Reply to RC1

The paper presents many important aspects in order to achieve optimal results in photogrammetric surveys from the planning stage to the final results. These aspects are mostly reviewed from a theoretical point view. The paper ends with two practical examples presenting the work and outcome for two flights at high latitude (~ 80° North).

The paper is a bit of a mixed bag. The first part (approx. 11 pages) summarizes many aspects of modern day photogrammetry, including MTF, diffraction, motion blur, rolling shutter, GNSS specifics, etc. Actually, everything is true for any photogrammetric endeavor -- irrespective of the terrain type (glacier or not).

The 2nd part (also 11 pages) presents the work and outcome for two flights at high latitude (~ 80° North). Following the very detailed description in the first part, one might expect that the practical part then showcases how the theory of the first part is considered in the planning stage of the two flights, and/or that it is demonstrated that neglecting certain aspects from the theory section produces certain errors in the results. But that's not the case. Both flights appear to be executed without having had many options. E.g. what could have been investigated:
- different apertures
- effect of flight speed on blur
- shooting not in raw; or bad Adobe Lightroom settings when "developing" the raw images
- impact of these settings on the different surface types: glacier (which is also part of the title) and non-glacier
- impact of changing the oblique viewing angle
- different flight patterns
- etc.

Of course, the possibilities of changing the parameters and investigations are almost endless, but in the present form the photogrammetric results are obtained in <<one>> way. Apparently, the outcome fulfilled the requirements, but it remains a bit unclear if less strict settings would have let to similar acceptable results. Still, the paper is very well and understandable written. The theoretical part gives a great theoretical summary on many aspects. And the description of the two flights shows the reality of conducting photogrammetric data acquisitions in such high altitude.

We thank the reviewer for their useful comments, and provide detailed responses below. In relation to the comments about the general structure of the paper, it is true that our field measurements were not undertaken in a systematic way to investigate all of the available imaging options. Rather, as stated by the reviewer, the flights were executed without having had many options due to the remote nature of our fieldwork and limited time (particularly helicopter availability) to undertake our measurements. Our aim, therefore, is to demonstrate that real-life surveys can still produce useful results even when undertaken is suboptimal conditions, and for our paper to provide a bridge between purely theoretically-based studies and purely field-based studies, since few other papers have previously been published in this area.

In order to accept the paper I would ask the authors to provide a few more details in the theoretical and practical parts:

row 78 "Along with focal length, sensor size also defines the ground sampling distance (GSD) and therefore"
--> This is not correct, equ (1) for the GSD has no sensor size, just the pixel size (pixel pitch).
Thanks for point out this error. Changed to: “Along with focal length, sensor size and pixel count also define …”

Table 1 "cy per in" should be "cy per mm". Corrected “cy ln^{-1}” to “cy mm^{-1}”.
How is the Nyquist limit calculated? Added a sentence in the table caption:
“The Nyquist limit, defined in units of cycles per mm (cy mm^{-1}), is related to pixel pitch (ρ) by $0.5/\rho$, and is equivalent to one half of the sensor cut-off frequency.”

Fig. 1+2+3 The paper should include the formulae that went into creating these figures.
Added the following text to captions for:

Fig. 1 "Detector and AA filter MTFs are modelled with: $MTF_{det} = sin(\pi\nu\rho)/(\pi\nu\rho)$ and $MTF_{AA} = cos(\pi\nu\rho)$ (Rowlands, 2017)."

Fig. 2 "Lens MTF is modelled with: $MTF_{diff} = \frac{2}{\pi}\left(cos^{-1}(\nu/\nu_c) - (\nu/\nu_c)\sqrt{1-(\nu/\nu_c)^2}\right)$ (Rowlands, 2017)."

