the grain. The (A) and (B) steps filter grains that have high overlap or are over-segmented, whereas (C) helps filter areas where multiple grains were combined in one mask or a non-grain was identified (e.g., remaining sand patch). ~~Only the remaining, unfiltered grains~~ Grains passing the test are taken as the final results, with the assumption ~~of higher uncertainties, but that the remaining~~ that misidentified grains are minimal~~compared to the good grains~~, particularly when up-scaling to large areas and tens-of-thousands of pebbles on high-quality (low-blur) images. The command-line variables for this method are shown in Table S2, and ~~the first steps are identical to Figure ??a,b~~command-line examples can be found in the manual.

We experimented with resampling (over- and under-sampling) the image prior to grain detection to increase smoothing and to improve the detection of larger grains at the cost of measuring fewer smaller grains. The majority of images achieved the best results using the original resolution, though we did find a slight improvement in results using under-sampling on some

unsharp images (see Section S3 in the supplement). The selection of other parameters like the maximum percent misfit is also covered in Section S3 in the supplement.

3 Calibration and Validation ~~Test~~ I: Controlled Experiment

3.1 Experimental Setup

To test the KMS and AIF approaches on a simple control we arranged three distributions of well-rounded, river pebbles with a-axis sizes from 3–130 mm in semi-overlapping patterns in a 0.5×0.5 m area (Fig. 4). As opposed to most studies that use b-axis lengths to measure the ~~GSD~~ grain-size distribution (Bunte and Abt, 2001), in the experimental setup we use a-axes since it was easier to hand-measure the longest axis of ~~each of~~ the > 200 grains ~~measured. Six size class~~ used. Six size-class bins (3–5, 10–20, 25–35, 40–50, 60–70, and 80–130 mm; all a-axis) were sampled to approximate two log-normal and one bimodal ~~GSD. These classes ensured the clear demarcation of sizes into the appropriate binned values, irrespective of small uncertainties in measurement.~~ grain-size distribution. The river pebbles ~~were selected to have~~ used had uniform intra-granular color with minimal striations (i.e., veins), low angularity, and a diverse array of inter-granular colors. Lighting was controlled by overhead fluorescent bulbs and the photos were taken without flash to limit cast shadows. ~~The choice of background was a textured carpet surface to provide enough match points around the pebbles in SfM-MVS processing.~~

3.2 ~~Camera Setup~~Orthomosaic Generation

We tested a Fujifilm X100F model camera with a fixed 23 mm focal length lens and a Sony α6000 model with a removable 35 mm fixed length lens. Both had the same advanced photo system type-C (APS-C) sensors (23.6 mm×15.6 mm) and both output photos at 24 ~~MP~~ megapixels in a 4000×6000-pixel format. Following initial tests, it became clear that the image quality and grain-size results were practically identical for these two cameras, so the results presented are only those for the Fujifilm, as the photo quality was slightly sharper throughout and less distorted at the image corners. To simulate reduced quality, the 24 ~~MP~~ megapixel Fujifilm picture dimensions were reduced to 75, 50, and 25%, resulting in 13.5, 6, and 1.5 ~~MP~~ megapixel images at pixel dimensions of 3000×4500, 2000×3000, and 1000×1500, respectively.

3.3 ~~Images~~

3.2.1 ~~Top-down Images~~

~~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.~~ For each test setup, we collected ∼10 images from ∼20°off-nadir (oblique) and at least 4 overhead near-nadir (tilts < 10°) pictures, for 12–16 photos in total. The collection of oblique images aided in removing doming effects from the resulting point clouds (e.g., James and Robson, 2014) and for capturing the pebble edges and sides (Fig. S1). 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:

$$R = \frac{(S \cdot h)}{(f \cdot I)} \quad (1)$$

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~~ near-nadir 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).

3.2.1 ~~Orthomosaic Images: SfM-MVS Processing~~

~~To ensure uniform resolution, we used multiple overlapping photos taken from different angles (up to 16 photos per setup, including at least 4 overhead shots) to~~ We used the 12–16 photos to generate SfM-MVS orthoimages in *Agisoft Photoscan v.1.4.2* (Agisoft, 2018) — renamed *Agisoft Metashape* in recent versions. This allows rapid output of additional information including point clouds, digital elevation models (DEMs), and the undistorted orthomosaics, with resolution recorded in the image metadata for direct input into *PebbleCounts* and *PebbleCountsAuto*. Detailed *Agisoft* processing ~~was carried out in the following steps :~~ steps are provided in the supplement Section S4.

1. ~~Image quality detection and the exclusion of photos with quality metric < 0.7. This step analyzes pixel contrast to estimate sharpness with values ranging from 0 (blurred) to 1 (sharp). We found 0.7 to be a sufficient lower cutoff upon visual inspection of results.~~
2. ~~Detection of 12-bit coded targets in the remaining photos, with two targets placed at each of the four corners of the area and ensuring that the diameter of the printed targets' center circle was limited to 10–30 pixels in image resolution for successful automated detection.~~
3. ~~Input of scale for the orthomosaic output, provided by the distances between the targets at each corner (resulting in four distance measurements) with 0.5 mm accuracy using a ruler with cm and mm demarcations.~~
4. ~~Photo alignment at high quality with a 40,000 key-point and 2000 tie-point limit.~~
5. ~~Dense cloud generation from the aligned photos at the medium output and with moderate depth filtering. Given the high quality of the photos more aggressive options did not improve results.~~
6. ~~DEM building from the dense cloud with default settings in a local coordinate system.~~
7. ~~Generation of an orthomosaic from the input imagery and DEM at the default settings.~~
8. ~~Output of the orthomosaic to a GeoTiff file with resolution provided in m/pixel.~~

3.3 Comparison Metrics

For the simple, controlled experiment, with relatively coarse grain-size bins, it is not appropriate to compare percentiles (e.g., D_{50}) or to run Kolmogrov-Smirnov (KS) tests and measure the difference in distributions between the AIF or KMS and control ~~GSDs~~grain-size distributions. Instead, we compared the counts in each bin between the control and algorithm and visually assessed the matching of the ~~GSDs~~grain-size distributions. This provides a reasonable baseline for checking the performance of the algorithm in a highly controlled setting.

3.4 Results I: Controlled Experiment~~Results~~

For each of the three 150–200 clast arrangements, the KMS *PebbleCounts* run time was ∼7 minutes on a laptop with 16 GB RAM and 2 cores (Intel i7-6650U 2.20 GHz) and no GPU, whereas the AIF *PebbleCountsAuto* run time was ∼1 minute. Both ~~the top-down and orthoimagery was~~ a single near-nadir image and the combined orthomosaic were used, but the results were entirely consistent aside from some inter-run variability in the KMS approach caused by the non-unique solution of k-means clustering. Given this consistency, we only present the results from the ~~top-down~~ single near-nadir images. Furthermore, the use of only 4 ~~top-down~~ overlapping near-nadir photos also generated the same results, albeit in about $1/6^{th}$ the *Agisoft* orthomosaic processing time of using all 12–16 photos (∼10 minutes versus ∼1 hour on the same laptop).

Across all three distributions, the KMS approach consistently undercounts the number of clasts in each a-axis bin (Fig. 4). However, and in agreement with previous research (Graham et al., 2010), this undercounting is uniformly distributed and thus the ~~GSDs~~ grain-size distributions do not show notable differences between the algorithm and control. For the two arrangements with increased fine (3–5 mm) and coarse (60–130 mm) pebbles (Fig. 4b,c), the undercounting is stronger at the finer end of the distribution leading to a slight underestimation of the ~~GSD~~ grain-size distribution by the KMS approach in this region. This is caused partially by the user missing more of the smaller grains (of which there are exponentially more), some smaller grains being partially hidden by the larger, and also by the smallest grains being only a few pixels in area and thus eliminated during mask-cleaning steps, or not captured at all. On the other hand, the AIF approach tends to overcount the fine pebbles, leading to overestimation of the ~~GSD~~grain-size distribution, because many small non-grain areas remaining in the masked image are automatically selected in the final result, rather than ignored as in the KMS approach.

As we reduced the resolution from 0.26–1.05 mm/pixel, the reduction in the finest size class increased dramatically for the KMS approach (Fig. 5). At the lowest resolution tested (1.5 ~~MP~~megapixel), this undercounting leads to severe discrepancies in the ~~GSD~~ grain-size distribution curve. As the resolution degrades it becomes more difficult to discern rocks in the smallest size class (3–5 mm), which correspond to an a-axis grain size of 12–19, 9–14, 6–9, and 3–5 pixels for the 24, 13.5, 6, and 1.5 ~~MP~~ megapixel resolution, respectively, indicating the necessity of a limiting lower measurement factor (e.g., Graham et al., 2005a).

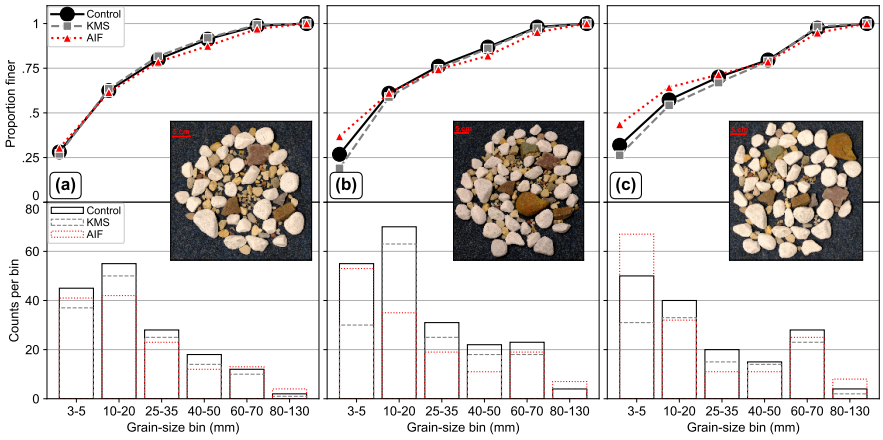

Figure 4. Result of KMS (gray, dashed lines) and AIF (red, dotted lines) on the three experimental lab setups (a-c) with known grain inputs in six size classes (black line), measured as the grain a-axis. (a) Log-normal, (b) log-normal with increased number of all classes, including fines, and (c) skewed bimodal with increased number of coarser grains. Bottom row shows the counts per bin and the top row shows the resulting ~~GSD~~grain-size distribution. The images are 0.26 mm/pixel (24 ~~MP~~megapixel).

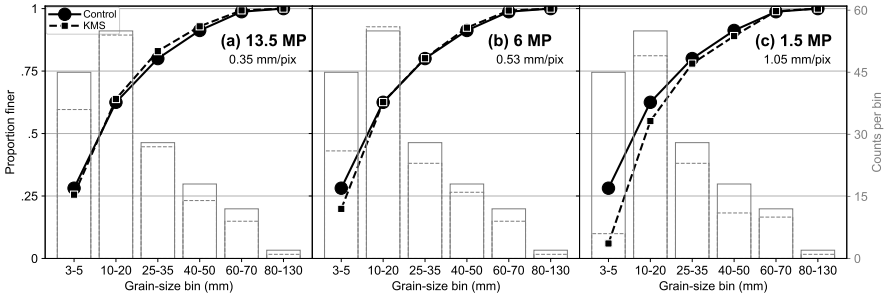

Figure 5. Results of reducing the image dimensions to (a) 75% (13.5 ~~MP~~megapixel), (b) 50% (6 ~~MP~~megapixel), and (c) 25% (1.5 ~~MP~~megapixel) and re-running the KMS approach on the distribution in Figure 4a. Control is shown as black (left y-axis) and gray (right y-axis) solid lines and KMS as the dashed lines.

4 Calibration and Validation ~~Test~~ II: Field Surveys

4.1 Field Setting

Having established the algorithms on control data, we sought to evaluate the performance on complex, natural photos. Field data provides the real-world application and detailed uncertainty analysis most useful for researchers seeking to apply the methods to their own sites. For this we turned to photo surveys carried out on gravel-bed river cross sections of the foreland and topographic transition zone of the northwestern Argentine Andes (Fig. 6). This is an area of strong precipitation, topographic, and environmental gradients, and the dynamic rivers surveyed are ~~dynamic environments~~ capable of transporting enormous quantities of sand, gravel, and boulders of various lithology (Bookhagen and Strecker, 2012; Purinton and Bookhagen, 2018). Catchment-average erosion rates from the area, based on cosmogenic nuclide inventories, suggest rates on the order of 0.6–1 mm/yr (Bookhagen and Strecker, 2012), with large variability during the Pleistocene and Holocene (Tofelde et al., 2017). The region is frequently affected by extreme hydrometeorologic events that lead to flooding and drainage-pattern ~~re-arrangement~~ rearrangement (Castino et al., 2016, 2017).

