Response to reviewer comment 2 (RC2)

David Mair[1], Ariel Henrique Do Prado[1], Philippos Garefalakis[1], Alessandro Lechmann[1], Alexander Whittaker[2], Fritz Schlunegger[1]

[1] Institute of Geological Sciences, University of Bern, Baltzerstrasse 1+3, 3012 Bern, Switzerland
[2] Imperial College, Department of Earth Science and Engineering, South Kensington Campus, London SW7 2AZ, United Kingdom

Correspondence to: David Mair (david.mair@geo.unibe.ch)

General response

We thank the reviewer for the constructive and insightful comments. We will first address the main comments before responding to the line-specific comments below. The reviewer raises 4 main concerns:

1) […] the manuscript needs restructuring and some significant shortening. For instance, the discussion should be shortened because it is partly repetitive, e.g., uncertainties are repeated. More importantly the manuscript needs shortening in regard of the SfM processing versions. The explanation of the flight setup and corresponding error results are repetitive to already existing literature (which is also emphasized by the authors themselves at 330-334). I would suggest to focus solely to what is new to the existing literature (chapter 3.2) and to strongly shorten the SfM processing display of methods, results and discussion in that regard (especially 4.1) and focus only on relevant aspects for the grain size estimations.

While we unreservedly agree on the need for eliminating repetitive statements to be as concise as possible, we consider the question how much UAV/SfM background being appropriate very tricky. The delicacy of this is also underlined by the comments of reviewer 1, which among other points, show that simple reporting of UAV survey choices without sufficient background might potentially mislead readers with the consequence that some further clarifications are even needed (see AR_RC1). The main problem we see here is that "*relevant aspects for the grain size estimations"* are almost impossible to disentangle from the overall UAV/SfM pipeline. For example, we consider it relevant to show that using the JPEG image format not only increased the systematic error in some models, but also led to failure of the photogrammetric alignment for some weak image network geometries (despite high overlap). This would not be conceivable for the reader without introducing the UAV/SfM strategy. Furthermore, we want to note here that we follow recommendations, i.e. James et al. (2019) and Eltner and Sofia

(2020), for choosing what to report in an effort to maximise the transparency and reproducibility. These guidelines build on the well-documented fact that for UAV/SfM workflows survey-specific traits can have significant impact on the model results (e.g., James et al. 2020; Sanz-Ablanedo et al., 2020; Hastedt et al., 2021). Therefore, we consider reporting flying conditions, SfM processing choices and resulting model quality important for readers to contextualize our results.

Nevertheless, to avoid repeating SfM literature and to increase conciseness, we shortened the incremented sections 3.2 (from ~400 to ~200 words) and 4.1 (from ~850 to ~520 words) significantly, amounting to roughly 35 lines of text being removed (with most of it, about ~18 lines, in section 4.1). However, while we shortened the method section somewhat, we think that we cannot omit any more statements without compromising the reproducibility of our study. Thus, the only way we can envision shortening the text further is by moving most of sections 2.2, 3.2 and 4.1 to an Appendix. However, this could force readers to scroll forth and back repeatedly to understand the context of the grain size data (e.g., what the different coloured data in Figs. 5, 6, 8 and 9 represent and why they exist at all). Furthermore, it would require such an Appendix to include method details, some results and a discussion thereof or, alternatively, force us to reference such an Appendix in the discussion section repeatedly.

Regarding *"[…] e.g., uncertainties are repeated":* We checked, but we could not find any uncertainty that was mentioned more than once in the text (just in case the comment was meant this way). We concede here that in section 4.2 we do give an overall range of median single percentiles and mean modelled key percentile values, which summarize results previously reported in sections 3.3 and 3.5. We do so because we think their overall trends and implications are discussed in the following lines and thus summarizing them facilitates the reading.

2) In addition, the authors might also consider different options to model interior geometry because it can have a strong influence on the 3D model (and thus orthomosaic) quality, especially considering Fourier models for DJI P4 UAV models with unique distortion patterns (Hastedt et al., 2021). But again, not displaying the SfM method itself in too much detail (better referring to the literature) is suggested, but instead focus on the relevance for the GSD.

