Reviewer 2

This article presents the calibration of the ORCHIDEE model against the MODIS derived snow albedo dataset. While the overall objective of improving albedo is very relevant, this particular study, in my opinion, is very limited. My specific concerns are outlined below.
We thank the reviewer for taking the time to read and comment on the manuscript. We believe addressing these comments will help widen the scope of the paper, especially by adding a section on the impact of different parameters. Although this study is an example over Greenland, the techniques used and model developments will be relevant to other applications. This study helps show how satellite data is vital for model development. We improve the model-data fit through robust Bayesian parameter calibration and identify further model developments needed to capture the GrIS processes fully. In ice sheet modelling, this type of robust Bayesian parameter estimation is rare, and its value in identifying missing processes is not well documented. Furthermore, we use the full area for parameter estimation (compared to using selected pixels), whereas passed examples focus on only a handful of in situ sites.

We have added the following text to the end of the Introduction to emphasise this:

> "Using MODIS snow albedo, in this study, we use DA for parameter estimation to improve the albedo parameterisation inside the ORCHIDEE LSM (Krinner et al., 2015). Instead of using a single or multisite approach which samples the space, here, to exploit the full spatial coverage of the satellite retrievals, we optimise over the whole area of the GrIS to obtain one best set of model parameters applicable over the full ice sheet. *Although this study is only over the GrIS, we can apply the method to other regions. We show how robust Bayesian parameter estimation is an important tool for model development. We further highlight the different limitations and considerations needed to apply such an approach.*"

The article reads like a description of the research in the way it was conducted. The authors describe all the methodologies the authors tried, which are sometimes distracting from the main objective of the paper. For example, Section 3.2. describes the results with two different optimization algorithms. As shown here (and as well known), gradient search methods have limitations in exploring complex decision spaces within an optimization context. The results presented here are not adding anything new to the key focus of this paper, and it is distracting. In Section 2.4.2 – It is not clear (at this point) in the manuscript what is meant by 'performing a sensitivity analysis of the model'. Typically, this is done ahead of the calibration step to reduce the number of parameters being optimized (as the authors acknowledge in Section 3.4.3). If that's the same context, it'll be good to describe that and present this section before 2.4.1. Similarly, Section 3.4.3 should be presented earlier (even though that's not how this work evolved). I appreciate the value of explaining all the steps, but there are lots of 'preliminary' setups (section 3.2, line 207) in this paper. A major recommendation is to restructure the paper so that it focuses on the finalized results, while presenting the intermediate results and steps only to support the main findings.
We thank the reviewer for their suggestion to restructure the paper. We agree that this will help widen the focus of the study. As suggested, we have moved the sensitivity analysis to the beginning of the study - as well as clarified in Section 2.4.2 what is meant by performing an SA.

> "*An SA tests the sensitivity of a model output (usually a physical variable). It tests how the output changes, with respect to different inputs - here the model parameters.*"

The defining edges section, which is needed before the SA, has been moved ahead of the "Performed experiments" section in "Methods and Data". "Defining edges" and "Performed experiments" are now under the umbrella of "Experimental setup". Although we believe that for the non-DA expert, we should keep the analysis of the different methods, this has been moved to the Appendix so as to not distract from the main results. Similarly, we have also moved the edge-only optimisation to the Appendix. Both sections are now under the "Preliminary optimisations" heading. To account for these changes, Sect. 2.5 (Performed experiments) and the Discussion section have been rewritten.

As the authors note in the summary, calibration has its problems in that adjusting certain model parameters may improve some parts of the model, while degrading others. The main objective of improving albedo is to improve the changes in the snow pack over GrIS, as noted in the intro. The paper needs to describe what the impact of the improved snow albedo formulation is on the snow simulations (and other model states). Does the improved albedo lead to better snow states?