4.2 ~~Surveying and~~ Orthomosaic Generation

All cross-section surveys were collected using the Sony α6000 camera model at 24 ~~MP~~megapixel resolution, and survey sizes ranged from $\sim$1000–5000 m^2. In this case, the standard zoom lens delivered with the camera was used at the shortest focal length of 16 mm to maximize the field of view. Also, to help cover the large survey sites, the camera was affixed to the end of a pole with a remote control trigger, allowing overhead shots to be collected from a height of 4.5–5 m (Fig. 7), giving a ground resolution of approximately 1.1–1.2 mm/pixel by eq. (~~3~~1). UAV flights have proven difficult in the windy conditions experienced in these valleys, but flights at 20–30 m heights with the 12 ~~MP~~ megapixel camera provided on the DJI Mavic and Phantom models (focal lengths of 3.6–4.3 mm, sensor dimensions of 6.17$\times$4.55 mm, and image dimensions of 4000$\times$3000 pixels) would result in image resolutions of $\sim$7–13 mm/pixel, and are thus inadequate for delineating cm-scale pebbles.

To generate georeferenced orthomosaics that could be tiled and passed directly to *PebbleCounts* and *PebbleCountsAuto*, survey sites on the dry river-bed were laid out with on average 18 coded targets (with a range of 10–24) and the position of each was measured with a differential GPS (Fig. 7). Kinematic post-processing with a permanent base station < 100 km away at the Universidad Nacional de Salta (UNSA) in Salta, Argentina, led to cm accuracy of XYZ target locations. The site was traversed in a cross-hatched pattern with a photo captured every 2–3 paces, so that each location appeared in $\sim$9 ~~top-down pictures from~~ near-nadir pictures from slightly different angles. ~~*Agisoft* processing is similar to that described for the experiment (see Section 5.3.2.) , with some key differences. Here,~~ We refer to the images as near-nadir, rather than nadir, due to the fact that during mast photo collection some unintentional tilting of the camera ($< 10°$) occurred. These near-nadir photos aided in removing doming effects, but did not allow us to capture the sides of pebbles as in the oblique images taken in the ~~scale was provided by the XYZ coded target locations in UTM zone 19S, WGS84 ellipsoidal datum. Given the increased complexity of the setting and imperfect photo collection, the dense point cloud was generated at high quality with aggressive depth filtering. The DEMs and orthomosaics were also output in UTM zone 19S projections, providing undistorted pixels with~~

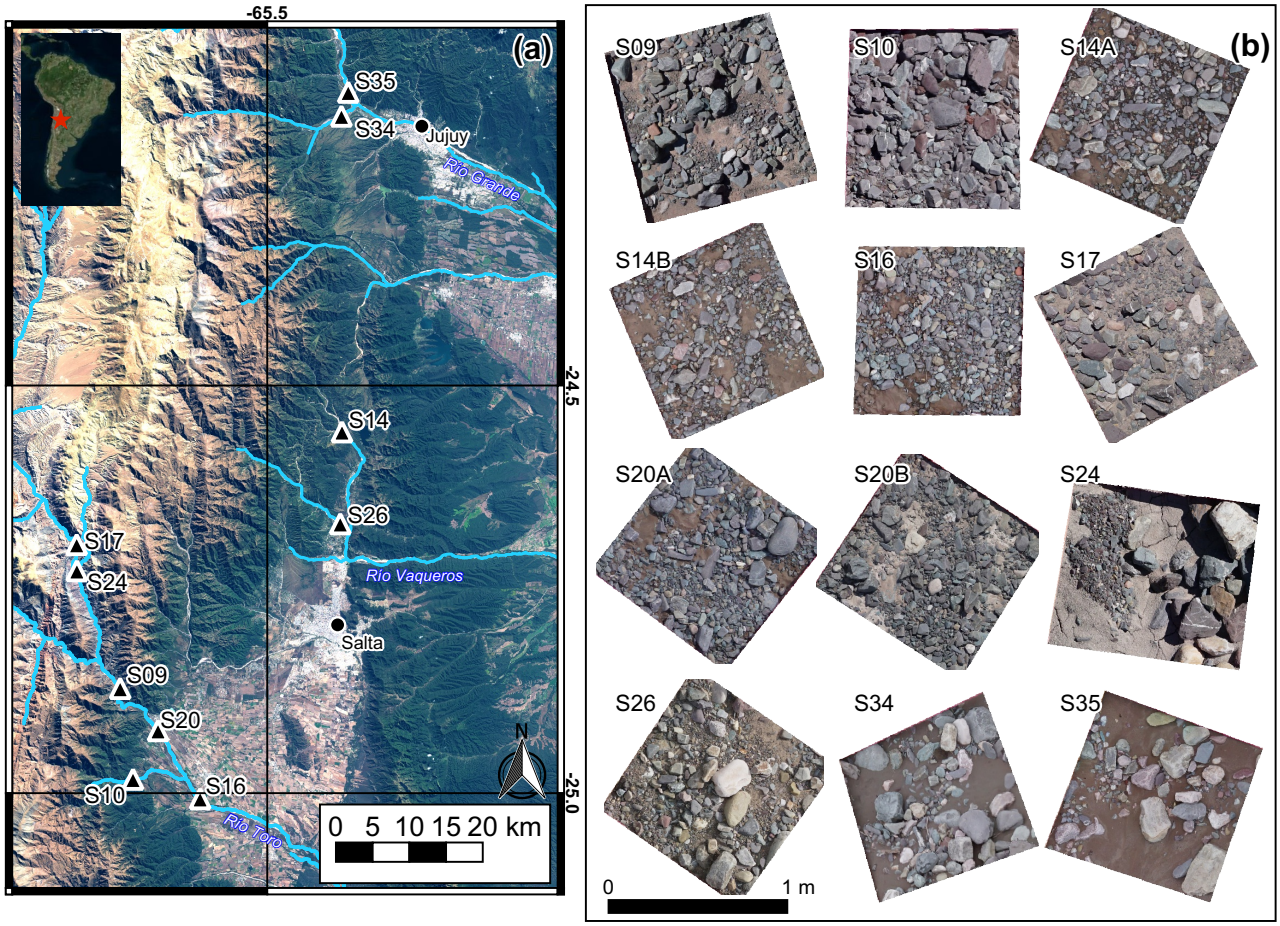

Figure 6. (a) Field cross-section survey sites (black triangles) in NW Argentina from three gravel-bed rivers (Toro, Vaqueros, and Grande) and their tributaries, draining from the sparsely vegetated mountains in the west towards the verdant foreland and city centers of Salta and Jujuy in the east. The Landsat 8 RGB composite satellite image (using bands 2, 3, and 4) from 12 June 2017 shows the climatic transition from wet foreland to dry mountains, demarcated by the green-brown transition zone, running approximately north-south, corresponding to vegetation changes~~running approximately north-south~~. (b) Detailed view of the 12 $\times$ $\sim$1 m^2 orthomosaic clips from each of the field sites with average resolution of 1.16 mm/pixel.

Figure 7. Sony α6000 24 ~~MP~~ megapixel camera affixed to mast for photo collection at a height of 4.5–5 m (left) and differential GPS measurement of coded targets (right).

~~resolution in m/pixel~~experimental setup (Fig. S1). Capturing oblique images of every patch in the field sites would require infeasible amounts of time and processing power.

Agisoft processing was similar to that described for the experiment, with some key differences (see supplement Section S4). Given the volume of photos (600–1300 per site), the sites were processed automatically using the Python API for *Agisoft*, with processing times consistently over 10 hours on an 80 core, 500 GB RAM server making use of 1 GPU NVIDIA Tesla K80 unit for some of the steps (e.g., dense matching).

From 10 of our full survey sites over three different river systems we selected 12 $\times$ $\sim$1 m^2 patches to clip out of the full orthomosaics and evaluate using the KMS and AIF approaches. The final resolution of these 12 GeoTiff orthoimages matched the theoretical value from eq. (~~3~~1), with an average of 1.16 mm/pixel and range of 1.08–1.24 mm/pixel (standard deviation of 0.05 mm/pixel). The patches (Fig. 6b) include variable amounts of sand and a large range of grain sizes, packing arrangements, and shadowing. From one site (S14A) there were ~~hand-held~~ handheld images available for the same selected patch from the same Sony α6000 camera zoomed to 20 mm focal length and taken from a height of $\sim$1.5 m, allowing for the generation of a complementary orthomosaic at 0.32 mm/pixel resolution.

4.3 ~~Control Data and~~ Comparison Metrics

For control data from the field we return to b-axis measurements (rather than a-axes as in the lab). In each patch, the b-axes of all grains visible to the naked eye were manually digitized. This generated a 5490 pebble control dataset across all 12 mast-surveyed sites. For the lone ~~hand-held~~ handheld patch at 0.32 mm/pixel, the control data was 1726 pebbles versus 621 from the same patch at the 1.12 mm/pixel mast resolution, as smaller grains could be manually measured on the image at a 4-times improved resolution.

The use of continuous control data, as opposed to discrete bins in the lab experiment, allows a more detailed investigation of the performance of both approaches, including biases and their correction. B-axis measurements of overlapping control and KMS grains were compared to look for sizing bias. This was followed by a search for the lower truncation limit (the lower cutoff in b-axis length in pixels that grains are reliably measured at) of the algorithm, also using the KMS results. For parts of the analysis, the size data were converted to the typical ψ scale ($\psi = -\phi = log2(mm)$) of grain-size measurement of coarse river sediments. This allows direct comparison of statistical results with other studies (e.g., Graham et al., 2005b)

We compared the ~~GSDs~~ grain-size distributions from the KMS and AIF approaches with the control using a two sample KS-test to check the null hypothesis that the two samples are drawn from the same distribution. Because sample sizes were at times small, leading to erroneous KS-test results, we also devised a second metric of ~~GSD~~ grain-size distribution comparison. Similar to the KS-test, which uses the maximum distance between the cumulative distribution functions (~~CDFs), or in our case the GSDs, our~~ grain-size distributions), our metric interpolates both distributions to the same lengths in 0.1 ψ steps and then sums the difference between the re-interpolated curve to give an approximate integral of the difference between the two ~~GSDs~~ grain-size distributions (AIF or KMS minus the control), which we term A_{diff}. Here, an A_{diff} value close to 0 indicates good matching, and positive or negative values indicate underestimation or overestimation, respectively.

We also examined the performance of some key percentiles ($D_{5,16,25,50,75,84,95}$). The metrics for comparison of control (P_C) and KMS or AIF (P_P) percentiles are consistent with other studies (Sime and Ferguson, 2003; Graham et al., 2005b, 2010). These are the mean ($m = \frac{1}{n} \cdot \Sigma(P_P - P_C)$), the mean squared ($ms = \frac{1}{n} \cdot \Sigma(P_P - P_C)^2$), and the irreducible random error ($e = \sqrt{ms - m^2}$). The bias of *PebbleCounts* is quantified by m, and e measures the scatter or precision after bias correction (Sime and Ferguson, 2003).

4.4 ~~Field Survey~~ Results II: Field Surveys

4.4.1 ~~Initial Results: Biases and Their Correction~~

The KMS *PebbleCounts* approach took ~10 minutes per 1 m^2 orthomosaic clip at 1.16 mm/pixel resolution, depending on the number of grains, and particularly the number of finer grains, present. Run time for the AIF *PebbleCountsAuto* approach was typically ~2 ~~minutes~~ minute per site. All run times refer to the same laptop with 16 GB RAM and 2 cores (Intel i7-6650U 2.20 GHz) and no GPU. For the 0.32 mm/pixel image the processing for KMS took ~45 minutes, as there were more fine grains to be identified (given the log-normal distribution) and so the clicking took exponentially longer, and the AIF took ~20 minutes given the longer time spent filtering the large number of grains. ~~These run times refer to the use of no lower truncation value~~

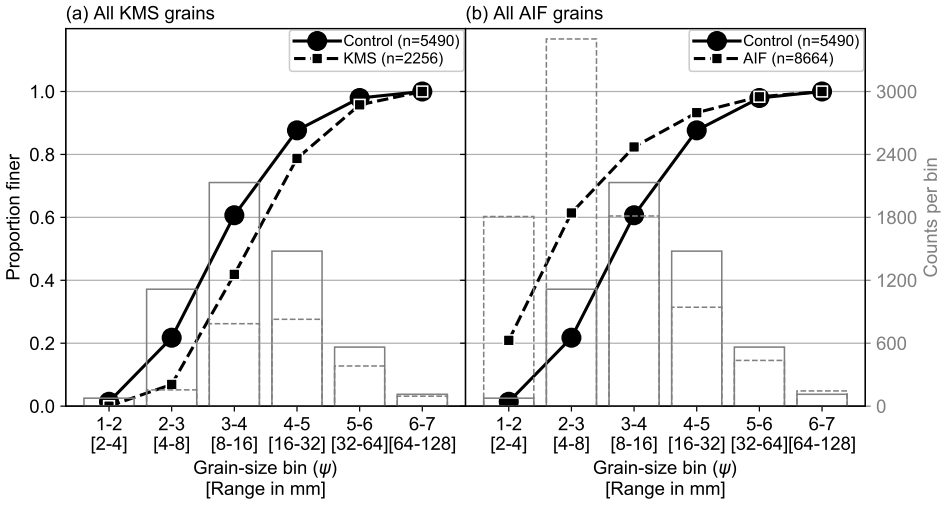

Figure 8. Comparison of (a) KMS and (b) AIF at the 12 field sites all aggregated and coarsely binned. Control is shown as black (left y-axis) and gray (right y-axis) solid lines and KMS and AIF as the dashed lines.

