Author's response

We are grateful to the editor and both the reviewers for their thoughtful comments and suggestions to our manuscript. In the following we give the complete author's response. Page and line numbers refer to locations in the original manuscript. We have included the response to both the reviewers and a revised manuscript. In this revised manuscript, direct changes made in accordance to the suggestions of the reviewers are marked in bold. Following the suggestions of both the reviewers, the entire paper has also been revised for its language and grammar/phrasing. These changes are not marked in bold and are not mentioned explicitly. In addition to all the changes made in response to the comments of the two reviewers, we have made the following notable changes:

P2 L10: We changed this sentence and added a reference to the review paper by Dozier et al. (2016).

P2 L20: We added a reference to the relevant and recently published paper by Kępski et al. (2017).

P7 L1: We changed the reference to the recently submitted discussion paper about the Bayelva climate and soil monitoring station by Boike et al. (2017).

P7 L4 and P9 L21: We added a reference to the relevant study of Bruland et al. (2001).

P13 L12: The lines *"due to faulty Aqua MODIS band 6 detectors and the subsequent use of band 7 in the NDSI causing greater pixel misregistration Salomonson and Appel (2006). If only Aqua MODIS fSCA retrievals are available for a given pixel, these are still used to maximize the temporal coverage of the snow cover depletion."* were removed as the Aqua band 6 misregistration problem has been fixed using the QIR algorithm in the MYD10A1 version 6 product (see Riggs and Hall, 2016).

Figures 4, 5 and 7: We have specified that these show the results from a single run with 100 ensemble members (including multiple iterations for the ES-MDA) and are not the average over multiple runs by adding *"These results are from a single run."* to the respective captions.

The reference to Westermann et al. (2016) was changed from the discussion paper to the final revised paper in Westermann et al. (2017).

We replaced the references to Evensen (1994) with the more recent and comprehensive Evensen (2009).

The section on computation expense (previously Section 5.3) was merged with Section 5.2 (Evaluation of data assimilation schemes) in the discussion.

In Section 5.3 (previously Section 5.4) we added a reference to the study of Peng et al. (2015) in a discussion concerning MODIS representativeness error.

Thank you once again for your time and consideration as well as all the helpful comments and suggestions,
On behalf of all the co-authors,
Kristoffer Aalstad

References

Boike, J., Juszak, I., Lange, S., Chadburn, S., Burke, E., Overduin, P., Roth, K., Ippisch, O., Bornemann, N., Stern, L., Gouttevin, I., Hauber, E., and Westermann, S. (2017). A 20-year record (1998-2017) of permafrost, active layer, and meteorological conditions at a High Arctic permafrost research site (Bayelva, Spitsbergen): an opportunity to validate remote sensing data and land surface, snow and permafrost models. *Earth System Science Data Discussions*, pages 1–86. doi: 10.5194/essd-2017-100.

Bruland, O., Sand, K., and Killingtveit, Å. (2001). Snow distribution at a high arctic site at svalbard. *Hydrology Research*, 32(1):1–12.

Dozier, J., Bair, E. H., and Davis, R. E. (2016). Estimating the spatial distribution of snow water equivalent in the world's mountains. *Wiley Interdisciplinary Reviews: Water*. doi: 10.1002/wat2.1140.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(C5):10,143–10,162. doi: 10.1029/94JC00572.

Evensen, G. (2009). *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03711-5.

Kępski, D., Luks, B., Migała, K., Wawrzyniak, T., Westermann, S., and Wojtuń, B. (2017). Terrestrial Remote Sensing of Snowmelt in a Diverse High-Arctic Tundra Environment Using Time-Lapse Imagery. *Remote Sensing*, 9(733):1–22. doi: 10.3390/rs9070733.

Peng, J., Liu, Q., Wang, L., Liu, Q., Fan, W., Lu, M., and Wen, J. (2015). Characterizing the Pixel Footprint of Satellite Albedo Products Derived from MODIS Reflectance in the Heihe River Basin, China. *Remote Sensing*, 7:6886–6907. doi: 10.3390/rs70606886.

Riggs, G. A. and Hall, D. K. (2016). MODIS Snow Products Collection 6 User Guide. Version 1.0.

Salomonson, V. V. and Appel, I. (2006). Development of the Aqua MODIS NDSI Fractional Snow Cover Algorithm and Validation Results. *IEEE Transactions on Geoscience and Remote Sensing*, 44:1747–1755. doi: 10.1109/TGRS.2006.876029.

Westermann, S., Peter, M., Langer, M., Schwamborn, G., Schirrmeister, L., Etzelmüller, B., and Boike, J. (2016). Transient modeling of the ground thermal conditions using satellite data in the lena river delta, siberia. *The Cryosphere Discussions*, pages 1–37. doi: 10.5194/tc-2016-130.

Westermann, S., Peter, M., Langer, M., Schwamborn, G., Schirrmeister, L., Etzelmüller, B., and Boike, J. (2017). Transient modeling of the ground thermal conditions using satellite data in the lena river delta, siberia. *The Cryosphere*, 11:1441–1463. doi: 10.5194/tc-11-1441-2017.

Reply to Reviewer 1

We are grateful to the reviewer for the thoughtful comments and suggestions to our manuscript. We have compiled a revised version and in the following provide a point-by-point reply to all issues raised. The reviewer's comments appear in bold font and our replies in normal font. Excerpts from and changes to the manuscript are quoted in italics. Page and line numbers refer to positions in the original manuscript.

Ensemble-based assimilation of fractional snow covered area satellite retrievals to estimate snow distribution at a high Arctic site

The Cryosphere, Aalstad et al., 2017

The paper presents a case study in which MODIS (and Sentinel-2) fractional snow cover area retrievals are assimilated into a simple snow model to infer peak snow water equivalent and subgrid snow estimates. The ensemble smoother with multiple data assimilation technique is used. The paper adds interesting new insights to the discipline of snow data assimilation and could be considered for publication, after addressing the topics below:

1) - There is a weird mix of poor sentences in a very weak English language (too much to list, 1 random example: p.8, L5: "In addition, not accounted for. . . contribute to. . ."") and parts with excellent English. My hope is that the excellent English parts are inserted by one of the co-authors and not copied from elsewhere. Please revise the entire paper for its language. Related to this, I would also suggest to revise the title to something like: "Ensemble-base assimilation of satellite-based fractional snow cover area to estimate the snow distribution at Arctic sites" (Arctic is already high-latitude; the sites do not have a particularly high elevation - I have no idea what "high" was referring to)

Following the reviewer's suggestion the entire manuscript has been revised for its language.

According to the reviewer's suggestion, the title has been revised to

"Ensemble-based assimilation of fractional snow covered area satellite retrievals to estimate the snow distribution at Arctic sites"

- Text: nice literature review, well done. The other text is a bit long in general, and has quite some repetitions in the discussion section in particular – please condense where possible. E.g. referring twice to the adaptive version of the ES-MDA, referring twice to the spatially distributed modeling, etc.

The potential for spatially distributed modeling and assimilation is now only mentioned in the outlook (Section 5.4). Furthermore, the adaptive ES-MDA is now first mentioned in Section 5.2 (Evaluation of data assimilation schemes) and once more briefly in the outlook. In addition, we have removed the statements concerning the novelty and basis of our work early in Section 3.3 as this was already mentioned in the introduction. Some necessary repetition between the outlook (Section 5.5) and other parts of the discussion remains, but this is kept as brief as possible.

- what are "patterned gorund features"? A Nordic term?

Thanks for spotting a typo. The term has been changed to

"patterned ground features"

which refers to a phenomenon in periglacial regions where patterns, such as sorted circles, form in the ground material.

2) MODIS, p.13, L.14: how exactly do you average the pixels for each study area? What to you mean by "all the pixels"? Would you include pixels with any (low) cloud fraction? (cloud fraction is given as additional information as a confidence measure for the fSCA retrieval estimates) What is the cloud cover limit? Mention explicitly that you are averaging the satellite data to the 1 km (I suspect), both for MODIS and Sentinel-2. Yet, it is mentioned that the Sentinel-2 data are averaged to the footprint of the snow surveys. . . so perhaps the latter is not right. If the resolution of the Sentinel-2 data, MODIS data and model are different, please explain how you reconcile the space-mismatch. Also mention the 1-km spatial resolution upfront in the modeling for clarity.

For each study site and day we simply take the mean fSCA over all the corresponding MODIS pixels shown in Figure 1. This average is only taken if none of the pixels are flagged as cloudy by the MODIS cloud mask. The MOD10A1 and MYD10A1 version 6 products only accept pixels flagged as cloud free (i.e. either "confident clear", "probably clear" or "uncertain clear") by the MODIS cloud mask (see Riggs and Hall, 2016). Cloud fraction is not given as a confidence measure in these products (Riggs and Hall, 2016). If cloud free pixels are available from both Terra and Aqua, then Terra pixels are chosen. To clarify how clouds are dealt with we have added the following sentence to Section 3.2.1:

'We average over all the pixels for each day and study site (see Figure 1). This average is only taken if cloud free (as determined by the MODIS cloud mask) retrievals are available for each of these pixels."