Yes, we agree that this information is missing; as such, we have added the following section to discuss the impact on the rest of the snow states:

> "***3.3.3 Impact of the different parameter sets on modelling the surface mass balance of the Greenland Ice Sheet***
>
> *In Fig. 7, we consider how the different parameter sets discussed in this study impact the modelled snow states. To assess the performance of the different ORCHIDEE parameter sets, we compare the model outputs to that of the MAR model. Although MAR is a model with its own biases and errors, it has been shown to have good estimations of the different snow states (*Fettweis et al., 2017, 2020) and so is a good product against which to compare.
>
> *In particular, we are interested in better modelling the surface mass balance (SMB). It measures the difference between mass gains and ablation processes, hence dominating the rates of mass change over the GrIS. Compared to MAR, the manually tuned version of ORCHIDEE performs best at simulating SMB. This can be seen both spatially and temporally. Spatially, the differences between MAR and the ORCHIDEE simulations are observed at the edges - especially in the north and west of the GrIS. The most noticeable difference in the ORCHIDEE runs can be seen at the west of the ice sheet, where the tuned model simulates SMB the best when compared to MAR, followed by the optimised model. In both the manually tuned and optimised models, the SMB is reduced at the west of the ice sheet compared to the default ORCHIDEE model. This is mirrored by an increase in runoff at the west of the ice sheet. Indeed, for simulated runoff, changes are mainly found at the west of the ice sheet, with the tuned model performing the best and the optimised model second when compared to MAR. Both models improve the fit compared to the default ORCHIDEE simulations. However, neither model is able to capture the magnitude of the runoff in summer, with the tuned model still only simulating half the expected magnitude of runoff.*
>
> *When we consider modelled sublimation, we get the most different results. By increasing the albedo over the ice sheet, we decrease latent heat over the area and hence sublimation. When considering the time series, we see that the optimised model gets the correct magnitude of sublimation during the summer months. All of the ORCHIDEE simulations have a delayed peak compared to MAR and no sublimation is simulated by*

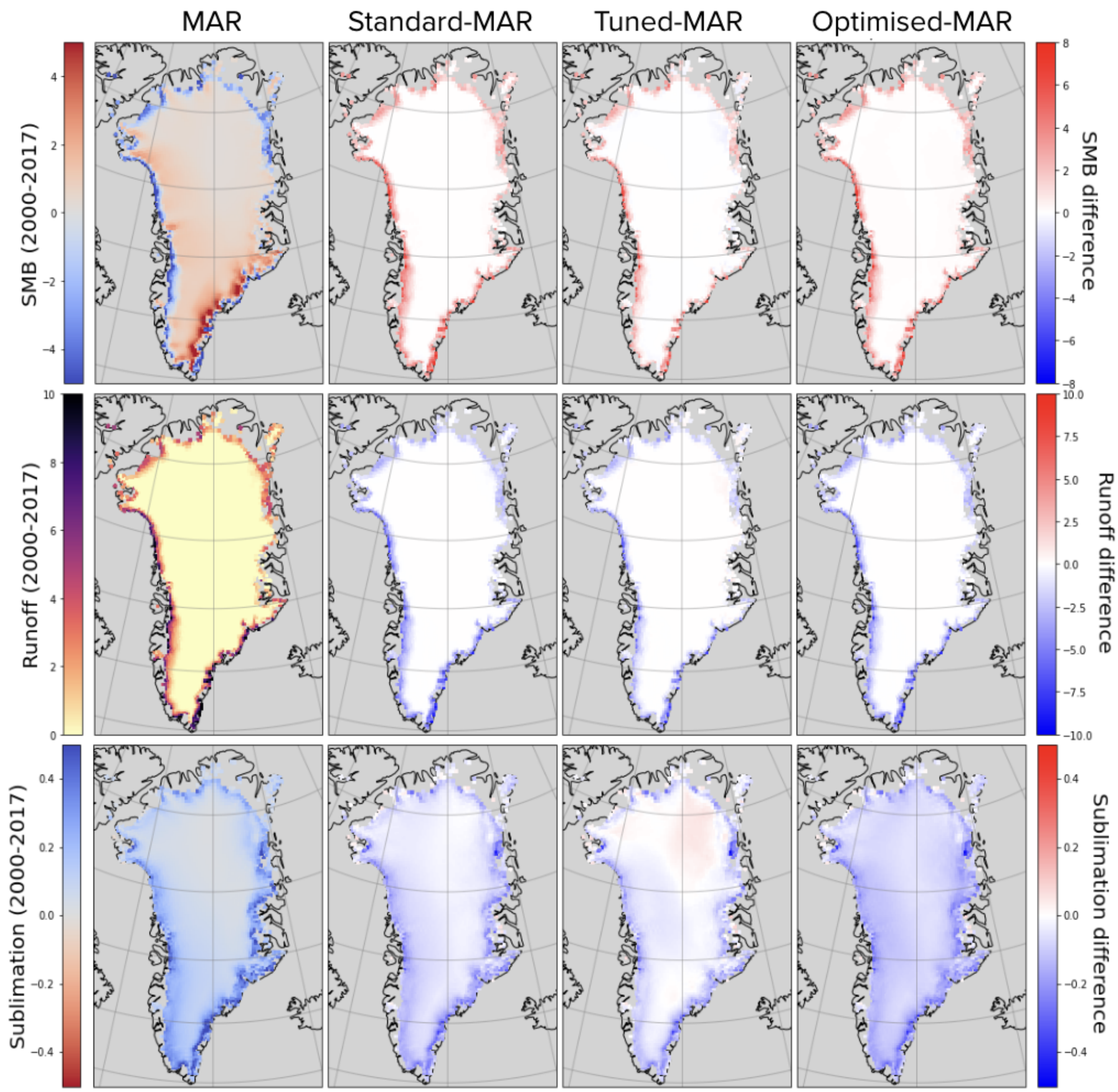
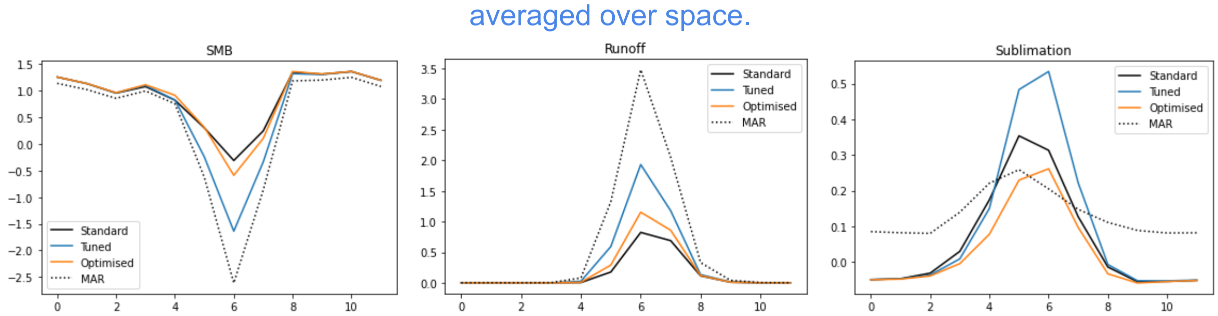