Added a reference to equations 1 and 2 in the caption for Fig. 3.

row 148 "the diffraction limit decreases with smaller apertures"
--> "the diffraction limit decreases with smaller apertures (f-numbers)"
ln 150. Added "(large f-numbers)."

row 149 "1/lambad N" should be "1/(lambad*N)".
ln 150. Corrected to "$1/(\lambda N)$" here, and in the caption for Fig. 2.

row 175-180 "more visible as the size of the Airy pattern"
"the distance between the centers of two disks is equal to their radius"
"when the circle of confusion reaches a size of twice the pixel pitch"
--> I would welcome that every time it is clearly specified if you mean diameter or radius (like in the 2nd quote).
Agreed. We replaced "size" with "diameter" throughout the text, and simplified Eq. 2 by:

(1) directly using the Airy diameter (2.44) instead of the radius (1.22×2) and,
(2) replacing the term for aperture diameter (D) by the equivalent expression (f/N).

Equation 2
"GSD", which is not the resolution of the image, is a well defined term (your equation 1).
Thus I object against introducing the term "diffraction limited GSD" in equation 2, because any later mentioning of GSD creates confusion whether you mean really "GSD" or use it as a short for "diffraction limited GSD". And, indeed, in the caption of Fig. 3 you write "Diffraction limited ground sampling distance (GSD)" which is ambiguous in the way that there GSD might refer only to "ground sampling distance" or the whole term "diffraction limited ground sampling distance". In the end of that caption you also mention another term: "theoretical GSD".
Furthermore, later, you refer to equ.2 as circle of confusion (which I think is better). Any term that really refers to the resolution (instead of the sampling = GSD) should by clearly discernable. Some authors use the term "GRD" (ground resolved resolution). Although, that term itself leaves open what effects are considered (just diffraction, or even the full point spread function), it still means obviously something else than GSD.
This is a good point. To make clear the distinction between sampling and resolution, we replaced "diffraction limited GSD" with "GRD" throughout the text, and added the following sentence (lns 177–8):

"In aerial (and satellite) photography, ground resolved distance (GRD) refers to the smallest resolvable detail on an image, given the limitations of the imaging system, including diffraction effects."

row 202 "Wide angle lenses exhibit negative (barrel) distortions which present as decreasing image magnification from the center of the frame towards the edges, while positive (pincushion) distortions are characteristic of telephoto lenses (70 mm or above)."
--> This is opposite to my experience which is:
1. That the sign of the radial distortion is not predictable from the angle of the camera.
2. The longer the focal length the more the lens shows no distortion at all.
Admitted, this is my personal experience, but currently, the quote is without reference. So, please, either provide a reference, or rephrase.
(On row 197 you write "the downside being that short focal lengths are more prone to distortions", if this is inverted to "longer focal lengths are likely to have less distortion", it fits to my experience.)
In general, distortions vary from barrel to pincushion as focal length increases from one extreme to the other, going from wide-angle to telephoto. This is often more evident for zoom lenses which show both types of distortion (e.g., Ray, 2002). Overall, however, the amount of distortion is more pronounced in lenses with short focal lengths and tends to be less visible with long focus lenses. In that sense, we agree with your comment. To clarify this point, we updated the text and added a reference to Ray (2002) (ln 215–8):

"Zoom lenses tend to display more complex distortions and a combination of both types, transitioning from positive to negative with decreasing focal length (Ray, 2002). The amount of distortion corresponds to the difference between the real image and the theoretical (undistorted) one, often reported as a percentage of image height, and is generally less pronounced in long focuses lenses but intensifies with increasingly short focal lengths (Ray, 2002)."

~~Fig 2.~~ Fig. 4
I am not sure, if (a) and (b) are derived from the very same RAW image, or only (a); and (b) is a direct JPG from the camera (and thus in principle a different photograph than the RAW for (a))?
They are versions of the same image, saved in-camera as both RAW and JPEG format. We charged the labels on the figure from "TIFF" to "RAW" to avoid confusion and rephrased the caption:
"Figure 4. Exposure adjustments performed on two versions of a single underexposed image captured in (a) 14-bit RAW, and (b) 8-bit JPEG formats. Both files were saved in-camera and imported into Lightroom for editing. Stronger adjustments are required for the JPEG (b1) to reach a comparable overall level of exposure and retrieve an equivalent amount of information to the RAW image (a1). With more extreme adjustments, the RAW image (a2) remains useable, while on the JPEG (b2) some information is lost in the darker shadows (bright blue pixels) and compression artefacts and false colour (purple patches) combine to degrade image quality. The RAW images were subsequently exported as 16-bit TIFFs for further processing."

row 250 "Including the affinity and non-orthogonality coefficients in the camera calibration matrix at the image alignment stage should partially compensate for this effect."
--> This works only (or the more) the flight speed is constant and the terrain is flat. You may wish to add/clarify that.
Added the following text to clarify (lns 262–3):
"... should partially compensate for this effect, although it is less likely to be effective with large and rapid changes in flight speed, direction, and height above ground."