As stated in the Section 3.2.2, the Sentinel-2 fSCA retrievals are mapped to the approximate footprint of (i.e. the area encompassed by) the snow surveys. The areal extent of the Sentinel-2 retrievals is close to those given in Table 1. As such, the following sentence has been added to Section 3.2.2:

"Therefore, the areal extent of the Sentinel-2 fSCA retrievals closely matches the areas of the corresponding study sites given in Table 1."

Consequently, there is a space-mismatch between the MODIS and Sentinel-2 fSCA retrievals. The latter have lower representativeness error as they provide a better match to the area covered by the snow surveys. The point of the exercise in Section 4.3 is to check whether or not this mismatch has a considerable effect on the assimilation results. Still, the space-mismatch in the MODIS pixels is not huge (Figure 1). Moreover, the space-mismatch is reconciled by the difference in the observation error variance ($RMSE^2$) between the MODIS and Sentinel-2 retrievals, determined based on the field measurements of fSCA, where the Sentinel-2 error is considerably lower.

As for the resolution of the modeling, the following has been added to the end of the modeling section (Section 3.1.2) for clarification:

"The model resolution is defined by the footprint of (area encompassed by) the snow surveys for each site (see Table 1 and Figure 1)."

In addition, the following has been added to the end of the forcing section (Section 3.1.3):

"While the resolution of the downscaled forcing data do not exactly match the model resolution (i.e. the footprint of the snow surveys, Section 3.1.2), the mismatch is small considering the gentle topography of the study sites (Section 2.1)."

3) Ensemble data assimilation:

P15, L4 explicitly name the "parameters" as "perturbation parameters" to avoid confusion with model parameters.

Done.

P16, L8-9: "remaining parameters", "prior parameter ensemble": again refine the word choice here or define "state", "parameter" and "perturbations" more clearly or efficiently upfront, because this sentence is referring to two types of parameters and confusing.

Since peak mean SWE is also treated as a state variable in the SSM, the sentence has been reformulated to

"We emphasize that through the perturbation parameters we effectively perturb the melt rate, precipitation rate and coefficient of variation. By performing a subsequent ensemble integration of the SSM we also get an ensemble of state variables that are consistent with the prior perturbation parameter ensemble."

Where we have also explicitly referred to the parameters that are perturbed as perturbation parameters. In addition, we have changed *"prior ensemble of parameters"* in Table 4 to *"prior ensemble of perturbation parameters"*.

P16, L18: identify here (again) what is in the state and parameter vector. I suspect that there are no "snow model parameters" included anywhere, but it would be good to explicitly mention this. Clarify upfront that perturbation parameters are updated and not state variables or model parameters, as is often the case in other hydrology/cryosphere research.

The section starting from the middle of P16 L16 running to P16 L21 has been modified to

"Let N_e, N_o, N_a, N_s, N_p and N_t denote the number of ensemble members, observations, assimilation cycles, state variables, perturbed parameters and time steps during an annual (September-August) model integration. $\mathbf{X}$ *is the* $(N_s \times N_t) \times N_e$ *matrix containing the ensemble of states (fSCA*$_{n,j}$, $D_{m,n,j}$, $\overline{D}_{n,j}$ *and* $\mu_{n,j}$*) and* $\mathbf{\Theta}$ *is the* $N_p \times N_e$ *matrix containing the ensemble of perturbation parameters listed in Table 4. The* $N_o \times 1$ *observation vector* $\mathbf{y}$ *contains all the fSCA satellite retrievals during the ablation season (Section 3.2),* $\mathbf{Y}$ *is the* $N_o \times N_e$ *matrix containing the ensemble of perturbed fSCA satellite retrievals and* $\widehat{\mathbf{Y}}$ *is the* $N_o \times N_e$ *matrix containing the ensemble of predicted fSCA observations. Additionally,* $\mathbf{H}$ *is the observation operator (mapping from the state space to the observation space) and* $\mathbf{R}$ *is the* $N_o \times N_o$ *observation error covariance matrix which is a diagonal matrix containing the observation error variances (Section 3.2)."*

To emphasize that the model constants (model parameters) in Table 2 are fixed and that only the perturbation parameters are updated in the analysis we have changed the sentence on P18 L9 to

"We emphasize that the analysis step (21) only updates the perturbation parameters and a consistent ensemble of states is found from the subsequent ensemble model integration. The model constants listed in Table 2 remain unchanged by the analysis and the integration.".

In addition, all further mentions of *"parameters"* have been corrected to *"perturbation parameters"*.

Fig.3: replace block with "Update states" $\longrightarrow$ "Propagate state" (I believe that the updating is done on the parameters in the ES-MDA Analysis step)

Done. For consistency, in this Figure the block *"Generate Prior Parameter Ensemble'"* was changed to *"Generate prior perturbation parameter ensemble"*.

4) Have you evaluated the updated parameters (result of Eq. 21) themselves? Do the bias estimates make sense? E.g. can you compare the re-analysis forcings, the bias-corrected ones and in situ observed meteorological data?

Snowfall observations are not available for our field sites (Boike et al., 2017) and are very difficult to conduct in the Arctic due to undercatch (e.g. Førland and Hanssen-Bauer, 2000). For the snowmelt, a field based estimate

was available from Westermann et al. (2009). We have added the following to the end of Section 4.1:

"The posterior bias parameters can be directly evaluated by comparing the bias corrected forcing to field measurements. Due to a lack of snowfall observations (see Boike et al., 2017), an evaluation of the precipitation bias parameter is not possible. However, the melt bias parameter can be evaluated by comparing the snowmelt flux (which is directly proportional to the perturbed melt depth) to field-based values. For June 2008, Westermann et al. (2009) estimate an average snowmelt flux of 27 Wm^{-2} , which compares well to the ES-MDA posterior median estimate (averaged for the same period) of 29 Wm^{-2} , while the prior median estimate is too low with 19 Wm^{-2}."

5) P.20, L17 and Fig 4: peak measurements. . . who or what knows that this measurement is taken at the peak? How can we know for sure that it is a peak measurement, if there is only one data point? Why is that peak measurement always located on May 1st in Fig 4? The peak measurement must be at a different time every year. . .

Only a single survey is available for each site for a given snow season. The times of these surveys (April/May), coincide closely with the peak SWE. We have added a quantification of the associated error using snow depth measurements from a sonic ranger at the Bayelva climate station to Section 2.2 (P8 L7):

"Although the snow surveys coincide closely with peak SWE, some accumulation (ablation) may occur after (before) the surveys. To assess the magnitude of this error source, we used snow depth measurements at the Bayelva station (Boike et al., 2017) to compare the snow depth at the survey dates to the maximum snow depth for each snow season. We found an average relative difference of 8% (maximum: 17%, minimum: 0.3%)."

Figure 4 has also been changed. It now shows the measured peak mean SWE μ (as determined by the snow surveys) over the whole snow season as a horizontal dotted black line. The caption of Figure 4 was changed to

"Time evolution of the prior (red) and ES-MDA ($N_a = 4$, $N_e = 10^2$) posterior (blue) fSCA (first column) and mean SWE ($\overline{D}$; second column); shading: 90th percentile range; solid lines: ensemble median; yellow dots: assimilated MODIS and Sentinel-2 fSCA retrievals; dotted black line: independently observed peak mean SWE (μ) from snow surveys (Section 2.2); x-axis: months. These results are from a single run."

6) Fig 5: these distributions seem not to be cross-masked, i.e. why are the prior and posterior estimates not cross-masked to the times and locations of the in situ observations?

We are not sure what the reviewer means by cross-masked. As discussed above, the times of the snow surveys coincide closely with peak SWE. In Figure 5 we are comparing the prior and posterior estimates of the peak subgrid SWE distributions with the corresponding empirical distribution measured in the field through the snow surveys for each study site and snow season. To clarify this, the caption in Figure 5 has been changed to

"Prior (red), ES-MDA ($N_a = 4$, $N_e = 10^2$) posterior (blue) and the corresponding independently observed (from snow surveys; dashed black) peak subgrid SWE distributions; shaded areas: 90th percentile range; solid lines: ensemble median; markers: mean value. These results are from a single run."

7) Table 4, Figure 7: clarify in the caption which validation data are used, where and during which period.

The caption of Figure 8 (previously Figure 7) was changed to

"Fractional improvement in RMSE through the analysis step (1 being perfect and 0 no effect) as a function of the number of ensemble members for the fSCA, peak mean SWE μ and coefficient of variation χ; top left: particle batch smoother, PBS; top right; ensemble smoother, ES; bottom left; ensemble smoother with multiple data assimilation, ES-MDA; bottom right: FI as a function of assimilation cycles in the ES-MDA. The FI for

$N_e \leq 100$ *are averaged over 100 independent ensemble model integrations. Errors were computed based on comparisons to all the corresponding field measurements presented in Section 2.2."*

The caption of Table 5 (previously Table 4) was changed to

"Summary of evaluation metrics, i.e. bias, RMSE and square correlation coefficient (R^2), for the fSCA, peak SWE (μ) and peak subgrid coefficient of variation (χ). These metrics are based on comparisons to all the field measurements presented in Section 2.2 with the number of observations for the comparisons in brackets next to the corresponding symbols. All the metrics are averaged over 100 independent runs each with 100 ensemble members. The ES-MDA was run with $N_a = 4$ assimilation cycles."