Figure 7: Impact of different parameter sets on ORCHIDEE simulations; “Standard” uses default parameter values, “Tuned” uses parameter values from the manual tuning and “Optimised” from the ORCHIDAS optimisation. The top panels show spatial maps averaged over time (March-October) of MAR (left) and ORCHIDEE-MAR; the bottom panels show monthly means averaged over space.



ORCHIDEE outside the summer months. When averaged over time, we see that MAR has high sublimation rates to the east of the GrIS. However, none of the ORCHIDEE simulations capture this. Instead, the sublimation over the centre of the ice sheet is what changes with the different parameter sets - with the optimised model lowering the rates the most. The strong impact that changing albedo has on simulated sublimation over the whole of the GrIS shows how coupled they are in the model.

Overall, with the optimised model, we do better than the standard ORCHIDEE model but not as well as the tuned model. During the manual tuning of the albedo parameters, the performance of the new parameters was assessed against several model outputs, including SMB, sublimation and runoff at each step of the trial and error procedure. We can think of this manual tuning as a multi-objective calibration. When performing the Bayesian optimisation, we get the best fit to the albedo. However, we overfitted to albedo with no other data, degrading the fit to other model outputs. As seen with the BFGS algorithm and the posterior parameters, parameter space is not smooth but has many local minima. As such, it is possible that a different solution exists, reducing the albedo to a similar extent whilst also improving the fit to other modelled outputs. To achieve this, we need to include more data in the optimisation to perform a multi-objective optimisation. If we cannot find such a parameter set, this would point to structural problems in the model, i.e., missing processes.