We thank the reviewer for raising this point and we are excited for any improvements in camera modelling during the bundle adjustment. Ideally, a reliable camera modelling during the photogrammetric alignment can correct distortion to a level that leaves local residuals in the order of ~1 px or less in the image space. However, to estimate uncertainties for grain sizes in orthomosaics, we

use a *shape error* (previously named pixel uncertainty) to model inter-pixel error globally in a conservative way (with values up to 6.4 px for some of our results). This quantity could be varied to account for larger camera model residuals and is independent on the camera model choice during the bundle adjustment. As for percentile uncertainty of grain size results, currently uncertainties from image errors are much smaller than the effect of counting statistics, segmentation or the ellipsoid fit for estimating the b-axis; therefore, we currently do not see the need to model uncertainty that varies locally. This is also the reason why we did not include uncertainty mapping for precision, doming in the model space or camera model residuals in the image space. However, we include a brief statement and the suggested reference in the method section 2.4 to highlight a potential inclusion of such methods in the future.

3) The decision for the error equations needs some more explanation in regard to how the authors derived them. And in general, it might be noted that the authors are not performing an error propagation according to a mathematical approach as they model the influence of different errors with MC and decide for equations, whose derivation is not obvious (please, see specific comments for more detail).

First and generally, we apologize for the sometimes-imprecise wording, which might lead to the impression that we considered our modelling to represent error propagation *sensu strictu*. We changed the concerning phrases accordingly throughout the text. Regarding the error parametrization, we thank the reviewer for catching inconsistencies and mistakes, which we gladly correct. We thoroughly reworked section 2.4, where we now more carefully explain our consideration upon error parametrization. Furthermore, we added statements on the applicability of our approach in light of camera geometry correction. For details, please see responses to the specific comments below.

4) The authors clearly highlighted why they did not consider AI based GSD calculation approaches in this study. Nevertheless, I think, it still is needed to discuss how the results in regard of the error impact from the SfM process at the GSD estimation might also be transferable to the techniques of AI that allow for direct grain size distribution estimation without the need of segmenting grains (Lang et al., 2021), which however still rely on SfM for scaled image assessments?

We thank the reviewer for opportunity to clarify this point further. We note here that we did not generally consider machine learning (ML) approaches. We did not find a suitable ML based model when we designed our study. Such a model would need to be capable of segmenting individual grains,

which was not available (except for the very recently published approach of Chen et al., 2022, which in theory could be used for our approach). We agree on the notion that there is a need to assess the error impact from the SfM process on texture-based models, such as proposed by Lang et al. (2021), since such models can only be as accurate/precise as the underlying data set. However, we cannot deliver such an assessment, because: 1) while our findings might also be transferable to the results of such models, our modelling method is not directly applicable to such approaches (as we need segmented grains). 2) Such ML models learn complex, non-linear relationships to predict distributions learned from training data of specific composition, quality and with a specific training schedule. All these factors make it very hard to predict exactly how much a specific uncertainty or error in the training data might influence the result. For example, a model might have learned to mitigate a specific error or it might regard it as predictive and thereby amplifying the effects of such models. Therefore, the only way to assess such effect is to compare the predictions of such models to independently referenced data sets. This is far beyond the scope of this study and we honestly could only speculate on the specific effects of SfM uncertainties on such ML models, from which we refrain. We note here that for the specific case of the Lang et al. (2021) model, the situation is even more unclear, as the model uses orthomosaics from several SfM surveys without GCPs from (as far we as we know) and neither the SfM models uncertainties, the resulting datasets (and their balance during training), nor the final model is openly accessible.

Line by line responses

L61: *Do the authors refer to indeed undistorted or ortho-rectified images? This has to be addressed thoroughly throughout the manuscript as there seems to be a mixed usage of the terminology undistorted images and undistorted orthoimages. If the image is an orthoimage it is undistorted by definition. However, an undistorted image does not need to be an orthoimage.*

We use single, undistorted nadir (precise to 0.1°) images, which are in that respect very similar to orthoimages (despite no rectification and under the assumption of negligible surface tilt). We clarified this throughout the text, where we now specifically explain what we mean with “single images” (undistorted, nadir images) and removed potentially misleading statements that called them orthoimages before. See also answers to corresponding comments below.

L99-100: *But if DL is used, should it not be transferable if the training data is large enough?*

This needs most likely a lot more research. While it might be possible that a suitable DL model with a large and diverse enough data set might be found, it is far from certain that this is possible. Instead, the natural complexity of such images (e.g., light conditions, vegetation, lithology differences) might be preventing such a model to find meaningful underlying similarities and thus preventing such a “one-shoe fits all” model. Therefore we restrain ourselves to state merely the current models have *so far* not been able to fully generalize.