For consistency, the caption in Table 6 (previously Table 5) was changed to

"Summary of evaluation metrics, i.e. fractional improvement in bias and RMSE as well as prior and posterior square correlation coefficient (R^2), using the ES-MDA ($N_e = 10^2$, $N_a = 4$) for peak mean SWE (μ) and coefficient of variation (χ) when assimilating only MODIS as well as assimilating both MODIS and Sentinel-2 observations. These metrics are based on a comparison to all the snow surveys conducted in 2016 (see Table 1) and are averaged over 100 independent runs each with 100 ensemble members."

8) P.23, L6: in situ fSCA retrievals – are these the camera-based data? Please clarify. It would also help to use the same term to refer to validation data throughout the paper – i.e. in situ vs ground-based vs ground truth vs "Field measurements" (section 2.2)

Yes, these are the fSCA field measurements based on data from the automatic camera system, UAV and GPS-based surveys discussed in Section 2.2. To clarify and avoid confusion with the satellite retrievals, the corresponding section of P23 L6 was modified to

"field measurements of fSCA (Section 2.2.)"

For consistency, all mentions of *"in situ"* (with one exception) or *"ground based"* or *"ground truth"* (with one exception) have been changed to *"field measurements"*.

9) P24, 15: "lowest improvements": how confident are you with this statement? With only 8 samples, it is hard to get statistical significance of any sort.

To qualify this statement, the sentence on P24 L15 was changed to

"For all schemes the available validation data suggests that the greatest improvements are achieved for fSCA, followed by peak mean SWE, while by far the lowest improvements are found for the peak subgrid coefficient of variation."

10) Table 5: not sure if this exercise has any value with only 3 observations as validation. . . what is the confidence interval on these metrics?

As the Sentinel-2 mission is relatively new (the first satellite, Sentinel-2A, was launched in June 2015), only 3 snow surveys (from 2016) are available for comparison at our study sites since the start of the mission. As such we have added the following cautionary statement to Section 4.3 (P26 L15):

"We emphasize that this evaluation is based on the only 3 available field measurements of μ and χ in 2016 (from the snow surveys), so that these preliminary results need to be consolidated by future studies with more validation data."

11) Section 4.3: "Effects of observation error and assimilation frequency" ?

The heading has been changed accordingly. The same change was made to the heading of Section 5.3 (previously Section 5.4).

12) Conclusions, P31: "For peak mean SWE. . . lower than in previously. . .": this is an apples to oranges comparison to published work over different regions and different periods and thus invalid; please remove.

Done. The end of the sentence has been removed so that it now reads

"For the peak mean SWE, the ES-MDA features an average RMSE of 0.09 m w.e. compared to field measurements"

Thank you once again for all the helpful comments and suggestions,
On behalf of all the co-authors,
Kristoffer Aalstad

References

Boike, J., Juszak, I., Lange, S., Chadburn, S., Burke, E., Overduin, P., Roth, K., Ippisch, O., Bornemann, N., Stern, L., Gouttevin, I., Hauber, E., and Westermann, S. (2017). A 20-year record (1998-2017) of permafrost, active layer, and meteorological conditions at a High Arctic permafrost research site (Bayelva, Spitsbergen): an opportunity to validate remote sensing data and land surface, snow and permafrost models. *Earth System Science Data Discussions*, pages 1–86. doi: 10.5194/essd-2017-100.

Førland, J. E. and Hanssen-Bauer, I. (2000). Increased Precipitation in the Norwegian Arctic: True or False? *Climatic Change*, 46:485–509. doi: 10.1023/A:1005613304674.

Riggs, G. A. and Hall, D. K. (2016). MODIS Snow Products Collection 6 User Guide. Version 1.0.

Westermann, S., Lüers, J., Langer, M., Piel, K., and Boike, J. (2009). The annual surface energy budget of a high-arctic permafrost site on Svalbard, Norway. *The Cryosphere*, 3:245–263. doi: 10.5194/tc-3-245-2009.

Reply to Reviewer 2

We are grateful to the reviewer for the thoughtful comments and suggestions to our manuscript. We have compiled a revised version and in the following provide a point-by-point reply to all issues raised. The reviewer's comments appear in bold font and our replies in normal font. Excerpts from and changes to the manuscript are quoted in italics. Page and line numbers refer to positions in the original manuscript.

Ensemble-based assimilation of fractional snow covered area satellite retrievals to estimate snow distribution at a high Arctic site

The Cryosphere, Aalstad et al., 2017

This paper shows an analysis of the results of three different assimilation algorithms applied to snow variables (in particular SWE, fSCA and sub-grid coefficient of variation) over an Arctic study site. The assimilation algorithms used are the Ensemble Smoother (ES), Particle Batch Smoother (PBS) and a newly introduced Ensemble Smoother Multiple Data Assimilation (ES-MDA) technique. The results show significant improvements in all evaluation metrics for the ES-MDA technique, matching or improving the results obtained using the ES and PBS. The ES-MDA is more robust as it avoids degeneracy and other problems of the other two techniques however this comes at an expended computational cost.

The paper is well written (albeit it needs more work in terms of grammar/phrasing, I recommend one final review by native-english speaker) and very clear. The methodology section might be improved by including examples of the method using figures. As it is right now is very mathematical, which is fine but reduces the possibility of understanding the workings of the method by other researchers on the field. The literature review is very comprehensive and might be improved if condensed. The paper further illustrates the extreme utility of data assimilation frameworks in the context of snow process estimation and I recommend it for publication after minor revisions.

Following the reviewer's suggestion the entire paper has been revised for its grammar/phrasing.

We agree with the reviewer that our presentation is quite mathematical. In a sense, this is unavoidable given the mathematical nature of data assimilation. With Figure 3 we provide a more schematic overview of the work flow in the methodology. We find that even after the revisions the manuscript is still quite long and thus chose to avoid adding additional figures.

The literature review has been condensed by removing many of the fine details regarding the results of different studies. We also refrained from stating the spatial resolution of the cited studies.

Specific comments:

Page 2

2-3: The amount of smoothing depends on the type of terrain - wouldn't expect this effect to be significant beyond smoothing microtopography (i.e., 1-2 m vertical scale).

We have removed this sentence.

13: Probably only precipitation and wind are space-time variant, topography and vegetation shouldn't be considered as dynamic. Radiation is also space-time variant however the direct component climatology might be relatively invariant every year - though I would expect that for high latitudes this is not necessarily true,

Yes, topography and vegetation are relatively fixed. In clear conditions the direct shortwave component of the

radiation is mainly fixed by solar geometry and the local topography and is as such relatively invariant from year to year in the absence of clouds. Still, clouds are common in the Arctic and along with variations in the surface albedo this can make net radiation a highly dynamic variable. To clarify, the sentences starting on P2 L12 and running up to the start of P2 L15 have been reformulated to

"The primary controls on the distribution and variability of SWE are topography, vegetation, precipitation, wind, radiation and avalanching (Sturm and Wagner, 2010; Clark et al., 2011). While topography and vegetation are relatively fixed in time, the other controls vary strongly over a range of spatiotemporal scales."

Page 4

17: Maybe it is worthy citing Cortés et al. (2016; 2017) for a more direct comparison with PBS metrics derived over similar study regions. Both papers include similar validation data (snow surveys), while Margulis et al is focused on point-data (stations).

We agree that both Cortés et al. (2016) and Cortés and Margulis (2017) are valuable references for probabilistic SWE reconstruction in sparsely instrumented regions and have thus added the following to P4 L19

"Cortés et al. (2016) applied the same PBS framework to construct a 30 year reanalysis of SWE over 6 instrumented basins in the Andes. Cortés and Margulis (2017) subsequently adopted this approach to perform a 31 year SWE reanalysis over the entire extratropical Andes."

Page 5

31: Define undulating.

By undulating we mean gentle topography with small hills. The sentence was changed to

"All sites feature gently undulating topography with small hills and surfaces characterized by patterned ground features, leading to strong differences in snow cover due to wind drift."

32: Typo (ground)

Thanks for spotting this typo. It has been corrected.

Page 9

7: Please clarify what do you mean by this term? Also - clarify what "external" processes are not considered (wind redistribution?)

By this we mean any process occurring inside the snowpack itself. Since refreezing is treated at sub-daily resolution and metamorphism is treated implicitly by the snow albedo parametrization we have reformulated the sentence to

"Many internal snow processes (occurring inside the snowpack), including heat conduction and melt water percolation, are omitted. In addition, several external processes such as sublimation and deposition are ignored."

Wind redistribution is treated implicitly by the probabilistic SDC of Liston (2004) through a non-zero peak coefficient of variation (χ) that accounts for a non-uniform peak subgrid SWE distribution (SSD). The shape of this peak SSD (and thus χ) will be partly controlled by wind redistribution during the accumulation season.

Page 10

26: Is there a range defined for this parameter?

Yes, in Table 4 (previously Table 3). A reference to the table was added to the text:

"Q_0 is a perturbation parameter (see Table 4) that is updated in the assimilation,"

Page 11

10: Is the daily time step a result from aggregating internal hourly calculations?

Yes, the forcing is aggregated from subdaily to daily resolution as discussed in Section 3.1.3.

Page 13

4-5: It would be useful to include a quantification of how many images were available per assimilation season for each site. How were clouds identified and masked out?