We have also added the following to the discussion:

"*We also showed the influence of the parameters on other model outputs by comparing simulated snow states to the MAR model. The optimised model was found to perform better than the original ORCHIDEE model but not as well as the tuned model for simulating SMB and runoff. For sublimation, the optimised model simulated the most accurate magnitude in summer; however, it still showed a bias when considered spatially.*"

The modeling setups use forcing data from MAR, which is a modeled estimate, presumably with its own associated biases and errors. In a calibration setup, the tuned parameters are then used as an error sink to 'hide' these boundary condition errors. This needs to be discussed in the article. Is there an evaluation of MAR data over GrIS? Are other 'observational' datasets available?

Yes, the reviewer is right to highlight this. Although we mention in the Conclusions that using a different meteorological forcing would help separate model structural errors from errors in the forcings, we do not explain that the data assimilation might correct bias in the data instead of errors in the land-surface model. This is made more explicit in the Conclusions:

> "Since we are running ~~the~~ ORCHIDEE offline - i.e., prescribing the meteorological forcing - it would also be beneficial to run the model with different forcings to separate model structural errors from the errors in the forcing. *This is important since MAR is a modelled estimate and, therefore, will be subject to its own biases and errors. We would want to ensure that we are correcting errors in the land surface model and not correcting atmospheric biases in the forcing data.*"

There are two stand-out studies evaluating MAR over GrIS. The first is Fettweis et al. (2017), where different forcings are tested. The second is Fettweis et al. (2020), where MAR was found to be one of the best models for simulating SMB in an intercomparison study. Since we also now use MAR to evaluate the different snow states of the optimised model, we have added these references to that section.

MAR is a regional atmospheric model that uses forcing data to prescribe the atmospheric boundary conditions outside the domain. Here ERA-Interim data is used, the predecessor of ERA5. This has been added to the MAR description in Sect. 2.2:

> The ORCHIDEE model was forced using meteorological outputs from the regional climate model Modèle Atmosphérique Régional (MAR; Gallée and Schayes (1994)) *version 3.11.4. MAR is a regional atmospheric model that uses 6 hourly ERA-Interim reanalyses data from the European Centre for Medium-Range Weather Forecasts (ECMWF, Dee et al., 2011) to prescribe the atmospheric boundary conditions outside the domain.* Outputs from the MAR have a resolution of 20 km *and a 3 hourly time step. In addition to the MAR meteorological outputs, we consider* runoff, sublimation and SMB outputs in this study to assess the impact of the optimisation on these simulated quantities.

We could consider using global reanalysis products to run the model e.g., ERA5. However, as global products instead of regional ones, most of these products have a coarser spatial resolution and are less understood over Greenland. As for observational datasets with which to evaluate the model, we have already listed a few in the conclusion, e.g. GRACE, LST_cci. However, evaluating against these data is outside the scope of this study.

Since ORCHIDEE is used in global setups, how are the results over this domain applicable in a general sense? Are these calibrated parameters limited to GrIS?

Ongoing work is looking at other regions to test the applicability of calibrated parameters. Although this is out of the scope of this paper, we have added the following to the conclusion to highlight this perspective:

> "For additional evaluation, we are testing the application of this model and parameters to other polar and non-polar regions, starting with other ice sheets such as Antarctica."

Minor comments:

Line 51. Need brackets around Krinner et al. (2005).
Done

Line 67: Change to 'the' instead of 'our'?
Done

Section 2.3 – It is important to clarify (here, early on in the paper, abstract, and title) that the snow albedo is being calibrated instead of the total albedo. MODIS has several different albedo products (blue-sky, black-sky etc.) Please clarify.
Yes, we agree with the reviewer that we need to clarify which product is used, given that MODIS has several. We have therefore added "snow albedo" to the title, abstract, introduction etc. as suggested. We have also expanded the MODIS description section (in line with comments from Reviewer 1):