Remark 1: (Again from my experience) affine parameters need to be introduced per image (and not per camera).
In theory, this would be our assumption as well, especially with large or rapid variations in camera motion. However, when dealing with our two datasets we found that letting these two coefficients vary led to over-fitting, giving unrealistic values for the focal length and principal point coefficients. Ultimately, using the 8-parameter camera model (omitting b1-b2) improved our results.

Remark 2: The developers of Pix4D have a paper about their method on rolling shutter compensation, which is better than the affinity in image space as it directly works on the change of the exterior orientation parameters per image:
https://s3.amazonaws.com/mics.pix4d.com/KB/documents/isprs_rolling_shutter_paper_final_2016.pdf
Thank you for bringing this useful paper to our attention. We have now added more information on the rolling shutter compensation to the text, along with a reference to the Vautherin et al., (2016) paper (lns 263–9):
"Various software, including Pix4D and Agisoft Metashape, have also implemented camera models to compensate for rolling shutter effects, estimating camera motion (translation and rotation) during exposure and modelling external orientation parameters per row of pixels on the sensor (instead of per image) (Vautherin et al., 2016). However, the performance of the correction is also sensitive to survey configuration, showing better results with more regular gridded flight patterns at relatively constant speed, and especially when combining nadir and oblique images. When correcting for rolling shutter, simultaneously solving for the affine distortion parameters has been shown to degrade accuracy due to an overparameterisation of the model (Zhou et al., 2020)."

row 254 "including" --> "included" (?) OK
row 266 "0.43" --> "4.3" OK

row 280 "The direct georeferencing method ... similar precision to the ground-based approach where camera position information is acquired with multi-frequency survey-grade GNSS equipment"
--> Does this last part ("where camera position ...") refer to the ground-based approach?

No, it refers to the direct georeferencing approach. We have moved the last part of the sentence higher up to avoid confusion (lns 297–300):

> "The direct georeferencing method, using airborne control measurements, represents a major logistical advantage for aerial surveys in remote locations as it eliminates the need for a network of GCPs and, where camera position information is acquired with multi-frequency survey-grade GNSS equipment, it has been shown to produce results of similar precision to the ground-based approach."

Remark: A photogrammetric survey that fully relies on direct georeferencing using GNSS, and thus without a single GCP, is prone to deliver results with a large height bias. Because if the camera calibration is considered unknown and thus is estimated during the bundle block adjustment, then any small bias in the estimated focal length causes a large height offset. This is especially true for vertical images, and may be mitigated using oblique images.
As you mention, this is predominantly an issue for nadir datasets, and especially for surveys over relatively flat terrain. Combining oblique images with nadir datasets has been shown to reduce GCP requirements (e.g., Nesbit and Hugenholtz 2019). Fortunately, our datasets are composed of oblique images captured in a convergent geometry, which minimises this issue: in our results, using no GCPs, RMS errors at checkpoints were roughly twice as high in the vertical than horizontal, but all were <0.7 m.