~~and only some morphological (e.g., erosion and dilation) cleaning operations.~~ We note that the use of a GPU for the filtering steps will significantly improve processing time. Improtatnly, these run times refer to the use of no lower truncation value, which leads to much longer processing time.

An aggregation and coarse binning of all b-axes in the control versus KMS and AIF data for the coarser imagery are presented in Figure 8. There is obvious undercounting in these data from the KMS results, similar to the experimental setup, and ~~it appears in this case to be causing~~ here it causes a significant discrepancy in the ~~GSD~~ grain-size distribution curves. Whereas the manual clicking found over 1000 grains in the smallest classes (1–2 and 2–3 ψ), the KMS approach found none in the smallest and only $\sim$100 in the second smallest. This skews the percentiles to the higher grain sizes, and thus overestimates them significantly. In opposition to this, but again in agreement with the experimental setup, the AIF results display significant overcounting at the finer sizes as many non-grains are identified, particularly when the algorithm is run with no lower truncation.

The skewed results from both the KMS and AIF approaches warrant detailed analysis of the algorithms' deficiencies and ~~GSD~~ grain-size distribution corrections. To begin, we examined the performance of *PebbleCounts* on grains manually digitized and the same grains selected during clicking in the KMS approach on the coarser imagery (Fig. 9). There is only a slight negative bias across all grain sizes, indicating underestimation of individual grains by *PebbleCounts*, however, this median shift varies with no apparent pattern and is likely caused by uncertainties in the manual b-axis digitization of thousands of grains. For instance, digitization with b-axis vector lines can achieve sub-pixel accuracy compared to the raster processing of *PebbleCounts*. The AIF approach measures grains identically to the KMS method and thus has the same misfit errors on correctly identified grains. From this we conclude that the algorithm is effective on a grain-by-grain basis and the skewing of

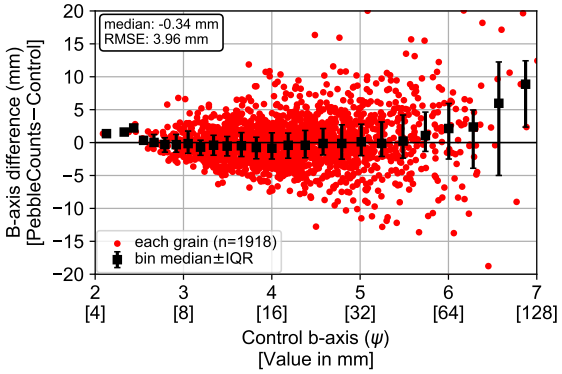

Figure 9. Measurement error of *PebbleCounts* (here the KMS results) versus control on a grain-by-grain basis for overlapping grains in the coarser (1.16 mm/pixel) imagery. There is an overall median shift, but the binned medians do not display a consistent pattern.

the ~~GSDs~~ grain-size distributions are instead caused by sampling errors related to the image resolution and ability to find small grains (see Figure 5).

The undercounting error can be explored on the full distribution of pebbles by gradually increasing the lower truncation value and assessing the error in percentiles versus the control data at each step (Fig. 10). As truncation is increased, the median percentile error decreases rapidly up to an inflecting value — manually chosen from the graph as a significant local minimum — where the median difference is near 0 mm. Truncating the KMS distributions at a minimum b-axis length of 23 mm (rounded to 20 pixels) improves the results significantly for the 1.16 mm/pixel imagery taken from the mast. Beyond this truncation, there is limited improvement. Regarding the 0.32 mm/pixel image, the 20-pixel (6.5 mm) truncation also results in a median difference near 0 mm, with subsequent truncation values leading to only ∼0.5 mm improvements. Supplying these truncation values directly to the KMS *PebbleCounts* tool results in reduced processing time to ∼5 minutes for the coarser imagery and ∼15 minutes for the finer, as many small grains were then ignored and left out of the clicking mask.

The same analysis for the AIF approach is complicated by the large number of false grains found and the extreme over-counting of fine grains. Given this, we instead make the assumption that the similarity of the two methods, particularly in the edge detection and ellipse fitting steps, leads to similar errors in both. Therefore, we assume the same 20-pixel truncation. For the AIF *PebbleCountsAuto* tool, processing times with the 20-pixel truncation reduced to < 1 minute and ∼3 minutes for the coarse and fine images, respectively.

4.4.1 ~~Results: Mast Images~~

The combined results before and after lower truncation for the coarser (∼1.16 mm/pixel) imagery taken from the mast surveys is shown in Figure 11. ~~For separate plots of the 12 different sites before and after truncation in the KMS approach see Section S2 in the supplement.~~ Without any lower truncation, the AIF tool results in significant overcounting and ~~GSD~~ grain-size distribution underestimation with a high $A_{diff} > 8$. The KMS tool instead shows undercounting and ~~GSD~~ grain-size distribution

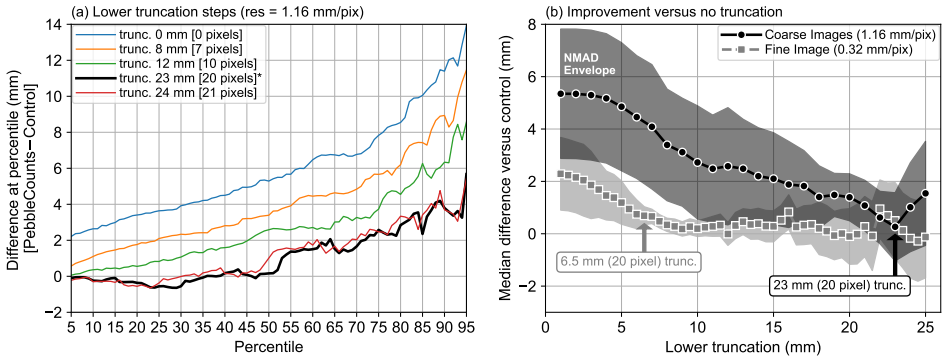

Figure 10. (a) Error in each percentile (5–95) as lower truncation value is increased in 1 mm steps for the 1.16 mm/pixel imagery. Only a few steps are plotted for clarity. (b) The median difference in percentiles compared with the control versus the lower truncation value, with the normalized median absolute difference (NMAD) shown as the error envelope (Höhle and Höhle, 2009). From this analysis, we select a lower truncation of 20 pixels. The analysis in (a) was repeated for the finer image (with 0.5 mm truncation steps) to get the gray squares line in (b), and is not shown here.

overestimation with a low $A_{diff} < -4$. Both have KS-test p-values < 0.0001. When we apply a 20-pixel truncation, both the AIF and KMS approaches achieve A_{diff} values near or below -1, with the manual KMS approach performing best and achieving a high KS-test p-value of 0.2398. The AIF approach retains a low p of 0.0008 with a ~0.1–0.2 ψ bias towards coarser values in the upper portion of the ~~GSD~~ grain-size distribution (> D_{50}).

In ~~Figure ??~~the supplement Section S5 (Fig. S7), we show the 20-pixel truncated KMS and AIF results on a site-by-site basis. For the KMS approach, following truncation 11 sites have p-values > 0.1 and one site (S16) has p=0.0971. A_{diff} values are also near 0 indicating close matching of the ~~GSDs~~grain-size distributions, aside from S24 and S34, which both show large discrepancies. The AIF results in Figure ~~??~~ S7 follow a similar trend to the KMS results~~. The main difference is that, for the AIF approach,~~ , but there is a bias towards coarser values, with many A_{diff} values < -1, and generally poorer results compared with the KMS approach, with ~~GSDs~~ grain-size distributions being overestimated by ~0.1–0.2 ψ.

In the KMS results, despite a high p-value, S24 demonstrates a stronger bias in the ~~GSD~~ grain-size distribution towards coarser grains (up to 0.5 ψ discrepancy), as indicated by the high A_{diff} value of -1.36. Here, the KS-test pass is likely caused by the small sample size remaining after truncation (n=24), the least of any site. The poor performance of S24 was expected given the large size range with many sub-cm pebbles and a few large boulders, strong cast shadows from the large grains, and intra-granular edges on angular boulders with quartz veins (see Figure 6b). 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 A_{diff}=−2.11. In this case, poor performance is due to significant blurriness of this image, and again a small sample size (n=47).

We also compared the individual percentiles of interest to assess the bias and accuracy of truncated results (Fig. 12). For the KMS approach, the bias (m) is 0.06 ψ with a precision (e) of 0.13 ψ. Excluding S24 and S34, m and e drop to 0.03 and 0.09 ψ, respectively. The AIF results have higher m and e values of 0.15 and 0.17 ψ, respectively, which are reduced to 0.13 and

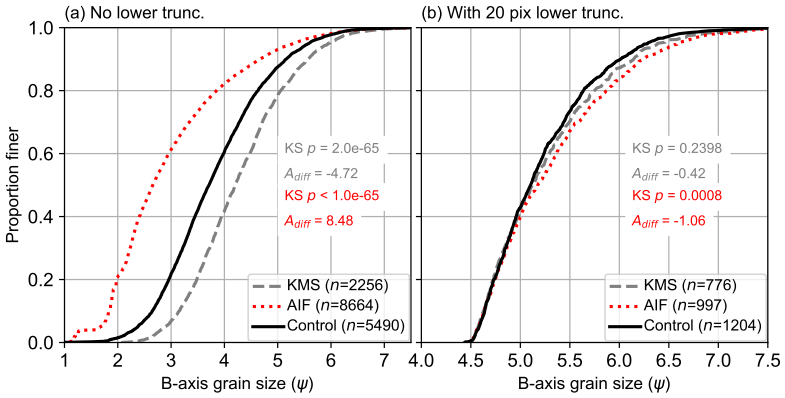

Figure 11. Results from hand-clicked control (black line), KMS *PebbleCounts* (gray, dashed line), and AIF *PebbleCountsAuto* (red, dotted line) with the initial non-truncated run (a) and the 20-pixel truncated run (b). In corresponding colors are the p-value results of a KS-test and the A_{diff} approximate integral between the curves for each approach versus the control data. The legend indicates the number of grains (n) making up each curve. Note the reduction in x-axis scale between the columns, where the right, truncated distributions are plotted on a narrower range to emphasize the remaining discrepancies. The curves separated by site (Fig. 6b) are shown in the supplement Section S5, Figure S7

0.15 ψ following exclusion of the same S24 and S34 sites, in addition to the S10 site, which was also somewhat blurry and with relatively few grains. For the AIF percentiles, we chose to include S16 despite large overestimation at higher percentiles (Fig. ??S7), as this was a sharp image with a relatively large sample size. The high uncertainties from this scene likely require some adjustment of the edge-detection variables (see Section S3 in the supplement) for improved segmentation, but the results presented are realistic for fast processing using the AIF method, with the caveat of higher expected uncertainties.

The uncertainties in Figure 12 are average values, and the inset plots also demonstrate the increasing uncertainty of larger percentiles. The maximum uncertainty for both at D_{95} is m=0.08 ψ and e=0.07 ψ for the KMS result and m=0.35 ψ and e=0.2 ψ for the AIF result. Importantly, since the ψ scale is logarithmic, the larger errors at higher percentiles correspond to similar percentage misfits as lower errors at smaller percentiles (e.g., 0.2 ψ precision at a grain size of 6.5 ψ (91 mm) is a 13–15% misfit, whereas, a 0.01 ψ precision at 4.5 ψ (23 mm) is a 4–10% misfit).

4.4.1 ~~Results: Handheld Image~~

As a final test for the KMS and AIF approaches, we turn towards our handheld imagery taken from S14A with a 4-times improved resolution of 0.32 mm/pixel (Fig. 13). We only show the 20-pixel truncated results, which displayed high KS-test p-values > 0.2 and A_{diff} close to 0 in both cases, with the AIF approach slightly underestimating (A_{diff}=0.6) and KMS slightly overestimating (A_{diff}=−0.77). For the KMS approach m and e are 0.07 and 0.05 ψ, respectively, and −0.06 and 0.05 ψ for AIF.