A new Table (Table 3) has been added that lists the number of available scenes per melt season for both MODIS and Sentinel-2 for each study site.

For MODIS, clouds were masked out automatically by the MODIS cloud mask (see Riggs and Hall, 2016). The following sentence has been added to Section 3.2.1:

"We average over all the pixels for each day and study site (see Figure 1). This average is only taken if cloud free (as determined by the MODIS cloud mask) retrievals are available for each of these pixels."

For Sentinel-2, clouds were masked out manually in the scene selection as specified in Section 3.2.2.

Page 15

7: Curious if the use of independent multiplicative biases for accumulation and melt would result in inconsistent accumulations? (For example b*M>b*P?)

This is an interesting point. We have added the following to the discussion (Section 5.1):

"An inherent equifinality problem (see Beven, 2006) exists in SWE reconstruction since different perturbation parameter sets can provide similar results. For example, if the prior fSCA melts out earlier than the observations this could be due to the prior precipitation having a negative bias, the prior melt having a positive bias or a combination of these two. The opposite would be true if the prior fSCA melts out too late. It is not possible to resolve this equifinality problem with observations of fSCA alone. A key assumption in deterministic SWE reconstruction is that the melt flux is more constrained than the precipitation so that uncertainty in the melt is ignored (Slater et al., 2013). We perturb both the precipitation and the melt, although the latter is assigned a lower uncertainty (Table 4). Through the assimilation we obtain snowmelts that are consistent with the observed snow cover depletion. The close match of the posterior peak mean SWE estimates to the independent field measurements (Figure 7) suggests that the assimilation yields consistent accumulations and that the inherent equifinality problem is of minor consequence."

8: When you mention constant multiplicative biases - does this mean the bias is unaltered throughout the year?

Yes, this has been clarified in the text where we have changed *"constant multiplicative biases"* to *"constant multiplicative biases (fixed throughout the annual integration)"*.

13: The PBS requires that the ensemble includes the observation, thus if no bias is assigned a priori then the PBS might not be applicable as some degree of bias correction is needed.

Both of the prior bias parameters are modeled as lognormal random variables with unit mean but a non zero variance. So a considerable bias is assigned a priori for some of the ensemble members. We agree that the PBS requires that the prior ensemble encompasses the assimilated observations and we view this as a weakness. For example in 2008 for Bayelva the prior fSCA ensemble is positively biased and does does not encompass the observations, so the PBS performs poorer than the ES-based schemes. We could have changed the prior mean of the bias parameters for this particular year but decided not to. In the application of Bayesian data assimilation the prior should always be set without knowledge of the observations that are considered in the likelihood otherwise it is by definition not a prior. See also the reply to the comment concerning Page 28 L3-4.

Page 20

6: The reduction in spread is a direct consequence of any assimilation algorithm, it would be more useful to assess if the constrain in uncertainty of the posterior is consistent with the observations (i.e., are you underestimating uncertainty after assimilation?)

We agree with the reviewer. As such, the following paragraph was added to the end of Section 4.1:

"In ensemble-based data assimilation the spread of the posterior ensemble should represent the uncertainty. To verify this one can compare two metrics: the residual, i.e. the instantaneous posterior RMSE of the ensemble relative to the corresponding independent field measurement, and the ensemble standard deviation (e.g. Evensen, 2009). For this comparison we define the relative residual as the ratio of the residual to the standard deviation. Ideally this ratio should have a value of 1 which indicates that the two metrics are equal so that the posterior ensemble spread accurately captures the estimation uncertainty. For the fSCA, peak mean SWE and peak subgrid coefficient of variation the average (over all available field measurements) relative residuals were 2.22, 1.53 and 1.66 respectively, so the posterior ensemble underestimates the uncertainty. This effect has been extensively described by Evensen (2009), it arises in part because of model structural errors related to neglected physical processes (Section 3.1). Still, the assimilation is generally able to simultaneously (but not to the same extent) reduce the spread and the error in the ensemble (Figure 4). "

Page 21

A scatterplot would be useful to compare the posterior results for all methods. Including the stats is correct but scatterplot allows for more context.

We agree and have included a scatter plot for a single run of all three schemes and the prior. This is included for orientation as a new Figure 6. The following line was included at the end of P23 L10:

"The scatter plots in Figure 6 visualize the performance of the prior and all the considered DA schemes relative to the field measurements."

Page 26

Table 5: A perfect correlation of 1.0 was obtained? Would be useful to have the scatterplots in order to inform the reader with more details on the results.

This perfect correlation is based on a comparisson to just 3 observations. Because of the low number of observations a scatter plot is not informative and would unnecessarily add to the length of the manuscript. We have added the following cautionary statement regarding the limited number of observations used in this evaluation to Section 4.3 (P26 L9):

"We emphasize that this evaluation is based on the only 3 available field measurements of μ and χ in 2016 (from the snow surveys), so that these preliminary results need to be consolidated by future studies with more validation data."

Page 27

32: It is very difficult to compare RMSE across studies due to the differences in methodology/data. I would stick to the comparison performed within the paper as it allows for more controlled conditions.

We agree and have removed this comparison.

Page 28

3-4: More than biased, if the prior ensemble doesn't cover the observations then the PBS would be unable to replicate the observation. Bias in the ensemble per se is not a problem for the PBS. The comparison between PBS/ES from previous papers with the current method is not as straightforward, asides from the obvious differences in regions there are differences in validation methodology and particularly in the number of fSCA measurements assimilated. Landsat fSCA assimilation results in 10-15 observations per year, while MODIS probably results in an order of magnitude greater.

To clarify, we have changed this sentence to:

"Thus, if the prior ensemble is so biased that it does not encompass the observations, the PBS is incapable of correcting the posterior towards the observations outside the bounds of the prior."

For the next part of the comment we assume that the reviewer is referring to the sensitivity analysis around the middle of P28 L18. All other comparisons to the results of previous probabilistic reconstruction schemes have been removed. Here we are just comparing the relative performance of the ES to the PBS. Of course the locations and assimilated data sets are different, with MODIS definitely having a higher temporal coverage. Still, it is positive to see that the results achieved from previous studies that the PBS generally outperforms the ES matches our own findings and so we do not see why this should not be included. We do not say that our studies are the same but simply that the results agree. The same applies to the study of Emerick and Reynolds (2013), in a completely different field, we still expect the same kind of relative performance for the data assimilation schemes in a sensitivity analysis with a non-linear model which is indeed what we find.

Thank you once again for all the helpful comments and suggestions,
On behalf of all the co-authors,
Kristoffer Aalstad

References

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320:18–36. doi: 10.1016/j.jhydrol.2005.07.007.

Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., Kerr, T., Hreinsson, E. O., and Woods, R. A. (2011). Representing spatial variability of snow water equivalent in hydrologic and land-surface models: A review. *Water Resources Research*, 47:1–23. doi: 10.1029/2011WR010745.

Cortés, G., Girotto, M., and Margulis, S. (2016). Snow process estimation over the extratropical Andes using a data assimilation framework integrating MERRA data and Landsat imagery. *Water Resources Research*, 52:2582–2600. doi: 10.1002/2015WR018376.

Cortés, G. and Margulis, S. (2017). Impacts of El Niño and La Niña on interannual snow accumulation in the Andes: Results from a high-resolution 31 year reanalysis. *Geophysical Research Letters*, 44:6859–6867. doi: 10.1002/2017GL073826.

Emerick, A. A. and Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15. doi: 10.1016/j.cageo.2012.03.011.

Evensen, G. (2009). *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-642-03711-5.

Liston, G. E. (2004). Representing Subgrid Snow Cover Heterogeneities in Regional and Global Models. *Journal of Climate*, 17(6):1381–1397. doi: 10.1175/1520-0442(2004)017<1381:RSSCHI>2.0.CO;2.

Riggs, G. A. and Hall, D. K. (2016). MODIS Snow Products Collection 6 User Guide. Version 1.0.

Slater, A. G., Barrett, A. P., Clark, M. P., Lundquist, J. D., and Raleigh, M. S. (2013). Uncertainty in seasonal snow reconstruction: Relative impacts of model forcing and image availability. *Advances in Water Resources*, 55:165–177. doi: 10.1016/j.advwatres.2012.07.006.

Sturm, M. and Wagner, A. M. (2010). Using repeated patterns in snow distribution modeling: An Arctic Example. *Water Resources Research*, 46:1–15. doi: 10.1029/2010WR009434.