L126-130: *What is the difference between internal consistency and systematic uncertainty? Systematic uncertainty and internal consistency are both influenced by e.g., insufficiently modelled interior camera geometries or an unfavourable image network geometry. Systematic errors can be caused by low internal consistencies.*

We used "internal consistency" to describe the "precision" in the sense of James et al. (2017; 2020). Therein it refers to the expected deviation of an estimated or measured value, i.e., the precision of the sparse tie points. We clarified the sentence accordingly. Regarding the mentioned relation between precision and systematic uncertainty, we refer to lines 127-135, where we briefly discuss this issue.

L135: *There are further errors that can be introduced during the generation of the final model, e.g., such as interpolation errors if a raster is derived or false matches during the dense matching in regions of repeating patterns or missing texture or in case of low image redundancy.*

We thank the reviewer for raising this point. Our statement refers to uncertainties relevant to grain size measurements, which are derived from a successful photogrammetric alignment during a SfM workflow. We realize that this was not reflected so far, therefore, we correct the text and add a small statement that cautions readers against the potential errors brought up by the reviewer. We note here that these errors can actually influence grain size measurements in orthophoto mosaics.

L143*: "or 3D point cloud roughness (Woodget and Austrums, 2017)"… The statement seems repetitive as it has already been mentioned before.*

While we acknowledge that the reference has been cited before in the general introduction before, we point out that this was always in the list with other methods. At this point (lines 143ff) we want to briefly, yet specifically discuss the two approaches and results by Woodget et al. (2018) and Woodget and Austrums (2017). Therefore, we opt to keep the short statement at this location.

L170: *"mechanical shutter"… Do you mean global shutter?*

Yes, as opposed to a rolling shutter. We changed the statement accordingly.

L180: *Please, change GNNS to GNSS.*

Changed.

Table 1: *Does QA refer to the manual removal of blurred images?*

It refers to the removal of images that were blurred, that were hard to align because of an insufficient depth of field due to too oblique camera angles, that were under- or overexposed, and images with insufficient contrast. A corresponding statement was added to the table caption.

L206: *Why did the authors not consider p2?*

We did not include p_2, because it has been shown that this parameter in combination with oblique camera angles might increase or even introduce systematic doming (see James et al. 2020). We did so to avoid potentially selectively biasing some models. We added a brief statement with reference to James et al. 2020 as explanation.

Figure 3: *Why is there an additional scaling needed? The orthophoto mosaic should already provide the information of scale.*

The scaling term (ε_{scale}) in Fig. 3 refers to a scale related uncertainty in the orthoimages on the scale of a measured length (i.e., b-axis). It is true that a general scale is provided by the orthoimage, which is largely governed by the quality of the underlying topographic model, i.e., by local model precision, systematic doming and local topography. However, for orthophoto mosaics the true scale of each mosaic part might differ due to errors in the SfM model. Therefore, we introduce the modelled scale uncertainty to the shape uncertainty, i.e., length uncertainty derived from a wrong scaling and incorrect stitching of images during the mosaicking. We do this because we cannot faithfully assume that for an orthomosaic, which was generated from a SfM mode with significant uncertainty, each final (and often super-sampled) pixel represents actually the same length. We added a clarification in section 2.4. See

also response to comment on lines 256-57 below. Furthermore, we removed the term “scale” from Fig. 3 to avoid confusion.

L223-224: *Please, briefly explain how James et al. (2020) consider the systematic doming to avoid that the reader needs to read the paper to grasp the concept.*

A brief explanation is now added.

L232-233: *Please, explain how you can estimate the camera height by considering “as distance of the camera centre to the corresponding centre points on the images”.*

Corrected the erroneous statement to state that we use “distance of the camera center to the horizontally closest 100 tie points using Euclidian distances”.

L234-235: *After which criteria did the authors choose the model regions? Was the selection performed randomly? And what is low and high confidence referring to?*

We now give information on the rationale of how we selected our region. We rephrased “high confidence” to “we selected areas with expected relative higher and lower model quality, with respect to image multiplicity, tie point precision and image noise due to water”. We note here that we did not randomly select regions.

L256-257: *“The image resolution, and thus the scale of single images, was estimated individually for undistorted and scaled single images”… This sentence is confusing in regard of the scaling. Why is the image scale estimated for an already scaled image? In case of the undistorted image, how did the authors account for perspective distortions,which leads to different scales across the image in the case of tilted images or non-planar surfaces? Or do the authors refer to an orthoimage? In the case of an orthoimage, those geometric effects would be accounted for.*

The image scaling refers to the calculation of the pixel scale from the image distance for the single images. As stated above, we use single, undistorted nadir (precise to 0.1°) images. Therefore, these single images do not include any significant perspective distortions for the purpose of this study. It is true that strongly tilted surfaces preclude such an approach. However, for our study, we used small and relative planar areas of our bar; thus, we consider the effect of potential tilting as minor here. We clarified the sentence.