~~Comparison of 20-pixel truncated GSDs between hand-clicked control (black line), KMS *PebbleCounts* (gray, dashed line), and AIF *PebbleCountsAuto* (red, dotted line) for the 12 × ~1.16 mm/pixel control sites. In corresponding colors are the p-value results of a KS-test and the A_{diff} approximate integral between the curves for each approach versus the control data. The legend indicates the number of grains (n) making up each curve.~~

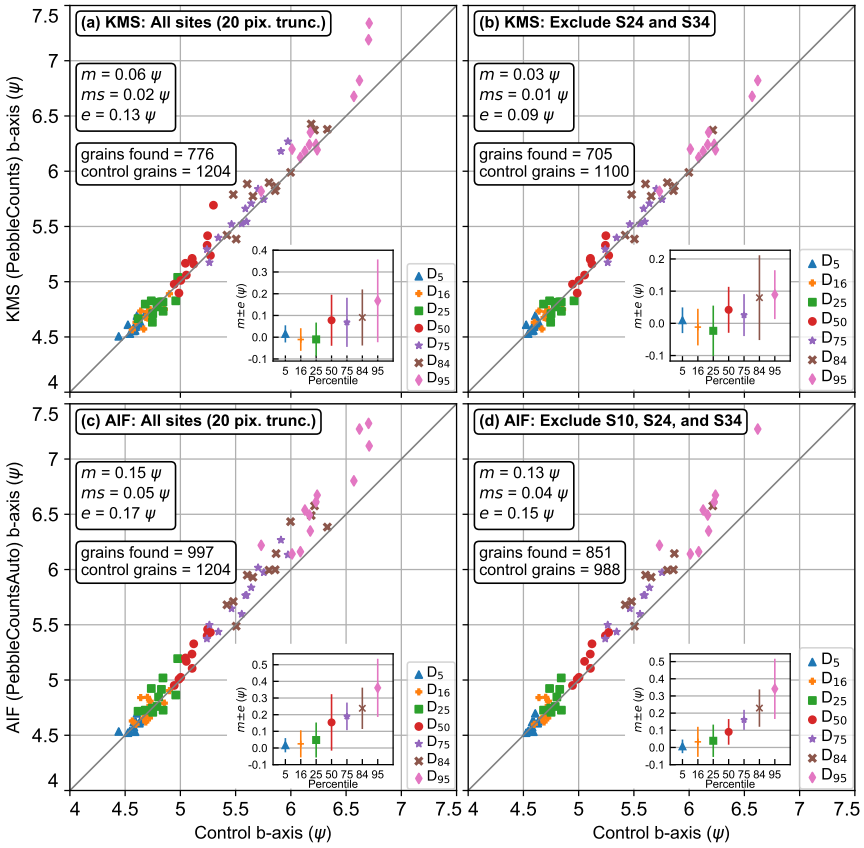

Figure 12. Comparing the key b-axis percentiles across all 12 field sites and between the KMS and AIF approaches with the 20-pixel truncation applied. (a) All 12 sites from KMS, (b) KMS improvement when excluding S24 and S34, (c) all 12 sites from AIF, and (d) AIF improvement when excluding S10, S24, and S34. For the main plot, each data point is a percentile value from a single site and the 1:1 relationship is the gray diagonal. The mean (m), mean squared (ms), and irreducible (e) errors are shown for each plot, taken as the average of all 7 percentile errors across the 9–12 sites plotted. The m and e are separately plotted for each percentile in the inset plot. The number of grains in the control ("control grains") and KMS or AIF results ("grains found") are also indicated. The individual site curves where these data points originate are shown in the supplement Section S5, Figure S7.

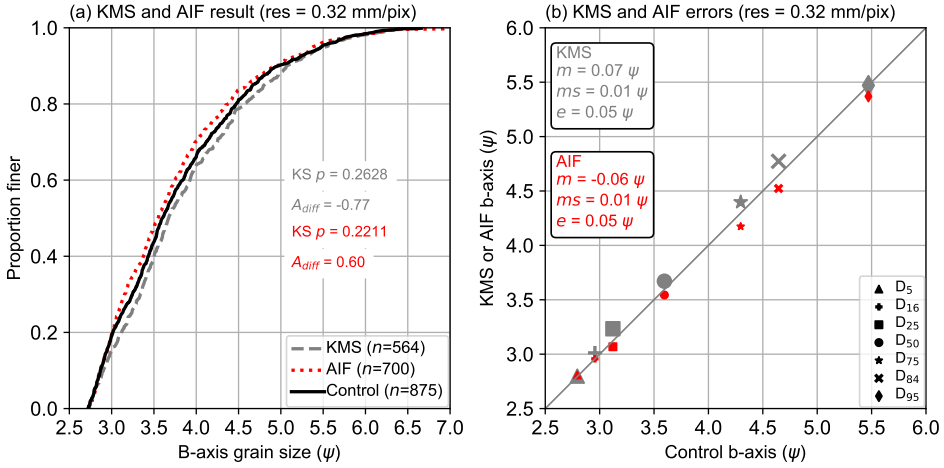

Figure 13. (a) Results from hand-clicked control (black line), KMS *PebbleCounts* (gray, dashed line), and AIF *PebbleCountsAuto* (red, dotted line) from the 20-pixel truncated run on the 0.32 mm/pixel handheld imagery. In corresponding colors are the p-value results of a KS-test and the A_{diff} approximate integral between the curves for each approach versus the control data. (b) Percentile comparison for both methods with KMS in gray and AIF in red, with inset box showing the uncertainties for each in the corresponding color.

4.5 Caveat of *PebbleCountsAuto* AIF

The promising results of the AIF approach shown in ~~Figure~~ Figures 11–13 come with some consideration of the grain-by-grain accuracy. In Figure 14, we analyze the percentage of grains found in the AIF approach that have a corresponding grain in either the hand-clicked control (based on a 6-mm buffer of the b-axis line) or the KMS results (based on a 6-mm centroid buffer). From this subset of grains, we consider the AIF grain to be a matching (or correct) result if the b-axis difference between it and the nearby "good" grain (from the control or KMS) is < 1 cm. From this we see that in the best-case scenario the percentage of correct grains identified by the AIF approach is only 70%, from the handheld 0.32 mm/pixel image. A number of sites (S10, S16, S20B, S24, S34, and S35) have < 50% matched grains. The two poorly performing sites (S24 with grain complexity and S34 with image blur) both demonstrate the lowest accuracy with < 40% matches. Notably, despite a significant number of false positives in the results, when comparing the overall ~~GSDs~~ grain-size distributions (Fig. 11), and on a site-by-site basis (Fig. ~~??~~S7), the distribution of the AIF results matches the hand-clicked control well.

~~Figure ?? demonstrates the issues~~ The errors associated with the AIF ~~approach in a few map-view examples of the results of the KMS approach versus the same pebbles in the AIF approach. On a grain-by-grain basis, there are many inaccuracies falling into three main categories: over-segmentation of grains with internal edges and the selection of each segment as a separate grain, under-segmentation and merging of neighboring grains that have weak edges sometimes caused by image blur, and misidentification of non-grain objects or clusters of small grains. It is clear from this analysis that caution must be used when interpreting AIF results, particularly in complex or blurry images.~~ method are demonstrated in the supplement Section S6.

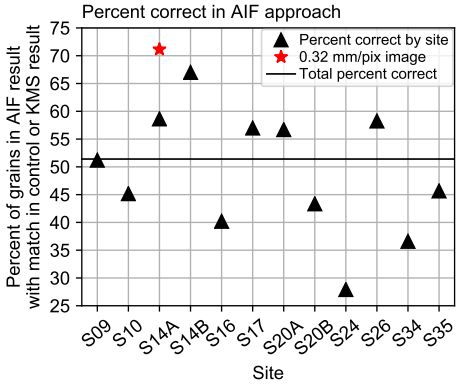

Figure 14. Percentage of grains from AIF results with a matching grain in either the hand-clicked control or in the KMS result. A match is defined as a grain within 5 pixels of the hand-clicked line or the KMS grain centroid for the 1.16 mm/pixel imagery, or within 20 pixels for the 0.32 mm/pixel image (corresponding in both cases to a distance of ∼6 mm), and with a 1 cm maximum b-axis difference between the AIF grain and the match. The total percent correct, taken across all black triangles, is 51%.

~~Resulting delineated grains using the AIF *PebbleCountsAuto* function (top row) versus the same area from the KMS *PebbleCounts* function (bottom row). Labels indicate the issues with the AIF results and improvement in KMS results. Note the poor results for the blurry image on the right (S34).~~

5 Discussion

In this study we developed two new methods for grain-size measurement with low uncertainties and the potential to deliver full ~~GSDs~~ grain-size distributions 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~~ grain-size distribution estimates, and practical guidelines for acquiring imagery and applying *PebbleCounts*, including the application of UAV surveys.

5.1 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 et al., 2018), since these ~~methods~~ are based on correlative relationships rather than physical measurement of each grain. Texture methods work well for homogeneous pebble arrangements in lower-energy settings, but high-energy mountain rivers with heterogeneous pebble arrangements and large ranges in sizes require segmentation approaches. Similar

Table 1. Comparison of *PebbleCounts* and *PebbleCountsAuto* results with other segmentation-based pebble counting studies.

Study / Technique	**Bias (ψ)**	**Precision (ψ)**
This Study / K-means with Manual Selection (KMS)	0.03–0.07	0.05–0.09
This Study / Automatic with Image Filtering (AIF)	−0.06–0.13	0.05–0.15
Butler et al. (2001) / Custom watershed segmentation	0.13–0.33[a]	-
Sime and Ferguson (2003) / Custom watershed segmentation	0.14[b]	0.22[b]
Graham et al. (2005b) / Custom watershed segmentation	0.007–0.03	0.07–0.09
Westoby et al. (2015)[c] / *Basegrain* (Detert and Weitbrecht, 2012)	0.16–0.82[d]	0.33–1.99[d]

[a] Taken from only three percentiles ($D_{50,84,95}$).

[b] Corrected value presented by Graham et al. (2005b).

[c] Comparison made in mm, converted to ψ units here.

[d] Large spread caused by significant disagreement at higher percentiles.

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 (Graham et al., 2010). This undercounting does not undermine the resulting ~~GSDs~~ grain-size distributions 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. 10), which explains the degradation in 3–5 mm counting in the reduced resolution lab images (Fig. 5)~~)~~, where the smallest pebbles were only a few pixels in size as resolution was decreased.

As shown in Figure 12, when we apply this cutoff and exclude poorly performing images we find an average m (bias) and e (~~spread~~precision) of 0.03 and 0.09 ψ, respectively, for the ∼1.16 mm/pixel imagery and 0.07 and 0.05 ψ for the 0.32 mm/pixel image. For the AIF approach these values are 0.13 and 0.15 ψ for the ∼1.16 mm/pixel imagery and −0.06 and 0.05 ψ 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 14 and ~~??~~S8, 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~~and precision values of −0.06–0.~~13~~ .15 ψ are ~~in the range~~ on the low end of previously published ~~absolute biases of 0.007–0.33 ψ~~ errors from similar techniques (~~see Table 2 in Graham et al. (2010)~~ Table 1). To our knowledge, the only study to compare *Basegrain* results to control data by Westoby et al. (2015), makes comparisons in mm rather than ψ units. Since the ψ scale is logarithmic, in our study the error in mm increases with ψ from ∼0.8 mm uncertainty at 4.5 ψ (23 mm) to ∼7 mm uncertainty at 6.5 ψ (91 mm) for the ∼1.16 mm/pixel imagery in the KMS case. Westoby et al. (2015) report ~~similar bias~~ even greater bias and lower precision from *Basegrain*, ~~again~~ with errors also increasing in magnitude at higher percentiles. ~~Regarding the error spread reported in the literature, our range of 0.05–0.13 ψ is less than the 0.25 and 0.14 ψ values reported by Sime and Ferguson (2003) and Graham et al. (2005b), respectively, for their image segmentation techniques.~~ We emphasize that the previous image segmentation techniques discussed ~~here~~ all rely on ~~the watershed segmentationstep, whereas~~ watershed segmentation, whereas neither of our algorithms use this step for the reasons demonstrated in Figures 1 and 2.

5.2 ~~Effect of~~ Lower Truncation~~on GSD~~

The issue of lower truncation on ~~GSDs~~ grain-size distributions and percentile estimates has received much attention in the literature (e.g., Fripp and Diplas, 1993; Rice and Church, 1996; Bunte and Abt, 2001; Graham et al., 2010). Previously, field geomorphologists were interested in all grains above 8–16 mm, simply because smaller grains were difficult to manually identify and thus underrepresented in the results (e.g., Fripp and Diplas, 1993; Rice and Church, 1998). Previous work suggests that truncation at the finer end of the distribution primarily increases the lower percentiles, while having less effect on the large ($> D_{50}$) percentiles (Bunte and Abt, 2001). We find significant shifts in all percentiles of > 0.5 ψ when applying a 20-pixel truncation. Graham et al. (2010) report truncation errors of < 0.3 ψ for all percentiles in 1, 3, and 5 ψ truncated distributions. Their better results at lower percentiles are likely because the data were collected manually grid-by-number style in the field with the ability to include smaller grain sizes. The measurement resolution presents the ultimate control on how accurately grain-size percentiles can be measured. The purpose of the KMS and AIF approaches introduced here is in acquiring ~~GSDs~~ grain-size distributions from a subset of the full grain-size range present in the river, namely the subset with > 20-pixel b-axis length in image resolution.