> "In this study, we used satellite-derived *snow* albedo from the NASA (National Aeronautics and Space Administration) MODIS (Moderate-Resolution Imaging Spectroradiometer) MOD10A1 product (Hall et al., 1995)*. This product uses data from the Terra satellite, which has a sun-synchronous, near-polar circular orbit crossing the equator at approximately 10:30 A.M. local time (Hall & Riggs, 2016) and providing global coverage every 1-2 days. MOD10A1 is a clear-sky daily product. When more than one retrieval is available on a given day, which is the case near the poles, the best value is kept. This best value is chosen based on solar elevation, distance from nadir and cell coverage (Hall & Riggs, 2016). In addition, pixels in the MOD10A1 with solar zenith angles greater than 70° are masked (night is defined as a solar zenith angle greater than 85°).*
>
> *The version of MOD10A1 we used in this study was further processed by Box et al. (2017). Using data from* collection 6 of MOD10A1 (Riggs et al., 2015; Hall and Riggs, 2016), Box et al. (2017) de-noised, gap-filled and calibrated the data into a daily 5km grid covering Greenland for the years 2000-2017. *This dataset was further validated against ground-based measurements from the PROMICE stations (Fausto et al., 2021) and the residual bias in the dataset based on the solar zenith angle corrected for using a linear regression according to time and latitude (Box et al., 2017). Finally, in this dataset,* when ~~MODIS retrievals are inaccurate due to~~ *there is* not enough solar illumination *to compute the albedo* during the winter months (January, February, November, and December), *Box et al. (2017)'s distribution swaps in the* April values~~ are swapped in~~.
>
> In this study, we *used the dataset created by Box et al. (2017),* further aggregating these data to the resolution of the ORCHIDEE outputs, imposed by the meteorological forcing files (20 km)."

Line 146: change to 'output' instead of 'writing'
Done

Line 151: remove 'However'

Done

How do the calibrated values influence the peak winter month simulations?
Since the retrievals in the winter months are uncertain, we have decided to remove all analysis of model-data fit to these months.

Figure 1 – this is the snow covered albedo? Is this average computed by excluding Nov-Feb?
This is retrieved snow albedo against simulated albedo. However, this can be ice albedo in areas where the snow has melted. This figure was computed over the whole year. However, this has now been changed to exclude Nov-Feb for consistency with the rest of the manuscript.

Section 3.1 – This is a very hand-wavy section. The authors need to spell out exactly what was changed in this manual calibration procedure. What parameters/physics were changed?
The parameter values in Table 1 were the values resulting from the manual tuning experiments. We have replaced the values with the default ORCHIDEE values, clarifying this in the caption of the Table. In Section 3.1, we explain how the parameters have been changed from these default values. We have added a table to the Appendix showing the resulting tuned values. This table also shows the posterior parameter values from the optimisation for completeness.

Below is the text we have added to Sect 3.1 to explain how the parameters were changed in the tuning experiment:

> "Before using ORCHIDAS to optimise the model parameters, the ORCHIDEE model was first tuned manually *through trial and error*. While not as robust as using a Bayesian framework, this initial step is common for land surface modellers and helps get a sense of the different parameter sensitivities. The primary focus of this manual tuning was to better capture the behaviour of the GrIS at its edges. *This was achieved by increasing the overall albedo of fresh snow (A_{aged} + B_{dec}) and the snowfall depth required to reset the snow age ($\boldsymbol{\delta}_c$), while also decreasing the albedo of aged snow and decreasing the rate of snow age decay ($\boldsymbol{\tau}_{dec}$). Furthermore, one of the tuning constants for glaciated snow-covered areas was decreased (ω). The rest of the parameters were kept as the default ORCHIDEE parameters (see Table B1 for full results).*
>
> This initial tuning helped the model to better simulate the albedo at the edges of the ice sheet, especially in the western part (Fig. 2)*, as well as other snow states such as SMB and runoff, which were also used to assess the success of the manual tuning*."

Section 3.2: How many iterations of GA were used here? Are these the results from the 'Both' approach (results in Figure 2)?
We used 15 iterations for the GA algorithm, which we found to be sufficient for convergence. This information has been added to the "Performed experiments" section (Sect. 2.5):

> "*We found that the genetic algorithm greatly outperformed the BFGS algorithm, reducing the cost function by 11% compared to a negligible reduction, and that 15 iterations of the genetic algorithm were sufficient for convergence.*"

Are these the results from the 'Both' approach (results in Figure 2)?
This is an optimisation over the whole ice sheet without any weighting. This section was about finding the right method. It led us to identify that the middle points dominated the optimisation, necessitating weighted edges in the main optimisation. This is now made explicit in the Preliminary experiments section:

> "*and over the whole of the GrIS (without weighting the edges).*"

Table 2: How are the albedo evaluated for ‘All months’? If you don’t trust the MODIS albedo during the winter months, how do you justify comparing back to them?
We agree that assessing the fit to the winter months is tricky and not precise, given the high uncertainties during these periods. We have removed all analyses using the winter months. This is also in line with the comments from R1.