Ensemble-based assimilation of fractional snow covered area satellite retrievals to estimate the snow distribution at Arctic sites

Kristoffer Aalstad[1], Sebastian Westermann[1], Thomas Vikhamar Schuler[1], Julia Boike[2], and Laurent Bertino[3]

[1]Department of Geosciences, University of Oslo, P.O. Box 1047, Blindern, 0316 Oslo, Norway
[2]Alfred Wegener Institute Helmholtz Center for Polar and Marine Research, Telegrafenberg A43, 14473 Potsdam, Germany
[3]Nansen Environmental and Remote Sensing Center, Thormøhlensgate 47, Bergen 5006, Norway

Correspondence to: Kristoffer Aalstad (kristoffer.aalstad@geo.uio.no)

Abstract. With its high albedo, low thermal conductivity and large water storing capacity, snow strongly modulates the surface energy and water balance which makes it a critical factor in mid to high-latitude and mountain environments. However, estimating the snow water equivalent (SWE) is challenging in remote sensing applications already at medium spatial resolutions of $1\,\mathrm{km}$. Here, we present an ensemble-based data assimilation framework to estimate peak SWE distributions at the $1\,\mathrm{km}$ scale. The scheme assimilates fractional snow covered area (fSCA) retrievals from remote sensors in a simple simple snow model forced by statistically downscaled reanalysis data. The basic idea is to relate the timing of the snow cover depletion (accessible from satellite products) to pre-melt SWE, while at the same time obtaining the subgrid scale distribution. Subgrid SWE is assumed to be lognormally distributed, which can be translated to a modeled time series of fSCA through the snow model. Assimilation of satellite-derived fSCA facilitates the estimation of the average SWE and coefficient of variation, while taking into account uncertainties in both the model and the assimilated data sets. As an extension to previous studies, our method makes use of the novel (to snow data assimilation) ensemble smoother with multiple data assimilation (ES-MDA) scheme combined with analytical Gaussian anamorphosis to assimilate time series of MODIS and Sentinel-2 fSCA retrievals. The scheme is applied to Arctic sites near Ny Ålesund (79°N, Svalbard, Norway) where field measurements of fSCA and SWE distributions are available. The method is able to successfully recover accurate estimates of peak subgrid SWE distributions on most of the occasions considered. Through the ES-MDA assimilation, the root mean squared error (RMSE) for the fSCA, peak mean SWE and peak subgrid coefficient of variation is improved by around 75%, 60% and 20% respectively when compared to the prior, yielding RMSEs of 0.01, $0.09\,\mathrm{m}$ water equivalent (w.e.) and 0.13 respectively. The ES-MDA either outperforms or at least nearly matches the performance of other ensemble-based batch smoother schemes with regards to various evaluation metrics. Given the modularity of the method, it could prove valuable for a range of satellite-era hydrometeorological reanalyses.

1 Introduction

The spatiotemporal distribution of the seasonal snow cover is a key control on the terrestrial surface energy and water balance in mid to high latitude regions and mountainous areas (Boike et al., 2003; Barnett et al., 2005). With its high albedo and large water holding capacity, snow is a modulator of the global radiation balance and hydrological cycle, making it one of the drivers

of the atmospheric circulation and the associated climate (Andreadis and Lettenmaier, 2006; Liston, 1999). Since the snow cover can exhibit considerable variability over small distances (Clark et al., 2011), mapping the snow water equivalent (SWE) distribution remains a difficult task (Dozier et al., 2016).

The primary controls on the distribution and variability of SWE are topography, vegetation, precipitation, wind, radiation and avalanching (Sturm and Wagner, 2010; Clark et al., 2011). While topography and vegetation are relatively fixed in time, the other controls vary strongly over a range of spatiotemporal scales. In unforested regions, snow tends to be affected by wind drift (e.g. Gisnås et al., 2014) leading to accumulation in areas with preferential deposition, such as topographic depressions or the lee-side of a ridge. The scale of such features can vary dramatically across the landscape. Nonetheless, the processes occurring at a given site are often consistent from year to year and so the SWE distribution is often quite similar to the climatological snow distribution pattern (Sturm and Wagner, 2010; Kępski et al., 2017). Manual measurement surveys are usually impractical for mapping SWE over large areas given their limited support, large spacing and small extent (Blöschl, 1999). Instead, modeling and remote sensing can be employed to map SWE.

Snow models range in complexity from relatively simple single layer models, such as the Utah Energy Balance model (UEB; Tarboton and Luce, 1996; You et al., 2014), to detailed multilayer snowpack models such as Crocus (Vionnet et al., 2012) and SNOWPACK (Bartelt and Lehning, 2002). Some snow models (e.g. ALPINE3D; Lehning et al., 2006) can also be run in distributed mode to simulate the snow distribution over large areas. The accuracy of the model results is limited by the hydrometeorological forcing data, be it from reanalyses or local measurements, whose errors are typically the major source of uncertainty in snow modeling (De Lannoy et al., 2010; Raleigh et al., 2015). In addition, snow models are generally developed as point scale models and even if they are run as distributed models the grid scale values predicted by the model may not be representative of the corresponding process scale (Blöschl, 1999). For example, if a snow model is forced by near point scale hydrometeorological measurements, the model results will only be representative for a grid cell if that particular point is representative of the mean conditions within the grid cell. To circumvent this problem, probabilistic snow depletion curve (SDC) parametrizations have been developed (Liston, 1999; Luce and Tarboton, 2004; Liston, 2004) in which a probability distribution function is assigned to the SWE within a grid cell at peak accumulation. Assuming uniform melt across the grid cell, this allows for a direct relationship between the mean SWE, melt depth and the fractional snow covered area (fSCA) of the grid cell. Liston (2004) used such a SDC parametrization in conjunction with land cover specific subgrid coefficients of variation of SWE with the ClimRAMS model to map the fSCA over North America. As a result, the total snow-covered area increased considerably compared to the control run. Aas et al. (2017) used a tiling approach to represent subgrid snow variability in the WRF model coupled to the Noah land surface scheme over southern Norway. The tiling reduced the cold bias in the modeled near-surface air temperatures and greatly improved the match to the observed fSCA evolution. Nevertheless, due to the inherently large uncertainties in the forcing, modeling alone is usually not an accurate tool for mapping SWE. Instead, models need to be combined with relevant data from remote sensing.

Snow related data sets can be acquired from a variety of remote sensing platforms with near global coverage. The Gravity Recovery and Climate Experiment (GRACE) twin satellites allow for the retrieval of terrestrial water storage (TWS), from which SWE can be recovered at around 100 km spatial resolution (e.g. Niu et al., 2007). Passive microwave (PM) satellite

sensors can retrieve SWE based on brightness temperature at a resolution of around 25 km. However, PM SWE retrievals have problems over forested areas and complex topography, as well as for wet and deep snowpacks (Foster et al., 2005). Both gravimetric and PM sensors are able to retrieve SWE independent of cloud coverage resulting in gap free time series. While not capable of measuring SWE, moderate resolution optical sensors such as MODIS can retrieve binary information on snow cover (i.e. snow or no snow), fSCA and snow grain size (Hall et al., 2002; Salomonson and Appel, 2004; Painter et al., 2009) at approximately 500 m resolution with a daily revisit frequency. In addition, higher resolution optical sensors, e.g. on board the Landsat and Sentinel-2 satellites, can map fSCA at around 30 m resolution (e.g. Cortés et al., 2014). Optical sensors can not see through clouds which results in data gaps over most snow covered regions. To obtain gap free time series, it is thus necessary to either interpolate optical remote sensing data in time and space, or ingest them in models.

Data assimilation (DA) methods can objectively fuse uncertain information from observations and models. Deterministic SWE reconstruction techniques (Girotto et al., 2014b) that directly insert remotely sensed fSCA data in models represent the simplest form of snow data assimilation. Such schemes back-calculate peak SWE from the disappearance date of the snow cover (as determined from fSCA retrievals) using snowmelt models. Martinec and Rango (1981) used Landsat fSCA retrievals during the melt season in conjunction with a simple degree day snowmelt model to estimate the peak mean SWE. Similarly, Cline et al. (1998) used Landsat fSCA retrievals combined with a distributed energy balance model to reconstruct the SWE distribution. More recently, Molotch and Margulis (2008) used fSCA information from multiple sensors for deterministic SWE reconstruction. Durand et al. (2008) introduced a probabilistic framework for SWE reconstruction. This was based on assimilating synthetic fSCA retrievals during the ablation into the SSiB3 land surface model coupled to the SDC of Liston (2004) using the ensemble smoother (ES; Van Leeuwen and Evensen, 1996) in batch mode (c.f. Dunne and Entekhabi, 2005). The assimilation of synthetic fSCA in this twin experiment was used to correct annual biases in the snowfall and facilitated the recovery of the SWE distribution. Using the Durand et al. (2008) framework, Girotto et al. (2014b) assimilated Landsat fSCA retrievals to recover the SWE distribution, yielding a significant reduction in RMSE relative to deterministic SWE reconstruction. Subsequently, Girotto et al. (2014a) used the same framework to perform a 27 year reanalysis of SWE distributions. Recently, Margulis et al. (2015) modified this probabilistic approach by adopting a particle batch smoother (PBS) as opposed to the ES for the assimilation of fSCA retrievals to estimate the SWE distribution. The PBS was found to outperform the ES, considerably reducing the RMSE. Based on this work, Margulis et al. (2016) adopted the PBS framework to conduct a 30 year reanalysis of SWE over the Sierra Nevada (USA) using Landsat fSCA retrievals. **Cortés et al. (2016) applied the same PBS framework to construct a 30 year reanalysis of SWE over 6 instrumented basins in the Andes. Cortés and Margulis (2017) subsequently adopted this approach to perform a 31 year SWE reanalysis over the entire extratropical Andes.**

Several other snow DA techniques have recently been employed. Andreadis and Lettenmaier (2006) assimilated MODIS fSCA retrievals into the VIC model through the ensemble Kalman filter (EnKF; Evensen, 2009) using a simple SDC for the SWE-fSCA inversion. However, the improvement compared to the open loop (OL, i.e. no DA) run was only modest which was also found in similar studies (Clark et al., 2006; Slater and Clark, 2006). A Bayesian technique was used by Kolberg and Gottschalk (2006) to assimilate Landsat fSCA retrievals into a snow model with a probabilistic SDC to estimate the peak SWE distribution. They found a significant reduction in uncertainty when retrievals were assimilated simultaneously as opposed to

sequentially. At the continental scale, a multisensor assimilation of both GRACE TWS and MODIS fSCA using the ES and EnKF for TWS and fSCA respectively yielded significant improvements relative to the OL (Su et al., 2010). De Lannoy et al. (2010) used the EnKF in a twin experiment to assimilate synthetic PM SWE retrievals and greatly outperformed the OL. This was extended to a real multisensor experiment by jointly assimilating PM SWE and MODIS fSCA retrievals (De Lannoy et al., 2012). Li et al. (2017) used the ES to assimilate PM SWE retrievals and estimate the SWE distribution, markedly outperforming the OL. Of late, particle filter (PF; see Van Leeuwen, 2009) schemes have been gaining popularity in snow DA studies (Charrois et al., 2016; Magnusson et al., 2017). For example, Charrois et al. (2016) assimilated synthetic optical reflectance retrievals into Crocus using the sequential importance re-sampling PF at a point scale and considerably outperformed the OL.