5.3 ~~Practical Considerations for~~ Image ~~Collection and Processing~~Acquisition

~~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.~~

5.3.1 ~~Acquisition Geometry and Resolution of Mast or Handheld Images~~

Ideally, collecting 9+ ~~top-down~~ near-nadir images/m^2 (as in our field surveys) or collecting an approximately 1:2 (or greater) ratio of ~~top-down~~ near-nadir to oblique imagery (as in our experiments with ~~point cloud~~ point-cloud data dimensions; see

supplement Section S1), leads to the highest quality ~~point cloud~~ 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~~ attained comparable results in reduced time using only 4 ~~top-down~~ overlapping near-nadir 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~~ point-cloud data in addition to the 2D orthomosaics, it is recommended that (many) more images closer to the surface be collected (e.g., Verma and Bourke, 2019) and from oblique viewing angles ~~(e.g., Verma and Bourke, 2019)~~(e.g., James and Robson, 2014).

As we find the difference in calculated resolution and subsequent grain-size measurement to be negligible between orthorectified and raw ~~top-down~~ near-nadir imagery at these scales, the use of ~~orthomosaic imagery~~ orthoimagery 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 α6000 (24 ~~MP~~megapixel, 15.6×23.5 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~~.

5.3.1 ~~On the Use of the UAVs~~

5.4 UAV Surveying

The > 20 m flight heights typical of UAV surveys lead to cm-scale imagery with currently available 12–24 ~~MP~~ megapixel 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,~~ Acquiring 0.5 mm/pixel imagery from a DJI Mavic drone with a 12 ~~MP camera , wide angle 4.3 mm focal length, and 4.55×6.17 mm sensor~~ megapixel camera requires a very low flight height of ∼1.4 m, giving a field of view of only ∼1.5×2 m. This may be ~~somewhat~~ improved using better cameras like on the Mavic 2 Pro (20 ~~MP camera). Regardless, acquiring~~ megapixel camera), but gathering 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,~~ . Given continual technology improvements (e.g., greater battery life, more accurate geo-tags from onboard dGPS, higher megapixel cameras, and reduced motion blur), it is within reason to expect hectare to multi-hectare SfM-MVS UAV surveys at mm resolution in seamless orthomosaics along entire river reaches in the near future. But, for the time-being, ~~the~~ a single, non-overlapping orthoimage workflow proposed by Carbonneau et al. (2018) has high potential to achieve large-areal results~~from~~ . Their workflow, building on Carbonneau and Dietrich (2017), uses a number of high and oblique overlapping flights to orthorectify a lower non-overlapping flight with mm-scale acquisition, with resulting single, scaled images passed to *Basegrain*, or, alternatively, to *PebbleCounts*~~using UAV imagery~~.

5.4.1 ~~Coverage and Processing Limits Using PebbleCounts~~

5.5 Coverage and Processing Limits

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~~sensor dimensions, focal length, and height). The potential to cover even larger ~~survey~~ sites up to or exceeding 100×100 m (~~10,000 m^2 =~~ 1 hectare) is feasible in a day of work by two people (with one measuring the targets and both sharing the photo-taking) 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 images 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 reduced motion blur. 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 (> 2,000×2,000 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 and processed into a georeferenced orthomosaic 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~~grain-size distribution. 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.

6 Conclusions

Using a k-means approach for pebble segmentation in the spectral and spatial domain combined with fast manual selection of good results, we developed a new semi-automated algorithm for grain sizing optimized for images taken over gravel-bed rivers (*PebbleCounts*). We also developed an automated algorithm that uses suspect grain filtering (*PebbleCountsAuto*), albeit with larger uncertainties in the results. The lower truncation of the methods (minimum b-axis length measurable) is limited to 20-pixels and above. These new methods were necessary to acquire grain-size distributions from dynamic high-mountain rivers with complexity from sources such as large ranges in grain size, intra-granular heterogeneity, grain overlap, irregular shadowing, and sand patches. Similar to previous methods, *PebbleCounts* is best applied at the patch scale (~~1–10~~ ∼1 m^2), however, *PebbleCounts* provides more realistic results in complex images without any post-processing steps in ∼~~5–20~~ 5–10 minutes per patch, assuming ∼1 mm/pixel resolution imagery. *PebbleCountsAuto* performs very well on high-quality (low-blur) imagery, though with remaining misidentification that must be approached with caution. Grain-sizing results can be upscaled to areas on the order of 10^2–10^4 m^2 when *PebbleCounts* results are used as ~~calibration and~~ validation for the automated *PebbleCountsAuto* function.

Code availability. *PebbleCounts* is a Python based program with the code and documentation available on GitHub at: https://github.com/UP-RS-ESP/PebbleCounts (Purinton and Bookhagen, 2019).

Author contributions. BB and BP defined the project. BP developed the algorithms with support from BB. BP carried out the analysis, produced the figures, and wrote the manuscript. BB provided funding, guidance in data analysis, and manuscript edits.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. Anna Rosner is thanked for assistance with fieldwork for mast surveys. Steffen Wellegehausen is thanked for aiding in the lab experiment setup. Funding was sourced from DFG funded IRTG-StRATEGy (IGK2018) and NEXUS funded through the MWFK Brandenburg, Germany, both for Bodo Bookhagen. We acknowledge the support of the Open Access Publishing Fund of the University of Potsdam. Constructive reviews from Patrice Carbonneau~~and Pascal Allemand~~, Pascal Allemand, and Eric Lajeunesse improved the structure of the manuscript.

References

Agisoft: AgiSoft PhotoScan Professional, http://www.agisoft.com/downloads/installer/, 2018.

Attal, M. and Lavé, J.: Changes of bedload characteristics along the Marsyandi River (central Nepal): Implications for understanding hillslope sediment supply, sediment load evolution along fluvial networks, and denudation in active orogenic belts, Geological Society of America Special Papers, 398, 143–171, https://doi.org/10.1130/2006.2398(09), 2006.

Attal, M., Mudd, S., Hurst, M., Weinman, B., Yoo, K., and Naylor, M.: Impact of change in erosion rate and landscape steepness on hillslope and fluvial sediments grain size in the Feather River basin (Sierra Nevada, California), Earth Surface Dynamics, 3, 201–222, https://doi.org/10.5194/esurf-3-201-2015, 2015.

Bertin, S. and Friedrich, H.: Field application of close-range digital photogrammetry (CRDP) for grain-scale fluvial morphology studies, Earth Surface Processes and Landforms, 41, 1358–1369, https://doi.org/10.1002/esp.3906, 2016.

Bertin, S., Groom, J., and Friedrich, H.: Isolating roughness scales of gravel-bed patches, Water Resources Research, 53, 6841–6856, https://doi.org/10.1002/2016WR020205, 2017.

Bookhagen, B. and Strecker, M. R.: Spatiotemporal trends in erosion rates across a pronounced rainfall gradient: Examples from the southern Central Andes, Earth and Planetary Science Letters, 327-328, 97–110, https://doi.org/10.1016/j.epsl.2012.02.005, 2012.

Brasington, J., Vericat, D., and Rychkov, I.: Modeling river bed morphology, roughness, and surface sedimentology using high resolution terrestrial laser scanning, Water Resources Research, 48, W11 519, https://doi.org/10.1029/2012WR012223, 2012.

Buades, A., Coll, B., and Morel, J.-M.: Non-Local Means Denoising, Image Processing On Line, 1, 208–212, https://doi.org/10.5201/ipol.2011.bcm_nlm, 2011.

Bunte, K. and Abt, S. T.: Sampling surface and subsurface particle-size distributions in wadable gravel- and cobble-bed streams for analyses in sediment transport, hydraulics and streambed monitoring, Tech. rep., US Forest Service, Rocky Mountain Research Station, Fort Collins, CO, https://doi.org/10.2737/RMRS-GTR-74, 2001.

Buscombe, D.: Transferable wavelet method for grain-size distribution from images of sediment surfaces and thin sections, and other natural granular patterns, Sedimentology, 60, 1709–1732, https://doi.org/10.1111/sed.12049, 2013.

Buscombe, D., Rubin, D. M., and Warrick, J. A.: A universal approximation of grain size from images of noncohesive sediment, Journal of Geophysical Research: Earth Surface, 115, F02 015, https://doi.org/10.1029/2009JF001477, 2010.

Butler, J. B., Lane, S. N., and Chandler, J. H.: Automated extraction of grain-size data from gravel surfaces using digital image processing, Journal of Hydraulic Research, 39, 519–529, https://doi.org/10.1080/00221686.2001.9628276, 2001.

Carbonneau, P., Bizzi, S., and Marchetti, G.: Robotic photosieving from low-cost multirotor sUAS: a proof-of-concept, Earth Surface Processes and Landforms, 43, 1160–1166, https://doi.org/10.1002/esp.4298, 2018.

Carbonneau, P. E.: The threshold effect of image resolution on image-based automated grain size mapping in fluvial environments, Earth Surface Processes and Landforms, 30, 1687–1693, https://doi.org/10.1002/esp.1288, 2005.

Carbonneau, P. E. and Dietrich, J. T.: Cost-effective non-metric photogrammetry from consumer-grade sUAS: implications for direct georeferencing of structure from motion photogrammetry, Earth Surface Processes and Landforms, 42, 473–486, https://doi.org/doi:10.1002/esp.4012, 2017.

Carbonneau, P. E., Lane, S. N., and Bergeron, N. E.: Cost-effective non-metric close-range digital photogrammetry and its application to a study of coarse gravel river beds, International Journal of Remote Sensing, 24, 2837–2854, https://doi.org/10.1080/01431160110108364, 2003.

Carbonneau, P. E., Lane, S. N., and Bergeron, N. E.: Catchment-scale mapping of surface grain size in gravel bed rivers using airborne digital imagery, Water Resources Research, 40, W07 202, https://doi.org/10.1029/2003WR002759, 2004.

Castino, F., Bookhagen, B., and Strecker, M.: River-discharge dynamics in the Southern Central Andes and the 1976–77 global climate shift, Geophysical Research Letters, 43, https://doi.org/10.1002/2016GL070868, 2016.

Castino, F., Bookhagen, B., and Strecker, M. R.: Oscillations and trends of river discharge in the southern Central Andes and linkages with climate variability, Journal of Hydrology, 555, 108–124, https://doi.org/10.1016/j.jhydrol.2017.10.001, 2017.

Chatanantavet, P., Lajeunesse, E., Parker, G., Malverti, L., and Meunier, P.: Physically based model of downstream fining in bedrock streams with lateral input, Water Resources Research, 46, W02 518, https://doi.org/10.1029/2008WR007208, 2010.

Church, M., Hassan, M. A., and Wolcott, J. F.: Stabilizing self-organized structures in gravel-bed stream channels: Field and experimental observations, Water Resources Research, 34, 3169–3179, https://doi.org/10.1029/98WR00484, 1998.

Cullen, N. D., Verma, A. K., and Bourke, M. C.: A comparison of structure from motion photogrammetry and the traversing micro-erosion meter for measuring erosion on shore platforms, Earth Surface Dynamics, 6, 1023–1039, https://doi.org/10.5194/esurf-6-1023-2018, https://www.earth-surf-dynam.net/6/1023/2018/, 2018.

de Haas, T., Ventra, D., Carbonneau, P. E., and Kleinhans, M. G.: Debris-flow dominance of alluvial fans masked by runoff reworking and weathering, Geomorphology, 217, 165 – 181, https://doi.org/10.1016/j.geomorph.2014.04.028, 2014.

Detert, M. and Weitbrecht, V.: Automatic object detection to analyze the geometry of gravel grains–a free stand-alone tool, in: River flow 2012 : Proceedings of the international conference on fluvial hydraulics, San José, Costa Rica, September 5-7, 2012, pp. 595–600, Taylor & Francis Group, London, 2012.

Dugdale, S. J., Carbonneau, P. E., and Campbell, D.: Aerial photosieving of exposed gravel bars for the rapid calibration of airborne grain size maps, Earth Surface Processes and Landforms, 35, 627–639, https://doi.org/10.1002/esp.1936, 2010.

Dunne, K. B. and Jerolmack, D. J.: Evidence of, and a proposed explanation for, bimodal transport states in alluvial rivers, Earth Surface Dynamics, 6, 583–594, https://doi.org/10.5194/esurf-6-583-2018, 2018.

Eltner, A., Kaiser, A., Castillo, C., Rock, G., Neugirg, F., and Abellán, A.: Image-based surface reconstruction in geomorphometry – merits, limits and developments, Earth Surface Dynamics, 4, 359–389, https://doi.org/10.5194/esurf-4-359-2016, 2016.

Ferguson, R., Hoey, T., Wathen, S., and Werritty, A.: Field evidence for rapid downstream fining of river gravels through selective transport, Geology, 24, 179–182, https://doi.org/10.1130/0091-7613(1996)024<0179:FEFRDF>2.3.CO;2, 1996.

Fripp, J. B. and Diplas, P.: Surface Sampling in Gravel Streams, Journal of Hydraulic Engineering, 119, 473–490, https://doi.org/10.1061/(ASCE)0733-9429(1993)119:4(473), 1993.

Gomez, B., Rosser, B. J., Peacock, D. H., Hicks, D. M., and Palmer, J. A.: Downstream fining in a rapidly aggrading gravel bed river, Water Resources Research, 37, 1813–1823, https://doi.org/10.1029/2001WR900007, 2001.