General comment to the theory (section 2): Interestingly, the problem of depth of field is completely neglected. Although, its importance increases with smaller viewing (focus) distance, it belongs into this theory part. Actually, you refer to defocus in row 591. Especially, the hyperfocal distance would be an interesting feature maybe not known to everybody of the target audience.
Thanks for pointing this out. We have added the following text describing the problem of depth of field and hyperfocal distance in section 2.2.2. (lns 203–10):

> "Focal length and aperture also define the hyperfocal distance, corresponding to the focus distance giving the maximum the depth of field (DOF), defined as the zone of acceptable focus. The hyperfocal distance decreases with focal length and aperture, with wide-angle lenses and large f-numbers maximising the DOF. Focusing a lens at infinity places the near edge of the DOF at the hyperfocal distance (Ray, 2002) which, for an effective focal length of 24 mm at f/5.6 is 3.4 m, meaning everything falling any further will be acceptably sharp. Further closing the aperture to f/11 reduces the hyperfocal distance by about half (to 1.7 m) but also impacts system resolving power by increasing diffraction softening. In aerial photography, where the height above ground exceeds the hyperfocal distance and DOF is not a concern, selecting an aperture minimising diffraction and motion blur is preferable."

page 14/15: I am a bit confused regarding your "off-nadir" images. On row 365 you say """>5°" and on row 372 "30-50°". Why use these rather different definition thresholds, and not just provide some information on the off-nadir angles your two sites used; e.g. 5th and 95th percentile, and add that info to table 2. Additionally, it is not clear in which direction the off-nadir angle is applied; as pitch or as roll, or something between?
Yes, this can be confusing. The definition varies between studies, but 5° seems to be a minimum to qualify an image as off-nadir, while 30–50° are more typically used to describe oblique aerial photography (e.g., Nesbit and Hugenholtz 2019), and closer to the angles in our two studies here. With the variable topography, the camera handheld, and without additional attitude information, it is difficult to give more precise estimates of the actual angles.

Following your suggestion, we replaced the general definition ">5° off-nadir" with the approximate angles used in the two surveys, and added a sentence specifying that the camera was pointed roughly 30–50° off-nadir, predominantly in the roll direction (lns 397–9):

> "In both surveys, the camera sensor was oriented with the short edge (vertical) parallel to the direction of aircraft travel (yaw 0°), and the viewing direction roughly orthogonal to the flightpath (pitch 0°), between 30–50° off-nadir to the right at TF (roll >0°) and to the left at ED (roll <0°)."

We also clarified the distinction between simply "oblique" and "convergent" imagery (lns 389–90):

> "… with a convergent image geometry with varying angles oriented around a central area of interest showing the biggest improvement (Sanz-Ablanedo et al., 2020)."

Here it would be super helpful to include the viewing direction in Fig. 5 for every 50th image or so.
In table 2 further some info on the GSD would be interesting to get some idea. I know with oblique angles and variable terrain there is no straight forward way to come up with a representative value, but currently there is no mentioning at all.
To help clarify this, we added a line in Table 2 with a range for the GSD based on the altitude above ground level information in the same table, also specifying the following in the caption:

> "The ground sampling distance (GSD), based on the indicated aircraft altitude a.g.l., represents an upper-bound estimate assuming nadir imagery."

Adding the viewing direction to Fig. 5 would make it very busy (especially the top panel) and harder to read, but we added information on the viewing direction in the caption to help with this:

> "The camera viewing direction was roughly orthogonal to the direction of travel (indicated by an arrow), looking left out of the aircraft in (b), and right in (c)."

row 392 You took pictures through the front passenger window. It would be interesting to know how the camera calibration was effected by that (in comparison to the other data set).
ln 414. Clarified that pictures were taken "out of an open window on the front passenger side." No extra glass was placed in front of the lens that could have affected camera calibration.

row 413 "PPP" is mentioned here the first time, please, add a reference.
ln 437. Added a reference to Kouba and Héroux (2001) and Kouba et al., (2017).

row 443 Here and elsewhere you mention an aperture of "f/5". Although, such value is possible in principle, it would be very unusual, because the usual f-stop numbers are the powers of sqrt(2), the closest being thus 4 and 5.6, which you also have in Fig. 2+3.
That is correct, f/5 (used in the EF survey) is one-third of a stop above the full stop of f/5.6 (used in the TF survey). While f/5 might not be particularly common, we are not aware of a reason not to use one half- or one third-stop increments in photogrammetry. However, this is still a good point, and so we updated ln 676 in the final recommendations from "aperture of f/5.0–5.6 …", to:

