The authors discuss the influence of errors in the SfM process on the grain size estimation based on segmentation approaches. Thereby, they evaluate different SfM process settings, image formats and the usage of either single images or an orthomosaic. This kind of assessment is clearly needed to go forward in the application of grain size estimation based on imagery.

However, I think, 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.

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.

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
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?

Please, see below for some more specific comments.

Specific comments:

61: 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.

99-100: But if DL is used, should it not be transferable if the training data is large enough?

126-130: 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.

135: 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.
143: “or 3D point cloud roughness (Woodget and Austrums, 2017)”... The statement seems repetitive as it has already been mentioned before.

170: “mechanical shutter”... Do you mean global shutter?

180: Please, change GNNS to GNSS.

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

206: Why did the authors not consider p2?

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

223-224: 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.

232-233: 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”.

234-235: After which criteria did the authors choose the model regions? Was the selection performed randomly? And what is low and high confidence referring to?

256-257: “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.

275-277: 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).
Eq. 2: What is the final unit of $\varepsilon_{\text{length}}$? At the current form it seems to be either pixel$^2$ 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?

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., $\sqrt{\text{sum}(\sigma^2)}$?

290: 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.

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?

330-331: 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)?

343-344: “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)?

351: GSD has already explained before.

358-359: 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?
Chapter 4.3 is a mixture of results and discussion and should be split, putting the results into chapter 3.5.

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.

555-557: 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)?

561-566: 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.

References:

- Hastedt, T. Luhmann, H.-J. Przybilla, R. Rofalski (2021): EVALUATION OF INTERIOR ORIENTATION MODELLING FOR CAMERAS WITH ASPHERIC LENSES AND IMAGE PRE-PROCESSING WITH SPECIAL EMPHASIS TO SFM RECONSTRUCTION. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2021 XXIV ISPRS Congress