Graham, D. J., Reid, I., and Rice, S. P.: Automated Sizing of Coarse-Grained Sediments: Image-Processing Procedures, Mathematical Geology, 37, 1–28, https://doi.org/10.1007/s11004-005-8745-x, 2005a.

Graham, D. J., Rice, S. P., and Reid, I.: A transferable method for the automated grain sizing of river gravels, Water Resources Research, 41, W07 020, https://doi.org/10.1029/2004WR003868, 2005b.

Graham, D. J., Rollet, A.-J., Piégay, H., and Rice, S. P.: Maximizing the accuracy of image-based surface sediment sampling techniques, Water Resources Research, 46, W02 508, https://doi.org/10.1029/2008WR006940, 2010.

Grant, G. E.: The Geomorphic Response of Gravel-Bed Rivers to Dams: Perspectives and Prospects, chap. 15, pp. 165–181, Wiley-Blackwell, https://doi.org/10.1002/9781119952497.ch15, 2012.

Haralick, R. M., Shanmugam, K., and Dinstein, I.: Textural Features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics, SMC-3, 610–621, https://doi.org/10.1109/TSMC.1973.4309314, 1973.

Höhle, J. and Höhle, M.: Accuracy assessment of digital elevation models by means of robust statistical methods, ISPRS Journal of Photogrammetry and Remote Sensing, 64, 398–406, https://doi.org/10.1016/j.isprsjprs.2009.02.003, 2009.

Ibbeken, H. and Schleyer, R.: Photo-sieving: A method for grain-size analysis of coarse-grained, unconsolidated bedding surfaces, Earth Surface Processes and Landforms, 11, 59–77, https://doi.org/10.1002/esp.3290110108, 1986.

James, M. R. and Robson, S.: Mitigating systematic error in topographic models derived from UAV and ground-based image networks, Earth Surface Processes and Landforms, 39, 1413–1420, https://doi.org/10.1002/esp.3609, 2014.

Kellerhals, R. and Bray, D. I.: Sampling procedures for coarse fluvial sediments, Journal of the Hydraulics Division, 97, 1165–1180, 1971.

Kondolf, G. M.: PROFILE: hungry water: effects of dams and gravel mining on river channels, Environmental management, 21, 533–551, https://doi.org/10.1007/s002679900048, 1997.

Kondolf, G. M. and Wolman, M. G.: The sizes of salmonid spawning gravels, Water Resources Research, 29, 2275–2285, https://doi.org/10.1029/93WR00402, 1993.

Lamb, M. P. and Venditti, J. G.: The grain size gap and abrupt gravel-sand transitions in rivers due to suspension fallout, Geophysical Research Letters, 43, 3777–3785, https://doi.org/10.1002/2016GL068713, 2016.

Langhammer, J., Lendzioch, T., Miřijovský, J., and Hartvich, F.: UAV-Based Optical Granulometry as Tool for Detecting Changes in Structure of Flood Depositions, Remote Sensing, 9, https://doi.org/10.3390/rs9030240, 2017.

Lloyd, S.: Least squares quantization in PCM, IEEE Transactions on Information Theory, 28, 129–137, https://doi.org/10.1109/TIT.1982.1056489, 1982.

Otsu, N.: A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics, 9, 62–66, https://doi.org/10.1109/TSMC.1979.4310076, 1979.

Paola, C., Parker, G., Seal, R., Sinha, S. K., Southard, J. B., and Wilcock, P. R.: Downstream Fining by Selective Deposition in a Laboratory Flume, Science, 258, 1757–1760, https://doi.org/10.1126/science.258.5089.1757, 1992.

Parker, G., Klingeman, P. C., and McLean, D. G.: Bedload and size distribution in paved gravel-bed streams, Journal of the Hydraulics Division, 108, 544–571, 1982.

Pearson, E., Smith, M., Klaar, M., and Brown, L.: Can high resolution 3D topographic surveys provide reliable grain size estimates in gravel bed rivers?, Geomorphology, 293, 143–155, https://doi.org/10.1016/j.geomorph.2017.05.015, 2017.

Purinton, B. and Bookhagen, B.: Measuring decadal vertical land-level changes from SRTM-C (2000) and TanDEM-X (~ 2015) in the south-central Andes, Earth Surface Dynamics, 6, 971–987, https://doi.org/10.5194/esurf-6-971-2018, 2018.

Purinton, B. and Bookhagen, B.: PebbleCounts: a Python grain-sizing algorithm for gravel-bed river imagery, https://doi.org/10.5880/fidgeo.2019.007, https://github.com/UP-RS-ESP/PebbleCounts, 2019.

Rice, S. and Church, M.: Sampling surficial fluvial gravels; the precision of size distribution percentile sediments, Journal of Sedimentary Research, 66, 654, https://doi.org/10.2110/jsr.66.654, 1996.

Rice, S. and Church, M.: Grain size along two gravel-bed rivers: statistical variation, spatial pattern and sedimentary links, Earth Surface Processes and Landforms, 23, 345–363, https://doi.org/10.1002/(SICI)1096-9837(199804)23:4<345::AID-ESP850>3.0.CO;2-B, 1998.

Rubin, D. M.: A Simple Autocorrelation Algorithm for Determining Grain Size from Digital Images of Sediment, Journal of Sedimentary Research, 74, 160, https://doi.org/10.1306/052203740160, 2004.

Russ, J. C.: The image processing handbook, fourth edition, CRC press, 2002.

Rychkov, I., Brasington, J., and Vericat, D.: Computational and methodological aspects of terrestrial surface analysis based on point clouds, Computers & Geosciences, 42, 64–70, https://doi.org/10.1016/j.cageo.2012.02.011, 2012.

Sculley, D.: Web-scale K-means Clustering, in: Proceedings of the 19th International Conference on World Wide Web, pp. 1177–1178, ACM, New York, NY, USA, https://doi.org/10.1145/1772690.1772862, 2010.

Shields, A.: Anwendung der Aehnlichkeitsmechanik und der Turbulenzforschung auf die Geschiebebewegung, Ph.D. thesis, Technical University Berlin, 1936.

Sime, L. and Ferguson, R.: Information on Grain Sizes in Gravel-Bed Rivers by Automated Image Analysis, Journal of Sedimentary Research, 73, 630, https://doi.org/10.1306/112102730630, 2003.

Sklar, L. S., Dietrich, W. E., Foufoula-Georgiou, E., Lashermes, B., and Bellugi, D.: Do gravel bed river size distributions record channel network structure?, Water Resources Research, 42, W06D18, https://doi.org/10.1029/2006WR005035, 2006.

Smith, M., Carrivick, J., and Quincey, D.: Structure from motion photogrammetry in physical geography, Progress in Physical Geography: Earth and Environment, 40, 247–275, https://doi.org/10.1177/0309133315615805, 2015.

Tofelde, S., Schildgen, T. F., Savi, S., Pingel, H., Wickert, A. D., Bookhagen, B., Wittmann, H., Alonso, R. N., Cottle, J., and Strecker, M. R.: 100 kyr fluvial cut-and-fill terrace cycles since the Middle Pleistocene in the southern Central Andes, NW Argentina, Earth and Planetary Science Letters, 473, 141–153, https://doi.org/10.1016/j.epsl.2017.06.001, 2017.

Tomasi, C. and Manduchi, R.: Bilateral filtering for gray and color images, in: Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), pp. 839–846, https://doi.org/10.1109/ICCV.1998.710815, 1998.

Verdú, J. M., Batalla, R. J., and Martínez-Casasnovas, J. A.: High-resolution grain-size characterisation of gravel bars using imagery analysis and geo-statistics, Geomorphology, 72, 73–93, https://doi.org/10.1016/j.geomorph.2005.04.015, 2005.

Verma, A. K. and Bourke, M. C.: A method based on structure-from-motion photogrammetry to generate sub-millimetre-resolution digital elevation models for investigating rock breakdown features, Earth Surface Dynamics, 7, 45–66, https://doi.org/10.5194/esurf-7-45-2019, https://www.earth-surf-dynam.net/7/45/2019/, 2019.

Ward, J. H.: Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58, 236–244, https://doi.org/10.1080/01621459.1963.10500845, 1963.

Warrick, J. A., Rubin, D. M., Ruggiero, P., Harney, J. N., Draut, A. E., and Buscombe, D.: Cobble cam: grain-size measurements of sand to boulder from digital photographs and autocorrelation analyses, Earth Surface Processes and Landforms, 34, 1811–1821, https://doi.org/10.1002/esp.1877, 2009.

Westoby, M. J., Dunning, S. A., Woodward, J., Hein, A. S., Marrero, S. M., Winter, K., and Sugden, D. E.: Sedimentological characterization of Antarctic moraines using UAVs and Structure-from-Motion photogrammetry, Journal of Glaciology, 61, 1088–1102, https://doi.org/10.3189/2015JoG15J086, 2015.

Wohl, E. E., Anthony, D. J., Madsen, S. W., and Thompson, D. M.: A comparison of surface sampling methods for coarse fluvial sediments, Water Resources Research, 32, 3219–3226, https://doi.org/10.1029/96WR01527, 1996.

Wolcott, J. and Church, M.: Strategies for sampling spatially heterogeneous phenomena; the example of river gravels, Journal of Sedimentary Research, 61, 534–543, https://doi.org/10.1306/D4267753-2B26-11D7-8648000102C1865D, 1991.

Wolman, M. G.: A method of sampling coarse river-bed material, Eos, Transactions American Geophysical Union, 35, 951–956, https://doi.org/10.1029/TR035i006p00951, 1954.

Woodget, A. S. and Austrums, R.: Subaerial gravel size measurement using topographic data derived from a UAV-SfM approach, Earth Surface Processes and Landforms, 42, 1434–1443, https://doi.org/10.1002/esp.4139, 2017.

Woodget, A. S., Fyffe, C., and Carbonneau, P. E.: From manned to unmanned aircraft: Adapting airborne particle size mapping methodologies to the characteristics of sUAS and SfM, Earth Surface Processes and Landforms, 43, 857–870, https://doi.org/10.1002/esp.4285, 2018.

Supplement — Introducing PebbleCounts: A grain-sizing tool for photo surveys of dynamic gravel-bed rivers

Benjamin Purinton[1] and Bodo Bookhagen[1]

[1]Institute of Earth and Environmental Science, Universität Potsdam, Potsdam, Germany

Correspondence: Ben Purinton (purinton@uni-potsdam.de)

S1. Additional Data Dimensions from Point Clouds

The results presented ~~here~~ in the main manuscript are similar to other studies segmenting grains from 2D imagery (e.g., Detert and Weitbrecl . This ignores the potential to exploit the third height dimension of the data from irregularly spaced SfM-MVS (or lidar) point clouds and associated DEMs. Many authors have already begun to look at patch-scale variance or roughness ~~(e.g., Rychkov et al., 2012)~~ (e.g., Rychkov et al., 2012; Brasington 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 near-nadir photos taken from a height of ∼~~5~~ 4.5 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 nadir+oblique 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 and Bourke, 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~~ nadir 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 $\sim$1 mm in multiple settings and distances (up to $\sim$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.

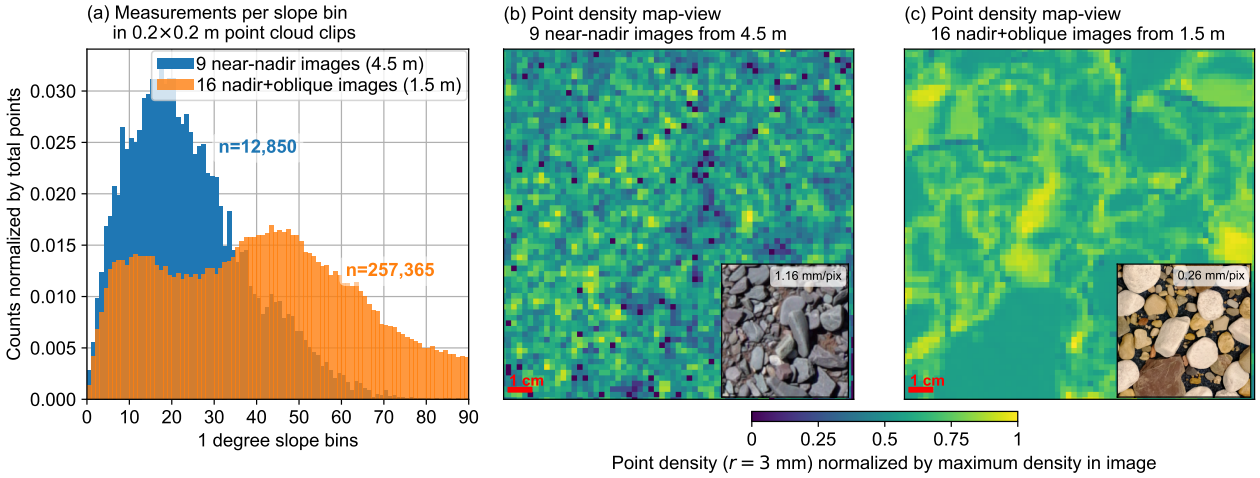

Figure S1. (a) Slope distribution in field (~~top-down~~near-nadir) and experimental (nadir+oblique) point cloud clips. The point cloud slope was calculated in *CloudCompare* (CloudCompare, 2018) by first calculating the normals at each point using the 6 nearest neighbors and then extracting the dip of each normal. (b) Map-view of point density normalized by the maximum for the 9 ~~top-down~~ near-nadir field images and (c) the same for the 16 nadir+oblique experimental images. Point density was calculated as the number of points in a radius of 3 mm. The clips were from a 0.2×0.2 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 nadir+oblique photo setting was 59 points/cm^2, whereas, in the field using 9 ~~top-down~~ near-nadir 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.