Line 220: Why were these three years chosen? How do you do these calibrations (separately for each year and somehow harmonize the calibrated parameters? Or are they calibrated from a single run, but the calibration data is withheld during all years except 2000, 2010, and 2012)?
As stated in the “Performed optimisations” section, these years were randomly selected. In section (L159-162), we also explain that the three years were optimised simultaneously. This means that the cost function is a sum of three cost functions, one for each year considered. This approach was chosen to reduce computation costs since even running ORCHIDEE over the GrIS for one year at this resolution takes 20 minutes of computation time on our computing cluster run in parallel over eight cores. We have expanded the text in the “Performed optimisations” to be more explicit:

> “We optimised over these three years simultaneously. *This means that, in this main experiment, we minimised a cost function comprising a sum of three cost functions, one for each year considered.* ~~and~~ The rest of the 2000-2017 time series was used for validation.”

and added to the words “randomly selected”, “simultaneously”, and a reference to Sect. 2.5.2 in L220:

> “For the main optimisation, the GrIS albedo was optimised over the *randomly selected* years 2000, 2010 and 2012 *simultaneously*, with a larger weight given to the edges *(see Sect. 2.5.2 for the full setup description)*.”

Line 223: Add a comma after ‘Indeed’.
Done

Line 232: Why is it that ‘We would not expect to lower the RMSD of the edges any further’?
Since we used the genetic algorithm in the edge-only optimisation, we can say with some confidence that the algorithm converged at a global minimum. We found the best set of parameters given the setup and, more importantly, the lowest cost function value we could achieve. Since the main optimisation included more data to fit, it was possible that the algorithm would reduce the cost function by fitting the other data (i.e., the middle) and not the edges since the middle has three times more data points. We have rephrased the text to be more explicit:

> “However, the RMSD reductions over the edge points are similar in magnitude to the reductions found in the preliminary optimisation where only the edge points were considered (Table A1). This means that the weighting used between the edge and middle points during the optimisation was sufficient - *we have achieved as low RMSD at the edges as in the edge-only experiment*. ~~We would not expect to lower the RMSD of the edges any further.~~”

References

Box, J. E., Van As, D., and Steffen, K.: Greenland, Canadian and Icelandic land-ice albedo grids (2000–2016), GEUS Bulletin, 38, 53–56, 2017

Box, Jason E., 2022, "MODIS Greenland albedo", https://doi.org/10.22008/FK2/6JAQPK, GEUS Dataverse, V1

Hall, D. and Riggs, G.: MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6. Greenland coverage., National Snow and Ice Data Center, NASA Distributed Active Archive Center, Boulder, Colorado USA., http://nsidc.org/data/MOD10A1/versions/6, accessed December 2016., 2016.

Hall, D. K., Riggs, G. A., and Salomonson, V. V.: Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data, Remote sensing of Environment, 54, 127–140, 1995

Fausto, R. S., van As, D., Mankoff, K. D., Vandecrux, B., Citterio, M., Ahlstrøm, A. P., Andersen, S. B., Colgan, W., Karlsson, N. B., 420 Kjeldsen, K. K., et al.: Programme for Monitoring of the Greenland Ice Sheet (PROMICE) automatic weather station data, Earth System Science Data, 13, 3819–3845, 2021.

Fettweis, X., Box, J. E., Agosta, C., Amory, C., Kittel, C., Lang, C., van As, D., Machguth, H., and Gallée, H.: Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model, The Cryosphere, 11, 1015–1033, https://doi.org/10.5194/tc-11-1015-2017, 2017.

Hall, D. K. and G. A. Riggs. 2016. MODIS/Terra Snow Cover Daily L3 Global 500m SIN Grid, Version 6. [Indicate subset used]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10A1.006. [Date Accessed].

Stroeve, J. C., Box, J. E., & Haran, T. (2006). Evaluation of the MODIS (MOD10A1) daily snow albedo product over the Greenland ice sheet. Remote Sensing of Environment, 105(2), 155-171. https://doi.org/10.1016/j.rse.2006.06.009