L275-277: *If the single images are orthoimages (and not solely undistorted images), then also in that case effects of image alignment errors would be present (e.g., due to artefacts in the 3D model).*

“Single images” refers to single, undistorted nadir images; therefore such model artefacts do not affect them. This has been clarified throughout the manuscript (see also related comments above).

Eq. 2: *What is the final unit of ε*length*? At the current form it seems to be either pixel² if a and pixel error are in pixels or cm*pixel if a is in cm (or mm) and pixel error is in pixel. Is that correct? Furthermore, how did the authors decide for 2a*sqrt(a)? How was the equation derived?*

We changed the previously incorrect Eq. 2 to represent our modelling approach correctly (it represented an older approach, where *px* was realized as dimensionless factor). Now we use a shape error in metric length units and specify this in the text. We added an explanation for using *2a*sqrt(a)*, stating that we consider two times the pixel diagonal as measurement uncertainty. More information on the limits of such an approach is added further down in section 2.4.

Eq. 3: *Why multiply with 1? This would not be needed in the equation. Furthermore, if the authors consider error propagation, why did they not propagate then the error, i.e., sqrt(sum(sigma²))?*

We apologize for the erroneous equation. The * should be a +, in order to obtain a dimensionless scaling factor that would center at b_i and be between 0 and 1 for values smaller than the initial one and between 1 and ∞ for values larger than the initial one. We do not consider classical error propagation (see comments on related unclear phrasing above). We favor a randomization approach over the

proposed form, because it does not require errors to be normally distributed, as not all our modelled uncertainty is parametrized as a normal distribution (we use a uniform distribution for the doming).

L290: *Why only in z-direction? Are x and y not relevant for the segmentation? I would argue that the lateral error is important for the errors in grain size estimations from images. And why not use the actual spatial information, and therefore get a spatially distributed modelled error, instead of averaging for the MC approach? Furthermore, please, shortly explain how the precision is estimated with the James et al. (2020) tool, thus the reader still can grasp the concept without needing to read the paper.*

We use the errors in the z-direction to assess the uncertainty introduced in the scale of both single images and orthophoto mosaics. While for orthophoto mosaics, the error components of doming/bowling in the X,Y directions could contribute to a length uncertainty as well, the magnitudes of such errors in nadir dominated image networks is magnitudes smaller than the error in the z-direction (see James et al. 2020). Furthermore, such an error would strongly vary spatially, thus an approach that utilizes the spatial information might be more appropriate (see corresponding responses above). However, for the time being, we did not implement such an approach because of the expected higher computational costs and the expected much higher contribution of counting statistics and segmentation performance to grain size uncertainty. We added a related statement. We included a brief description of the precision export from Metashape by James et al. (2020).

Fig. 4: *DNG reveals a lower contrast than JPG images. However, as DNG refers to raw data (with i.e., 12 or 16 bit?), did the authors use some image processing to enhance the contrast?*

We use the 16 bit format from our UAV, directly imported into *Metashape*, which uses the "As Shot parameters" from the file. We converted all images to JPEG images before grain size measurements. We do not consider the lower contrast being in general an issue for metashape for the photogrammetric processing, therefore, we did not vary the contrast value at the SfM stage. We refrained from any additional image processing before determining the grain size in order to avoid any potential bias thereby (see also related comment and response below).

L330-331: *Why is the error so high in the z-direction (over 200 m)? Is that related to false GNSS-heights assigned to the UAV camera trigger locations (which is a known issue for some DJI models)?*

We consider this reported error in DJI GNSS models the cause of this offset. We moved the related statement from the discussion to the caption of Table 2. We removed the short paragraph that previously contained a summary of the SfM model quality to comply with the main concern 1 (see above).

L343-344: *"produce the highest uncertainties across all metrics. Models that are based on raw format images"… Is this due to issues of distortion estimations by DJI, which are not describable by a standard Brown model (James et al. 2020)?*

We suspect so, but we do not know how the generic on-board preprocessing is done. We note here that we do see some models with low uncertainties, where all uncertainties are within the expected margin (e.g., L2_2_C1). Therefore, we refrain from such a general statement.

L351: *GSD has already explained before.*

Explanation removed.