It is worth emphasizing that the most popular schemes in the snow DA community, both the EnKF and the PF, are filters (i.e. sequential techniques). As such, they are Markovian of order 1 (memoryless): the future state at a given point in time depends only on the present state. Furthermore, observations are assimilated sequentially with only the current observation affecting the current state. Batch smoothers (Dunne and Entekhabi, 2005), on the other hand, take into account the entire history of a model trajectory within a batch (observation window) and as such have memory (non-Markovian) so that they are better suited for reanalysis problems.

In this study, we build on the probabilistic SWE reconstruction technique outlined in Girotto et al. (2014b) to recover subgrid SWE distributions (SSD) for a study area in the Arctic based on fSCA retrievals from MODIS and Sentinel-2. The novelty of our study lies in the use of an iterative batch smoother scheme, namely the ensemble smoother with multiple data assimilation (ES-MDA; Emerick and Reynolds, 2013). To update physically bounded parameters we make use of analytical Gaussian anamorphosis (Bertino et al., 2003). We investigate the performance of the ES-MDA in terms of SWE reconstruction and compare it to the ES and the PBS employed by Girotto et al. (2014b) and Margulis et al. (2015) respectively. The results are evaluated against independent field measurements of fSCA and snow surveys conducted over six snow seasons.

2 Study area

2.1 Physical characteristics and climate

The study area is located in NW Svalbard close to the research town of Ny-Ålesund ($78°55'$N, $11°50'$E) on the Brøgger peninsula. Field measurements are available from three sites (Figure 1). "Bayelva" about two kilometers west of Ny-Ålesund is the main study site where multiyear in situ records on e.g. the surface energy balance, permafrost thermal regime and snow distribution are available (Westermann et al., 2009; Gisnås et al., 2014; Boike et al., 2017). In addition, snow surveys for a single season (2016) are available from "Steinflåen plateau" and "Kvadehuksletta". **All sites feature gently undulating topography with small hills and surfaces characterized by patterned ground features, leading to strong differences in snow cover due to wind drift.** Bayelva and Kvadehuksletta are located between 10 and 50 $\mathrm{m\,a.s.l.}$, while the Steinflåen plateau is at a higher elevation of around 200 $\mathrm{m\,a.s.l.}$. Kvadehuksletta is exposed to most wind directions, whereas Bayelva and Steinflåen plateau are partly sheltered by mountains. The sites are located within the continuous permafrost zone (Boike et al., 2003) with a maximum active layer depth of around 1.5 m at the Bayelva site (Westermann et al., 2009).

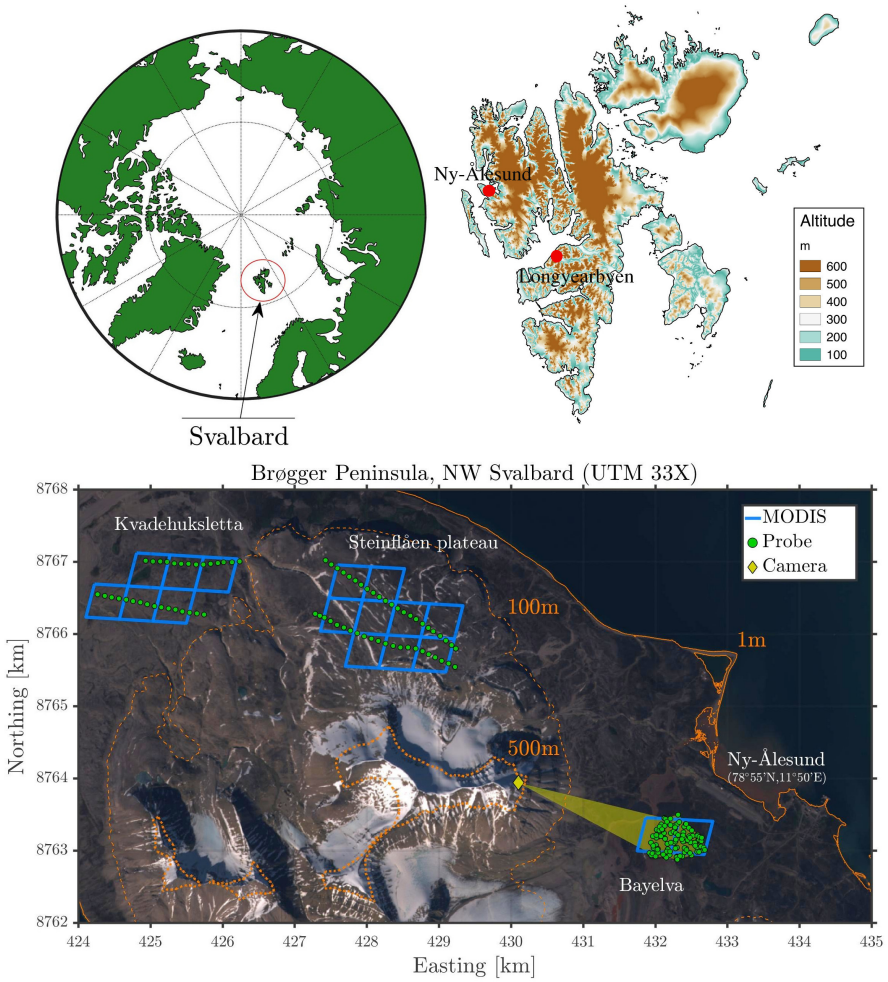

Figure 1. Top left: location of Svalbard (in red) in the Arctic. Top right: map of Svalbard, the study area is close to Ny-Ålesund. Bottom: Sentinel-2A true color image (taken 02.07.2016) of the western Brøgger Peninsula with the three study sites Kvadehuksletta, Steinflåen plateau and Bayelva; green dots: snow survey probe locations; blue polygons: MODIS pixels; yellow diamond: automatic camera system on Scheteligfjellet; yellow shading: field of view of the camera; contour lines courtesy of the NPI (2014) DEM.

The Bayelva site is located around the heavily instrumented Bayelva climate and soil monitoring station (Boike et al., 2017). This area has been the subject of extensive studies spanning permafrost (Roth and Boike, 2001; Boike et al., 2008; Westermann et al., 2011a), the surface energy balance (Boike et al., 2003; Westermann et al., 2009), CO_2 exchange (Lüers et al., 2014; Cannone et al., 2016), ecology (Kohler and Aanes, 2004), snow (Bruland et al., 2001; Gisnås et al., 2014; López-Moreno et al., 2016), hydrology (Nowak and Hodson, 2013) and satellite retrieval validation (Westermann et al., 2011b, 2012). The surface cover at Bayelva and Kvadehuksletta alternates between bare soil, rocks and sparse low vegetation (Westermann et al., 2009), while the more elevated Steinflåen plateau is predominantly covered by loose rocks.

The climate of western Svalbard is influenced by the relatively warm West Spitsbergen ocean current causing a maritime climate with mild winters and cool summers for this latitude (Esau et al., 2012). At Ny-Ålesund the winter, summer and annual (1981-2010) average air temperatures were -12.0 °C, 3.8 °C and -5.2 °C, respectively, while the average annual precipitation was 427 mm (Førland et al., 2011). Between September/October and May the precipitation mainly falls as snow, although rain on snow events have become more frequent due to the warming of the local climate (Nowak and Hodson, 2013; López-Moreno et al., 2016). The seasonal snow cover usually forms in late September or early October and lasts until mid June to early July, with a melt season of around one month (Winther et al., 2002). The dominant energy source during the snowmelt is radiation (longwave and shortwave), while the heat flux required to warm the frozen ground underlying the snow is an important energy sink (Boike et al., 2003; Westermann et al., 2009).