> "Use an aperture two to three stops up from the maximum, ideally around f/4.0–5.6 …"

row 456 "at or below the size of the circle resulting from diffraction"
--> you mean "diameter"? Also add a reference to the equation in the theoretical part.
lns 480–1. Replaced "size" with "diameter" and added a reference to Eq. 2 (GRD).

row 478 "Lastly, a variable exposure gain was automatically applied to all images to brighten underexposed areas and match total exposure of successive images."
--> Can your provide a bit more information on how this exposure gain is controlled?
To provide more information on the variable exposure gain, we replaced the above sentence and the next one with the following text (lns 502–7):

> "Lastly, a variable exposure gain was applied to all images to brighten underexposed areas and increase the level of detail and available information for feature extraction. Here, Lightroom automatically adjusts the total exposure (EV) of successive images captured with different in-camera exposure settings (i.e., shutter speed and ISO), to match a selected reference image. This was performed in batches, selecting overlapping images with similar content to that of the reference image, to even out differences in illumination between images and enable a more uniform orthophoto reconstruction."

row 486 "the most time-intensive task in post-production is masking extensive swaths of sky and any terrain beyond the area of interest."
--> What would have happened if you would not have done this masking? Would the bundle block adjustment have completely failed for all images, or just for the affected images? Or would the adjustment have worked, but the

subsequent dense point cloud extraction would have failed (if so, why not simply define a bounding box prior to deriving the dense point cloud)?
Without masking the sky, the bundle adjustment still works, and most of the bad tiepoints can be filtered using gradual selection. The main issue then is the large amount of noise in the dense point cloud. Because on some photos the glacier surface meets the sky (or some distant mountain), defining a bounding box to exclude the background is virtually impossible. The masks are then useful to exclude part of images from the depth maps, and so from the dense cloud.

General comment on the bundle block adjustment in Metashape:
- What accuracies were assumed for the GNSS image positions?
- What GNSS residuals were obtained after the adjustment?
- What reprojection error was obtained?
Since the focus here is on the acquisition and optimisation of source data, not on the photogrammetry processing itself, this detailed information is included in the PhD thesis that this paper is based on (Medrzycka, 2022).

row 575 "Ideally, horizontal accuracy should be higher or equivalent to the spatial resolution of the final gridded products. Here, both DEMs and orthomosaics were gridded at 0.5 m resolution and horizontal checkpoint misalignment errors remain below that level for both reconstructions."
--> I never heard of this rule and, actually, do not subscribe to it. The gridding of the results (DEM from dense image matching and orthomosaic) should fit to the image resolution, i.e. the finest details in the image should also be included in these results, independently of the spatial accuracy of the georeferencing. Even if accuracy and grid width do not match, as in your case, the results provide valuable information about these fine details, however, you only are able to derive the location with a certain limited accuracy. For that reason, the accuracy always should be communicated together with the results.
The gridding was done independently of the accuracy of the results. What we mean here is that ideally the horizontal accuracy would be better than the resolution so as not to degrade the quality of the final product. In this case, final resolution was 0.5 m and RMS errors for the checkpoints are given in the previous sentence (lns 600–1).

In more general terms, when GSD is better constrained, we would aim to minimise the ratio of RMSE/GSD, which is the same as minimising RMSE/DEM and orthophoto resolution since final pixel size should be roughly equivalent to the GSD.

Furthermore, the error in the georeferencing is usually a global one, meaning that your result could be improved simply by shifting in 3D to obtain a much better georeferencing. The latter can be done even on the results themselves; e.g. in case of time series where one epoch (with the best georeferencing quality) serves as reference (and enough corresponding stable areas are present, of course). Thus it would be a waste of potential to set the gridding of the result to the obtained georeferencing accuracy.
Assuming survey-wide systematic error only, simply shifting the reconstruction could indeed improve uniform registration accuracy, but this is not the case with spatially-uncorrelated random error as we had in our study.