S2. Command-line Variables and Example Screenshots for *PebbleCounts*

Table S1 shows the command-line variables for *PebbleCounts* (KMS approach) and Table S2 shows the command-line variables for *PebbleCountsAuto* (AIF approach). Examples of the command-line interface and manual clicking steps are shown in Figure S2 and Figure S3, respectively.

Table S1. Command-line variable flags in *PebbleCounts* and their meaning. The default values are effective for most images.

Variable Flag	**Meaning (units)**	**Default Value(s) and Suggested Range**
im	Image to run, including path to folder	No default
ortho	Georeferenced orthoimagery flag	No default, 'y' for orthoimagery, 'n' for ~~top-down~~ nadir
input_resolution	Input resolution if not orthoimage (mm)	No default, calculate from eq. (3)
subset	Interactively subset image	Default no ('n')
*sand_mask**	Name, including path, to a sand mask if one already exists	No default
*otsu_threshold**	Percentage of Otsu value to threshold shadows by (percentage of 100)	No default, suggested value of 50
*maxGS**	Expected maximum a-axis grain size (m)	Default 0.3
*cutoff**	Minimum b-axis length to be counted (pixels)	Default 20, can be raised
*min_sz_factors**	Factors to multiply *cutoff* at each scale, used to cleanup masks for easier clicking	Default [50, 5, 1] for three scales (large to small) for $\sim$1 mm/pixel imagery, double for < 0.8 mm/pixel
*win_sz_factors**	Factors to multiply *maxGS* by at each scale	Default [10, 3, 2] for three scales (large to small), can be changed $\pm$0.5–1.5 to get more or less windows
*improvement_ths**	Improvement threshold values that tell k-means when to halt (fraction of 1)	Default [0.01, 0.1, 0.1] for three scales (large to small), can be varied from 0.01–0.2
*coordinate_scales**	Fraction to scale *x,y* coordinates (fraction of 1)	Default [0.5, 0.5, 0.5] for three scales (large to small), can be varied from 0.3–0.7
*overlaps**	Fraction of overlap between windows (fraction of 1)	Default [0.5, 0.3, 0.1] for three scales (large to small), can be varied from 0–0.5 at each scale
*first_nl_denoise**	Strength of first non-local means denoising	Default 5, can be varied $\pm$1
*nl_means_chroma_filts**	Strength of windowed non-local means denoising	Default [3, 2, 1] for three scales (large to small), can be varied $\pm$1
*bilat_filt_szs**	Size of bilateral filtering windows (pixels)	Default [9, 5, 3] for three scales (large to small), can be varied from 3–9
*tophat_th**	Upper percentile threshold to take from top-hat filter for edge detection (fraction of 1)	Default 0.9, can be varied from 0.8–0.95
*sobel_th**	Upper percentile threshold to take from sobel filter for edge detection (fraction of 1)	Default 0.9, can be varied from 0.8–0.95
*canny_sig**	Canny filtering sigma value for edge detection	Default 2, can be varied from 1–2
resize	Value to resize windows by (fraction of 1)	Default 0.8, can be varied from 0.5–0.99 if you want a smaller (0.5) or larger (0.99) pop-up window

*Influence on results

Table S2. Command-line variable flags in *PebbleCountsAuto* and their meaning. The default values are effective for most images.

Variable Flag	**Meaning (units)**	**Default Value(s) and Suggested Range**
im	Image to run, including path to folder	No default
ortho	Georeferenced orthoimagery flag	No default, 'y' for orthoimagery, 'n' for ~~top-down~~ nadir
input_resolution	Input resolution if not orthoimage (mm)	No default, calculate from eq. (3)
subset	Interactively subset image	Default no ('n')
*sand_mask**	Name, including path, to a sand mask if one already exists	No default
*otsu_threshold**	Percentage of Otsu value to threshold shadows by (percentage of 100)	No default, suggested value of 50
*cutoff**	Minimum b-axis length to be counted (pixels)	Default 20, can be raised
*percent_overlap**	Maximum allowable overalp between neighboring ellipses for filtering suspect grains (percentage of 100)	Default 15, can be varied from 5–30
*misfit_threshold**	Maximum allowable misfit between ellipse and grain mask for filtering suspect grains (percentage of 100)	Default 30, can be varied from 10–50
*min_size_threshold**	Minimum area of grain, used to clean the mask (pixels)	Default 10 for $\sim$1 mm/pixel imagery, 40 for $<$ 0.8 mm/pixel
*first_nl_denoise**	Strength of first non-local means denoising	Default 5, can be varied $\pm$1
*tophat_th**	Upper percentile threshold to take from top-hat filter for edge detection (fraction of 1)	Default 0.9, can be varied from 0.8–0.95
*sobel_th**	Upper percentile threshold to take from sobel filter for edge detection (fraction of 1)	Default 0.9, can be varied from 0.8–0.95
*canny_sig**	Canny filtering sigma value for edge detection	Default 2, can be varied from 1–2
resize	Value to resize windows by (fraction of 1)	Default 0.8, can be varied from 0.5–0.99 if you want a smaller (0.5) or larger (0.99) pop-up window

*Influence on results

S3. GSDs Separated by Site

Figure ?? and ?? show the results of the KMS approach to grain sizing on a site-by-site basis before and following a 20-pixel lower truncation of the distribution. On the left columns are the initial runs of KMS *PebbleCounts* without any lower truncation, and on the right columns are the re-running with a 20-pixel truncation. Aside from three sites (S09, S10, and S24) the p-values for the KS-tests without truncation were < 0.01, indicating a robust rejection of the null hypothesis. Following truncation, all 12 sites have p-values > 0.1, except S16, which has p=0.09.

Despite a high p-value, S24 demonstrates a stronger bias in the GSD towards coarser grains (up to 0.5 ψ discrepancy), as indicated by the high A_{diff} value of −1.36. Here, the KS-test pass is likely caused by the small sample size remaining after truncation (n=24), the least of any site. The poor performance of S24 was expected given the large size range with many sub-cm pebbles and a few large boulders, strong cast shadows from the large grains, and intra-granular edges on angular boulders with quartz veins (see Figure 9b in the main manuscript). 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. The site S34 also had a high A_{diff}=−2.11. In this case, poor performance is due to significant blurriness of this image, and again a small sample size (n=47).

Figure S2. ~~Results from KMS~~ Example of command-line and pop-up interface for *PebbleCounts*~~on~~ . (a~~site-by-site basis with the initial run in the left columns~~) Interactive Otsu thresholding using percentage of Otsu value and yes (‘y’) or no ~~truncation~~(‘n’) ~~and the truncated~~ confirmation. (~~at 20-pixels~~b) ~~run in the right columns~~Interactive color masking by yes (‘y’) or no (‘n’) and resulting color mask after selection. ~~The control data are given as a solid black line with the number of pebbles~~ (~~n~~c) ~~shown in~~ K-means clustering and pop-up window for pebble selection by left clicking, with black ~~. KMS *PebbleCounts* results are~~ arrows measured in ~~gray~~ final output and ~~dashed~~red arrows ignored after right-click removal (see Fig. ~~The *p*-value results of a KS-test are also shown~~S3). ~~A_{diff} is the approximate integral between the curves. Note the reduction in x-axis scale between the columns, where the right, truncated distributions are plotted on a narrower range to emphasize the remaining discrepancies.~~

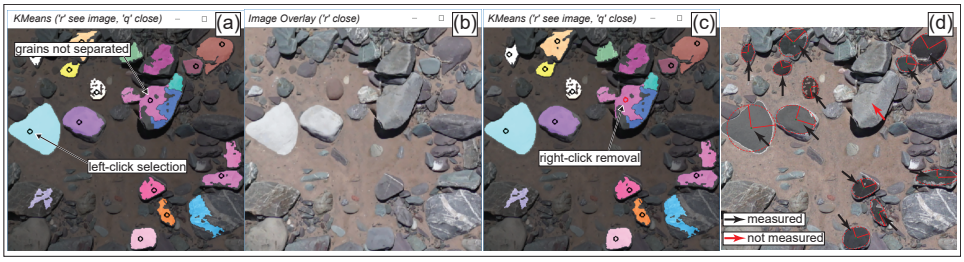

Figure S3. ~~Results~~ Clicking tutorial continued from ~~KMS *PebbleCounts* on~~ Figure S2c. Following k-means clustering at each scale a ~~site-by-site basis with~~ mask overlaid on the ~~initial run in the left columns~~ original image is presented (~~no truncationa~~), and ~~the truncated (at 20-pixels) run in the right columns. The control data~~ grains are ~~given as~~ selected by a ~~solid black line with~~ left click anywhere in the ~~number of pebbles (n) shown~~ segmented area, resulting in a black circle at the click location. ~~KMS *PebbleCounts* results are in gray and dashed. The p-value results of a KS-test are also shown. A_{diff}~~ When clicking is finished the ~~approximate integral between the curves~~mask is closed by pressing 'q'. ~~Note~~ To view the ~~reduction in x-axis scale between~~ original unmasked image the ~~columns, where~~ user may press 'r' (b). Using this switching the ~~right, truncated distributions~~ user can see which grains are ~~plotted on~~ poorly delineated and remove the last click with a ~~narrower range~~ right click on the mouse (c). The original black circle selection turns to ~~emphasize~~ red to signify this grain is off and will not be measured in the ~~remaining discrepancies~~final output (d).

~~S4~~S3. Resampling and Parameter Selection in AIF Approach

Figure S4 demonstrates the percentage of grains with a match found in the AIF approach ~~(matches are defined as in main manuscript Figure 18)~~ when increasing resampling from a factor of 0.6–2.6 by 0.1 steps using Lanczos resampling (Lanczos, 1950). As the resampling factor increases, there is progressive reduction in the number of found grains after filtering, therefore we selected the original resolution (resampling factor of one). Figure S5 and Figure S6 demonstrate two cases where the resampling slightly improved the resulting ~~GSD~~grain-size distribution. Both images were of relatively low quality with significant blurring and the presence of many weak edges between grains of similar color.

We selected a maximum percent misfit between the ellipse and grain of 30% as the 90^{th} percentile of misfits for the KMS approach was 30%. Furthermore, we allowed a maximum overlap between neighboring ellipses of 15%, visually selected to minimize overlapping grain measurement and over-segmentation of discrete grains. For the higher resolution imagery it was necessary to use a lower sobel and top-hat threshold (0.85), since we consider all the edges at once in the AIF approach, rather than in a windowed subset as in the KMS approach, and many edges are not found when using the 0.9 threshold given the increased number of pixels under consideration.

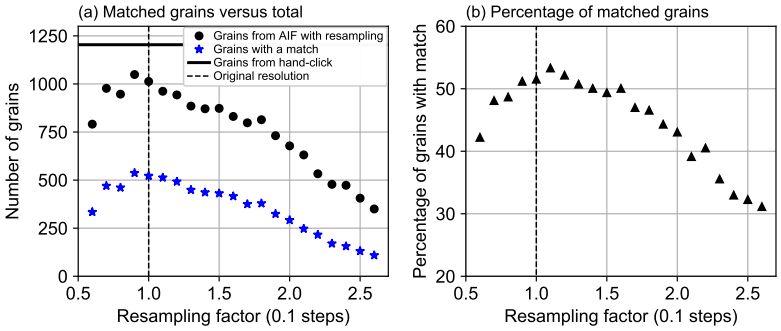

Figure S4. Matching grains found in each filtered mask versus the resampling factor (where 1 is the original image) for the ∼1.16 mm/pixel resolution images. Matches are defined as an AIF grain within 5 pixels of the hand-clicked line or the KMS grain centroid and with a 1 cm maximum b-axis difference between the AIF grain and the match.