2.2 Field measurements

Manual surveys of snow depth and density were carried out in April/May for six years at the Bayelva site and for one year (2016) at the two other sites (Table 1). At this time, the snow depth is near its maximum but the snowpack is still dry. The snow density was sampled in vertical layers at every fifth point. As no systematic stratification of the snow density was found, SWE was finally calculated from snow depth and the average snow density at each site in a given year. At Bayelva, the snow density was generally confined to a range of $350 \pm 50\,\mathrm{kg\,m^{-3}}$ for all the surveys, while the snow density was found to be around $450\,\mathrm{kg\,m^{-3}}$ at Steinflåen plateau in 2016. At Kvadehuksletta and Steinflåen plateau, the surveys were conducted along transects with regular sample intervals (see Figure 1). A randomized array of sample points was employed for Bayelva in most years, except for the first two years where transects were used.

Location	$\bar{z}$ [m a.s.l.]	σ_z [m]	Area [km^2]	Survey years	Samples per survey
Bayelva	23	9	0.5	2008, 2009, 2013-2016	853^t, 617^t, 105^r
Steinflåen plateau	210	11	1.1	2016	45^t
Kvadehuksletta	55	6	0.9	2016	30^t

Table 1. Overview of the study sites and snow surveys. $\bar{z}$ is the mean elevation and σ_z is the standard deviation in the elevation both based on the NPI (2014) DEM. Key: t=transect, r=randomized array.

Basal ice layers resulting from rain on snow events (Kohler and Aanes, 2004; Westermann et al., 2011a) occur in the area and can constitute a major source of uncertainty for SWE measurements. In 2016, the depth of basal ice layers was measured using ice screws and their contribution to the SWE was accounted for. In addition, internal ice layers and the spatial variability of average snow densities (see above) contribute to the uncertainty of the measurements. Furthermore, only a limited number of sampling points is available, so that the obtained snow distributions are expected to deviate to a certain extent from the true snow distributions in the area. **Although the snow surveys coincide closely with peak SWE, some accumulation (ablation) may occur after (before) the surveys. To assess the magnitude of this error source, we used snow depth measurements**

at the Bayelva station (Boike et al., 2017) to compare the snow depth at the survey dates to the maximum snow depth for each snow season. We found an average relative difference of 8% (maximum: 17%, minimum: 0.3%).

In 2012, 2013 and 2016, an automatic time-lapse camera was deployed near the summit of Scheteligfjellet (694 m a.s.l.; c.f. Figure 1) overlooking the Bayelva site. The camera was a standard digital camera triggered by a Harbortronics time-lapse system, delivering daily images except for prolonged periods with low cloud cover. The raw camera images were orthorectified at a 1 m resolution and snow was mapped for each pixel using a threshold on the intensity, so that fSCA could be determined for each image. The orthorectified images for two of the years are freely available in Westermann et al. (2015a).

In 2008, aerial images were obtained for the Bayelva site for five dates in June during the beginning of the snowmelt period. This was accomplished by mounting a digital camera to a UAV flying at an altitude of 100 to 250 m above ground which took between 700 and 1000 images per mission at nadir angles. As the images were taken in a near random fashion over the entire area, fSCA was calculated by averaging over the fSCA determined for each image using a simple threshold criterion. GPS-based surveys of the remaining snow patches were available for five additional dates, so that a complete fSCA time series is available for the snowmelt period in 2008.

3 Method

3.1 Simple snow model

To efficiently run a large number of ensemble members, a simple snow model (SSM) is employed, which computes snowmelt rates according to surface energy balance formulations (as in the CryoGrid 3 ground thermal model; Westermann et al., 2016). The model is a blend of a single layer mass balance scheme, based on the UEB model (Tarboton and Luce, 1996; You et al., 2014), and the Liston (2004) SDC. **Many internal snow processes (occurring inside the snowpack), including heat conduction and melt water percolation, are omitted. In addition, several external processes such as sublimation and deposition are ignored**. In the following sections, we describe the governing equations of the SSM (see Table 2 for the model constants).

3.1.1 Snow depletion curve

We use the SDC parametrizations discussed in Liston (1999), Luce and Tarboton (2004) and Liston (2004) which parametrize the relationship between fSCA, melt depth and SWE by using a probability density function (pdf) to represent the peak subgrid SWE distribution (SSD). A key assumption is that the melt rate is spatially uniform within each grid cell. The relationship between the accumulated melt depth (D_m), the peak SSD pdf (f_P), and the fSCA within the grid cell at time t is given by

$$\mathrm{fSCA}(t) = \int_{D_m(t)}^{\infty} f_P(D)\, dD\,. \tag{1}$$

Similarly, the mean SWE depth is given by

$$\overline{D}(t) = \int_{D_m(t)}^{\infty} (D - D_m(t))\, f_P(D)\, dD\,. \tag{2}$$

Following Liston (2004) we parametrize the peak SSD using a two parameter lognormal pdf $f_P = f_P(D|\mu,\chi)$ where μ is the peak mean SWE and $\chi = \sigma/\mu$ is the peak subgrid coefficient of variation (σ is the standard deviation). Our choice of parametric distribution was motivated by independent measurements of the SSD which fit reasonably well to a lognormal distribution (Bruland et al., 2001). Equations (1) and (2) can both be solved analytically as presented in Liston (2004).

3.1.2 Mass and Energy Balance

To obtain the instantaneous net accumulation rate, $\mathcal{A}(t)$, we follow the UEB model through (You et al., 2014)

$$\mathcal{A}(t) = P(t) - M(t)\,, \tag{3}$$

where $P(t)$ is the precipitation rate and $M(t)$ is the melt rate. Sublimation is not considered as it is a relatively small contribution to the mass balance at our study area (Westermann et al., 2009). We use a linear transition to delineate between snowfall and rainfall (You et al., 2014), with thresholds given in Table 2. We only consider rainfall as a positive contribution to the mass balance during non-melting conditions when the rainwater generally refreezes in the snowpack (Westermann et al., 2011a). For melting conditions (where $D_m > 0$), we assume that rainfall directly becomes runoff.

The melt rate, M, is calculated based on a simplified snow energy balance defined by

$$Q_M(t) = Q_R^*(t) + Q_P(t) - Q_H(t) - Q_E(t) - Q_G(t)\,, \tag{4}$$

where Q_M is the snowmelt flux, Q_R^* is the global radiation, Q_P is the heat advected by precipitation, Q_H is the sensible heat flux, Q_E is the latent heat flux and Q_G is the ground heat flux. The last three fluxes are defined as positive when directed away from the surface and vice versa for the first two on the right hand side of (4). The SSM differs from UEB in that we calculate the surface energy balance for a melting snowpack, i.e. isothermal at 0°C, at all times. In this case, the global radiation is

$$Q_R^* = (1 - \alpha_S)\, S^{\downarrow} + L^{\downarrow} - \varepsilon_S \sigma_{SB} T_0^4\,, \tag{5}$$

in which $S^{\downarrow}$ and $L^{\downarrow}$ are the downwelling shortwave and longwave irradiances and the last term is the upwelling longwave radiation for the assumed isothermal snowpack at T_0 =273.15 K. The snow albedo (α_S) is parametrized prognostically through the continuous reset formulation following Dutra et al. (2010), which computes the albedo for time increments Δt by distinguishing between accumulating, steady and ablating conditions:

$$\alpha_S(t+\Delta t) = \begin{cases} \alpha_S(t) + \min\left(1, \mathcal{A}(t)\Delta t/\tau_S\right)\left(\alpha_{\max} - \alpha_S(t)\right)\,, & \mathcal{A}(t) > 0\,, \\ \max\left(\alpha_S(t) - \tau_A \Delta t, \alpha_{\min}\right)\,, & \mathcal{A}(t) = 0\,, \\ \left(\alpha_S(t) - \alpha_{\min}\right)\exp\left(-\tau_F \Delta t\right) + \alpha_{\min}\,, & \mathcal{A}(t) < 0\,. \end{cases} \tag{6}$$

Here, $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum snow albedo values respectively, while τ_A and τ_F are aging (decay) rates for non-melting and melting snow respectively. τ_S is a threshold for daily snowfall which if exceeded leads to a reset of the snow albedo to its maximum value. This simple decay and reset type of snow albedo parametrization has been shown to perform reasonably well at Bayelva (Pedersen and Winther, 2005). The heat advected by rainfall (Q_P) is computed as in Tarboton and Luce (1996), while the turbulent fluxes of sensible (Q_H) and latent (Q_E) heat are evaluated following Westermann et al. (2016). The ground heat flux (Q_G) is parametrized through a simple e-folding relationship during the melting period, i.e.

$$Q_G = Q_0 \exp\left(d_H t_m / z_E^2\right) , \tag{7}$$

where Q_0 is the initial ground heat flux, d_H is the thermal diffusivity of the ground, z_E is the effective depth of the heat transfer below the base of the snowpack and t_m is the number of days with melting conditions after peak accumulation. Q_0 **is a perturbation parameter (see Table 4) that is updated in the assimilation,** d_H is selected according to field measurements (Westermann et al., 2009) and z_E is set so that the ground heat flux decays to near zero a month into the melt season.