L358-359: *Indeed, I would expect more grains to be identified in the images compared to the ortho-mosaic due to the missing impact of smoothing, interpolation or general errors during the ortho-mosaic calculation process.*

Fig. 8: *What is the information provided by this figure? I am not able to see what it is supposed to tell in regard of the relation between flight pattern, image format and percentiles?*

Figure 8 shows the difference and magnitude of percentile uncertainty between grain size data from all different SfM models for 3 key percentiles, when comparing SI and OM. The figure is the only plot that

shows the data from all models (Figs. 5, 6 show a selection of model due to visual clarity). Furthermore, it shows that for models with no GCPs the percentile results were off the farthest when comparing SI to OM data (see blue and cyan data). It further shows that this is not always the case (e.g. K1_B).

L518: *Please, change "we" to "were".*

Changed.

Chapter 4.3: *is a mixture of results and discussion and should be split, putting the results into chapter 3.5.*

We discuss here the accuracy of the results where grain size data was collected on images, and we compare it with the results where the measurements were accomplished in the field. We acknowledge that we mention some results, however, we do not see a way to accomplish this discussion without briefly mentioning these results.

Fig. 9: *The DNG images seem to show higher errors than JPG images. Did the authors process the DNG to improve the contrast (as 12 or 16 bit(?) raw images are given) or did they stretch the grey values uniformly to 8 bit? This is an important aspect because if the latter is the case, then lower accuracy is not necessarily due to the image format but insufficient image processing.*

We use 16 bit DNG images, directly imported into metashape with "As Shot parameters" from the file, which we used for the photogrammetric processing. We then exported the all DNG (similar to JPEG) images as standard 24 bit JPEG (using the camera white balance) before using results of grain size measurements accomplished with *PebbleCounts* (as stated in section 2.2; see also Fig. 3 and related response above).

Moreover, we refrained from additional contrast enhancement for the DNG images, because while it is easy to increase by an arbitrary value, it is not straightforward to decide such a value without introducing potential additional bias. In fact, several strategies (e.g., histogram equalization, adaptive histogram equalization, CLAHE, SWAHE) for optimizing image contrast exist to adjust contrast images to an optimal level. Deciding for a particular strategy in itself, or for a single global value of contrast, might influence some images more than others. Additionally, one could also try correcting brightness or saturation, leading to more decisions, which would need careful consideration. To avoid these, we stuck to a minimal level of image processing.

However, we acknowledge here that, indeed for DNG based models, *PebbleCounts* segments generally fewer grains than for JPEG images (see also Table S4), which might be influenced by the lower contrast values in the DNG images. This might contribute to a slightly lower precision from counting statistics, in turn leading to slightly larger percentile uncertainties. However, we do not think we can actually infer that either format is producing better results with *PebbleCounts*, because under-segmentation occurred for all images, independent of the acquisition format. If at all, we would even consider the DNG based SI data (Fig. 9a) the most accurate if compared to the field data, despite slightly lower percentile precision. We note here that we do not see systematic differences between results from DNG and JPEG derived images, other than the variation in number of grains segmented. We further point out that for most of our images the relative difference in number of segmented grains is not strongly relevant, as this would at largest add a few additional percent to percentile uncertainty (cf. Eaton et al. 2019; e.g., their Figs. 10, 11), due to the still high number of segmented grains (> 300). Nevertheless, we realize now this might be important for readers, therefore we added a brief statement in section 2.2 to inform about our decision not to apply a contrast correction. We further added a brief statement to section 4.2 to indicate the potential effect of low contrast in DNG images on the number of grains found by *PebbleCounts.*

L555-557: *How does an in-accurate SfM model result in a view shift? What is meant by that? And how does a different view influence the segmentation? An orthophoto should always lead to a "Nadir" view (the same as for the orthoimage)?*

An in-accurate SfM model might lead to measuring grain sizes for different areas of a bar. Statement clarified.

L561-566: *I do not agree with the reasoning that the orthomosaic error might be larger due to the automatic PebbleCount application. The authors state, as well, that the error also occurred for the single images. Thus, the third reason is not a reason in regard of the orthomosaic but a general reason for the under-segmentation with PebbleCount. The orthomosaic errors already discussed under the second reason also lead to errors with the automatic PebbleCount.*

We wanted to state here that segmentation performance might amplify the uncertainties introduced for the reasons 1 and 2. For low image quality, PebbleCounts, or actually most likely any segmentation method, might accidentally find pebbles of a specific size fraction less or more often than expected. We rephrased the statement accordingly.