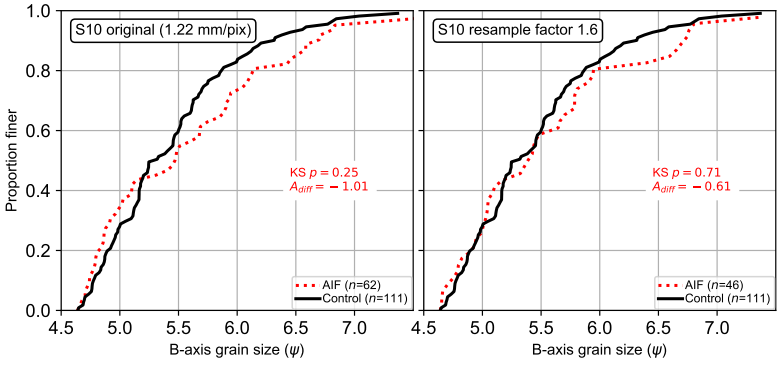

Figure S5. Slight improvement (increase in p and decrease in A_{diff}) in result using a 1.6-times resampling factor prior to running the AIF algorithm for the difficult (somewhat blurry, weak edges) S10 orthoimage.

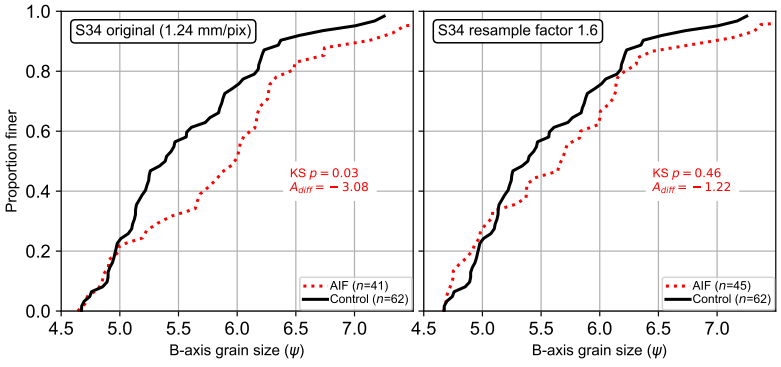

Figure S6. Slight improvement (increase in p and decrease in A_{diff}) in result using a 1.6-times resampling factor prior to running the AIF algorithm for the difficult (very blurry, weak edges) S34 orthoimage.

S4. Agisoft Orthomosaic Generation

Agisoft (Agisoft, 2018) processing was carried out in the following steps for the indoor handheld imagery (with field-gathered mast imagery differences in parantheses following the step):

1. Image quality detection and the exclusion of photos with quality metric < 0.7. This step analyzes pixel contrast to estimate sharpness with values ranging from 0/blurred to 1/sharp. We found 0.7 to be a sufficient lower cutoff upon visual inspection of results.

2. Detection of 12-bit coded targets in the remaining photos, with two targets placed at each of the four corners of the area and ensuring that the diameter of the printed targets' center circle was limited to 10–30 pixels in image resolution for successful automated detection.

3. Input of scale for the orthomosaic output, provided by the distances between the targets at each corner, resulting in four distance measurements, with 0.5 mm accuracy using a ruler with cm and mm demarcations. (For the field images: The scale was provided by the XYZ coded target locations in UTM zone 19S, WGS84 ellipsoidal datum.)

4. Photo alignment at high quality with a 40,000 key-point and 2000 tie-point limit.

5. Dense cloud generation from the aligned photos at the medium output and with moderate depth filtering. Given the high quality of the photos more aggressive options did not improve results. (For the field images: Given the increased complexity of the setting and imperfect photo collection, the dense point cloud was generated at high quality with aggressive depth filtering.)

6. DEM building from the dense cloud with default settings in a local coordinate system. (For the field images: The DEMs and orthomosaics were also output in UTM zone 19S projections, providing undistorted pixels with resolution in m/pixel.)

7. Generation of an orthomosaic using the DEM for orthorectification at the default settings.

8. Output of the orthomosaic to a GeoTiff file with resolution provided in m/pixel.

S5. KMS and AIF Results Separated by Site

Here we show all of the results (following 20-pixel truncation) for each of the 12 sites in Figure S7. These results are aggregated in curves shown in the main manuscript Figure 11 and a comparison of the individual percentiles of interest is shown in the main manuscript Figure 12.

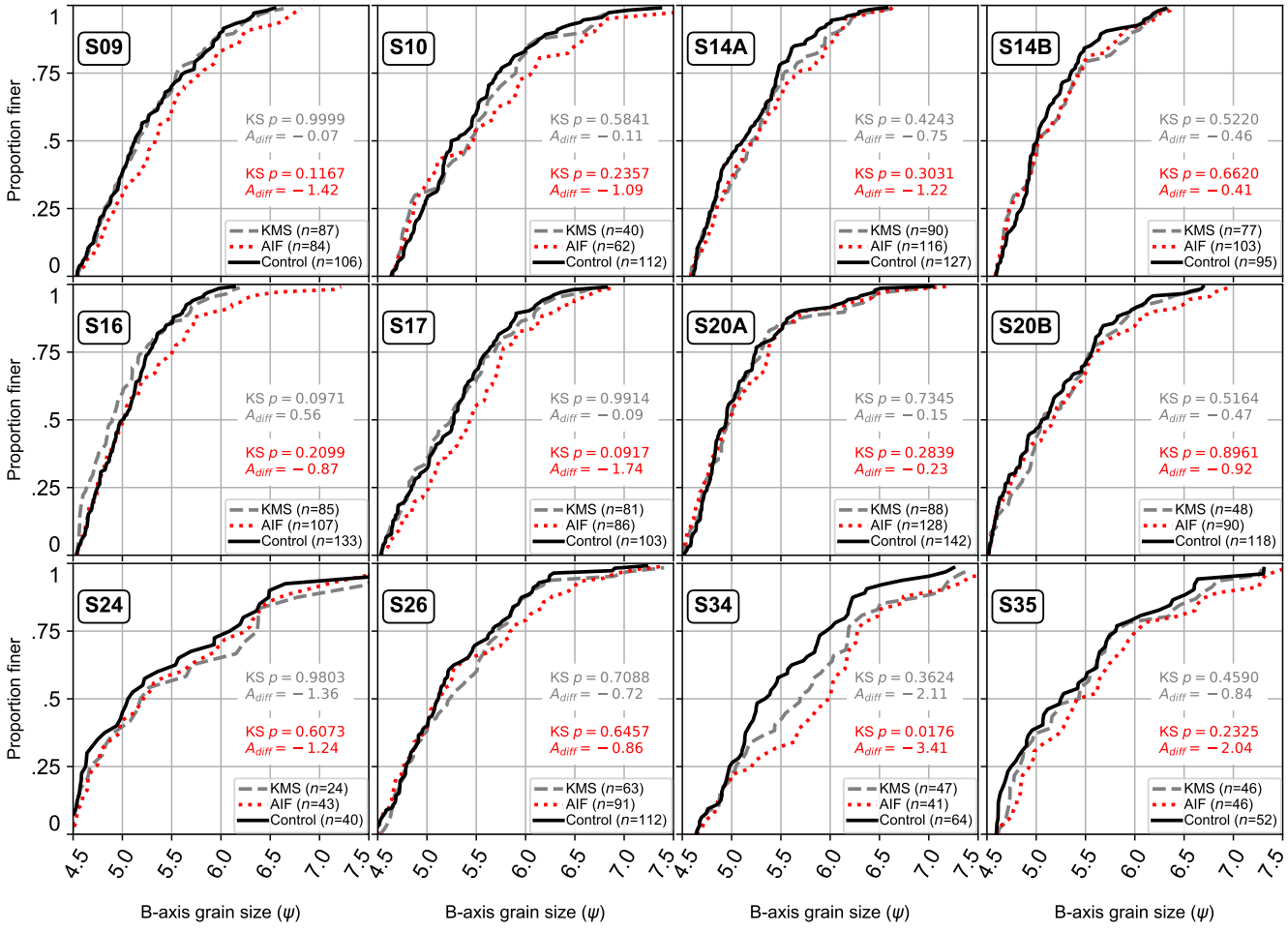

Figure S7. Comparison of 20-pixel truncated grain-size distributions between hand-clicked control (black line), KMS *PebbleCounts* (gray, dashed line), and AIF *PebbleCountsAuto* (red, dotted line) for the 12 × ~1.16 mm/pixel control sites. In corresponding colors are the *p*-value results of a KS-test and the A_{diff} approximate integral between the curves for each approach versus the control data. The legend indicates the number of grains (*n*) making up each curve. See Figure 6b in the main manuscript for sites.

S6. Misidentification in the AIF Approach

Figure S8 demonstrates remaining issues with the AIF approach in a few map-view examples. On a grain-by-grain basis, there are many inaccuracies falling into three main categories: over-segmentation of grains with internal edges and the selection of each segment as a separate grain, under-segmentation and merging of neighboring grains that have weak edges sometimes caused by image blur, and misidentification of non-grain objects or clusters of small grains. It is clear from this analysis that caution must be used when interpreting AIF results, particularly in complex or blurry images.

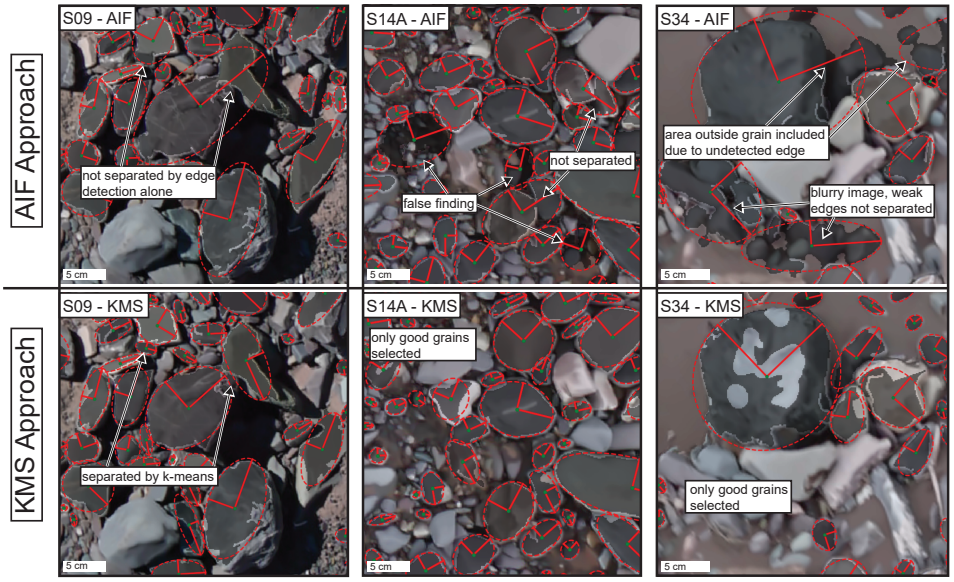

Figure S8. Resulting delineated grains using the AIF *PebbleCountsAuto* function (top row) versus the same area from the KMS *PebbleCounts* function (bottom row). Labels indicate the issues with the AIF results and improvement in KMS results. Note the poor results for the blurry image on the right (S34).

References

Agisoft: AgiSoft PhotoScan Professional, http://www.agisoft.com/downloads/installer/, 2018.

Alonzo, M., Bookhagen, B., McFadden, J. P., Sun, A., and Roberts, D. A.: Mapping urban forest leaf area index with airborne lidar using penetration metrics and allometry, Remote Sensing of Environment, 162, 141–153, https://doi.org/10.1016/j.rse.2015.02.025, 2015.

Brasington, J., Vericat, D., and Rychkov, I.: Modeling river bed morphology, roughness, and surface sedimentology using high resolution terrestrial laser scanning, Water Resources Research, 48, W11 519, https://doi.org/10.1029/2012WR012223, 2012.

Chen, Q., Baldocchi, D., Gong, P., and Kelly, M.: Isolating Individual Trees in a Savanna Woodland Using Small Footprint Lidar Data, Photogrammetric Engineering & Remote Sensing, 72, 923–932, https://doi.org/10.14358/PERS.72.8.923, 2006.

CloudCompare: CloudCompare Software, http://www.cloudcompare.org/, 2018.

Cullen, N. D., Verma, A. K., and Bourke, M. C.: A comparison of structure from motion photogrammetry and the traversing micro-erosion meter for measuring erosion on shore platforms, Earth Surface Dynamics, 6, 1023–1039, https://doi.org/10.5194/esurf-6-1023-2018, https://www.earth-surf-dynam.net/6/1023/2018/, 2018.

Detert, M. and Weitbrecht, V.: Automatic object detection to analyze the geometry of gravel grains–a free stand-alone tool, in: River flow 2012 : Proceedings of the international conference on fluvial hydraulics, San José, Costa Rica, September 5-7, 2012, pp. 595–600, Taylor & Francis Group, London, 2012.

Lanczos, C.: An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. Natl. Bur. Stand. B, 45, 255–282, 1950.

Rychkov, I., Brasington, J., and Vericat, D.: Computational and methodological aspects of terrestrial surface analysis based on point clouds, Computers & Geosciences, 42, 64–70, https://doi.org/10.1016/j.cageo.2012.02.011, 2012.

Verma, A. K. and Bourke, M. C.: A method based on structure-from-motion photogrammetry to generate sub-millimetre-resolution digital elevation models for investigating rock breakdown features, Earth Surface Dynamics, 7, 45–66, https://doi.org/10.5194/esurf-7-45-2019, https://www.earth-surf-dynam.net/7/45/2019/, 2019.