Last not least, photogrammetry works even without GCPs and GNSS (one only needs a known distance for scaling). In this case no meaningful (absolute) georeferencing accuracy can be obtained, but the resolution of the images still serves as a guide. As you outlined in the theory the actual resolution of the images is not easily determinable, because it depends on so many factors. However, the GSD is well defined and easily obtainable (at least in case of nadir images over "flat" terrain). So the usual way in photogrammetry is to adopt the GSD as gridding value for the DEM and the orthophoto.
In your case I am not sure what the GSD is, but from the given point density values of around 15 pts/m^2, we see that the average distance between these points is 1/sqrt(15) = 26 cm. Thus you could at least create your DEM and the orthomosaic with 25 cm grid width and should thus be able to get a bit more out of your results than compared to the chosen 50 cm (provided the images were not dramatically effected by blur). I could imagine, that the orthomosaic could be created with an even smaller pixel size, because with dense image matching only in optimal cases one really gets a 3D point per image pixel. Furthermore, in your case, you will have a high variability of the

GSD, and in order to get the details even in the images with the smallest GSD, one thus could base the orthomosaic pixel size not on the average GSD but some smaller statistical value, like the 10th percentile.
We agree with all of the above and we initially did build the DEMs with ~0.25 m grid spacing (based on the average point density relationship you mentioned), but any smaller cell size resulted in patchier reconstructions. Ultimately, 0.5 m was chosen as a compromise between processing time and level of detail required by the specific project. In this case, the final products were compared with much lower resolution reconstructions from historical aerial photographs. Working with the DEM and orthophotos at higher resolution was computationally heavy and provided very little improvement (if any) to the final results.

We updated the text and added more information to clarify this point (lns 601–8):

> "Ideally, horizontal accuracy would be higher or equal to the spatial resolution of the final gridded products which, for surveys with more regular geometry and constant height above ground, should be roughly equivalent to the GSD. In. In this case, where GSD is not easily constrained, point density is useful to define an appropriate pixel size for the gridded products. Here, DEMs and orthomosaics were gridded at 0.5 m, or roughly half the achievable resolution based on the average point spacing of 0.27 m at TF, and 0.22 m at EF. The 0.5 m represents a compromise between processing time and resolution and, in this case, is sufficiently detailed to answer the requirements of the specific project. Horizontal checkpoint misalignment errors remain below the 0.5 m cell size for both reconstructions."

row 665 "Due to data gaps, 28 cameras from the 10 Hz EF survey were disabled (~5 %), compared to 129 cameras (or 13 %) from the 15 s TF survey."
--> Here you use the wrong terminology from Metashape. You mean 28 and 129 "images" not "cameras". Actually, here do you mean that the entire images were disabled, or that the GNSS locations were disabled (due to big interpolation error)?
(Actually, Metashape should be able to link images without GNSS location information to their neighboring images (with GNSS location), provided the image content allows for enough feature points.)
That's correct, Metashape can align images with no valid coordinates. The 28 and 129 images were still used, but the associated camera position estimates were ignored in the bundle adjustment (i.e., disabled). In the original sentence, the term "cameras" comes from the "Reference" tab in Metashape which lists the imported photos in the "Cameras" column, along with the camera position coordinates. In this context, the terminology seems to be correct, although we agree that the wording is confusing, so we changed it to (lns 723–5):

> "Due to data gaps in the GNSS observations, camera position estimates for 28 images (~5%) from the 10 Hz EF survey, and 129 images (or 13%) from the 15 s TF survey, were marked as invalid and were omitted from the SfM workflow."

It would be interesting to list the numbers of images regarding: originally taken vs. disabled images (classified for whatever reasons (e.g. blur)).
We have added more information to Table 2 with the number of used images and those with valid coordinates (also corrected a mistake with the total number of images from the TF survey on lns 405 and 508).

Finally, a general comment: If you cite a book of several hundreds of pages, then please include the page number in the quotes; e.g. Rowlands, 2017.
The norm in the geosciences, and associated style guides, is to only provide a page number when text is directly quoted from a book or paper (e.g., https://blog.apastyle.org/apastyle/2015/03/when-and-how-to-include-page-numbers-in-apa-style-citations.html). We have therefore not included page numbers with our references, but would be happy to do so if the editor agrees with the reviewer's suggestion.