The snowmelt flux Q_M can now be evaluated through (4). We recall that an isothermal snowpack at $0^\circ\mathrm{C}$ is assumed for (3) and (4), which is only justified for a melting snowpack. In this case, positive Q_M correspond to melting and SWE reduction, while negative values correspond to refreezing of melt water and thus SWE increase. For a dry snowpack (as is generally the case before the snowmelt), negative Q_M values would lead to a cooling of the snowpack, which is not considered in this simple snowmelt scheme. To discard unphysical values (negative melt rates) we only consider days with net melting conditions, i.e. positive daily average snowmelt fluxes. Thus, the daily averaged melt rate M_n at day n (lasting from t_n to t_{n+1}) is given by

$$M_n = \max\left(\frac{1}{\rho_w L_f \Delta t} \int_{t_n}^{t_{n+1}} Q_M(t)\, dt, 0\right) , \tag{8}$$

where ρ_w is the density of fresh water, L_f is the latent heat of fusion and Δt is the daily time step. We emphasize that the effects of refreezing are still considered at a subdaily time resolution in (8). Similarly, the daily averaged precipitation rate is

$$P_n = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} P(t)\, dt . \tag{9}$$

Now the daily averaged net accumulation rate can be obtained through

$$\mathcal{A}_n = P_n - M_n , \tag{10}$$

and the accumulated melt depth D_m is accounted for through

$$D_{m,n+1} = \max\left(D_{m,n} - \mathcal{A}_n \Delta t, 0\right) H\left(\mu\right) . \tag{11}$$

The peak mean SWE μ is updated via

$$\Lambda = \mu_n + \max\left(\mathcal{A}_n \Delta t - D_{m,n+1}, 0\right) , \tag{12}$$

through

$$\mu_{n+1} = \Lambda H\left(\Lambda - \tau_S\right), \tag{13}$$

where the alternative Heaviside function is defined through $H(x) = 0$ if $x \leq 0$, and $H(x) = 1$ otherwise. Consequently, in (13) the peak mean SWE μ is only non-zero if Λ exceeds the threshold τ_S. Note that the formulation in (11) gradually resets the melt depth towards zero in the case of snowfall after the onset of melt, following Liston (2004). This means that fSCA is not reset to unity in the case of new snowfall after a melting period unless the new snowfall leads to an increase in the peak SWE. In the study area, snowfall events occurred rarely during the snowmelt period and the new snow cover lasted only a short time. At sites where such events are more frequent, Durand et al. (2008) presents an alternative solution, albeit at an increased computational cost. The annual model integrations start in the beginning of September when the surface is assumed to be snow free so that both μ and D_m are initialized as zero. Both μ and D_m are reset to zero following the complete disappearance of the snowpack, defined as when the fSCA decreases below 0.01 to account for the infinite tail of f_P. **The model resolution is defined by the footprint of (area encompassed by) the snow surveys for each site (see Table 1 and Figure 1).**

Symbol	Name	Value	Units (SI)	Reference
α_{max}	Maximum snow albedo	0.85	$-$	Dutra et al. (2010)
τ_S	Threshold snowfall	0.01	$\mathrm{m\,(w.e.)}$	Dutra et al. (2010)
τ_F	Aging constant for melting snow	2.78×10^{-8}	s^{-1}	Dutra et al. (2010)
τ_A	Aging constant for non-melting snow	9.26×10^{-8}	s^{-1}	Dutra et al. (2010)
T_R	Threshold temperature for rain	276.15	K	You et al. (2014)
T_S	Threshold temperature for snow	272.15	K	You et al. (2014)
ε_S	Emissivity of snow	0.99	$-$	Westermann et al. (2016)
d_H	Thermal diffusivity of the ground	6×10^{-7}	$\mathrm{m^2\,s^{-1}}$	Westermann et al. (2009)
z_E	Effective transfer depth	1	m	-
Δt	Daily time step	86400	s	-
L_f	Specific latent heat of fusion	3.35×10^5	$\mathrm{J\,kg^{-1}}$	Tarboton and Luce (1996)
ρ_w	Density of fresh liquid water	10^3	$\mathrm{kg\,m^{-3}}$	Tarboton and Luce (1996)
σ_{SB}	Stefan-Boltzmann constant	5.67×10^{-8}	$\mathrm{W\,m^{-2}\,K^{-4}}$	Tarboton and Luce (1996)

Table 2. List of the model constants used in the simple snow model runs along with the corresponding reference studies.

3.1.3 Forcing

Forcing terms in the form of precipitation, air temperature, relative humidity and wind speed, as well as downwelling longwave and shortwave radiation are required to diagnose the mass and energy balance in (3) and (4). These terms are obtained by downscaling ERA-Interim reanalysis data (Dee et al., 2011) at 0.75° resolution following Østby et al. (2017). This method

uses the linear theory of orographic precipitation in Smith and Barstad (2004) to downscale precipitation and a modification of the TopoSCALE approach (Fiddes and Gruber, 2014) for the remaining fields. The reanalysis forcing is downscaled onto 1 km resolution DEM grid cells centered on each of the study sites. The downscaling is performed based on the mean physiographic conditions (elevation, slope and aspect) within each of these grid cells. The resulting values at 1 km spatial and 6 hourly temporal resolution are linearly interpolated in time to facilitate a stable computation of the time evolution of turbulent energy fluxes following Westermann et al. (2016). From these fluxes, and the remaining surface energy balance fields, diurnally averaged melt rates are calculated. Similarly, diurnally averaged rainfall and snowfall rates are computed by delineating between rain and snow in the time interpolated precipitation rate (You et al., 2014) and then taking diurnal averages. **While the resolution of the downscaled forcing data do not exactly match the model resolution (i.e. the footprint of the snow surveys, Section 3.1.2), the mismatch is small considering the gentle topography of the study sites (Section 2.1).**

3.2 Satellite retrievals

We make use of satellite retrievals between May and September which contain the snowmelt period for all the investigated years. Only retrievals that fall inside the melt season are assimilated as these contain information about the snow cover depletion. Due to frequent cloud cover, the effective revisit frequency of fSCA retrievals is irregular, with prolonged data gaps occurring regularly. An overview of the number of available scenes is given in Table 3.

Location	Melt season	# of MODIS scenes	# of Sentinel 2 scenes
Bayelva	2008, 2009, 2012, 2013, 2014, 2015, 2016	8, 9, 8, 9, 6, 14, 11	−, −, −, −, −, −, 7
Steinflåen plateau	2016	5	8
Kvadehuksletta	2016	11	7

Table 3. Number of MODIS and Sentinel-2 scenes per melt season with field measurements available for the three study sites.

3.2.1 MODIS

We employ version 6 of the level 3 daily 500 m resolution fSCA retrievals from the Moderate Resolution Imaging Spectroradiometer (MODIS) on board the satellites Terra (MOD10A1 product; Hall and Riggs, 2016a) and Aqua (MYD10A1 product; Hall and Riggs, 2016b). The retrieval algorithm is based on an linear fit of the normalized difference snow index (NDSI) measured by MODIS to fSCA retrievals from ground truth Landsat scenes as described in Salomonson and Appel (2004). The NDSI exploits the fact that snow is highly reflective in the visible, but a good absorber in the shortwave infrared which sets it apart from other natural surfaces such as clouds, vegetation and soil (Painter et al., 2009).

We average over all the pixels for each day and study site (see Figure 1). This average is only taken if cloud free (as determined by the MODIS cloud mask) retrievals are available for each of these pixels. If both Terra and Aqua retrievals are available for a given day only the former are used. Despite small deviations in the measurement footprint (see Figure 1),

we compare MODIS fSCA retrievals to the field measurements of fSCA obtained from the automatic camera system, UAV and GPS surveys (Section 2.2). From this comparison, we estimate a RMSE of $\sigma_{\mathrm{MOD}} = 0.13$ for the MODIS fSCA. We use σ^2_{MOD} as the observation error variance in the corresponding diagonal entries of the observation error covariance matrix (Section 3.3.2).

3.2.2 Sentinel-2

For the year 2016, we complement the MODIS fSCA retrievals with aggregated $20\,\mathrm{m}$ resolution retrievals from the Sentinel-2A mission (Drusch et al., 2012). fSCA estimates are derived from the level 1C orthorectified top of the atmosphere reflectance product, with cloud-free scenes manually selected. For this purpose, NDSI is computed from reflectances (r) from a visible (b3, centered on $0.56\,\mu$m) and a shortwave infrared band (b11, centered on $1.61\,\mu$m) through

$$\mathrm{NDSI}_{\mathrm{S2}} = \frac{r_{\mathrm{b3}} - r_{\mathrm{b11}}}{r_{\mathrm{b3}} + r_{\mathrm{b11}}}\,. \tag{14}$$

Each pixel is then either classified as snow covered (NDSI≥ 0.4) or snow free (NDSI< 0.4), where the NDSI threshold was chosen in line with Hall et al. (2002). The binary (snow/no snow) pixels are then aggregated to the approximate footprint of the independent snow surveys conducted at each site (Figure 1) to obtain Sentinel-2 derived fSCA estimates. **Therefore, the areal extent of the Sentinel-2 fSCA retrievals closely matches the areas of the corresponding study sites given in Table 1.** The retrieval process is illustrated schematically in Figure 2. By comparing the Sentinel-2 retrievals to the field measurements of fSCA from the automatic camera system in 2016 we estimate a RMSE of $\sigma_{\mathrm{S2}} = 0.09$. We use σ^2_{S2} as the observation error variance in the corresponding diagonal entries of the observation error covariance matrix (Section 3.3.2).

3.3 Ensemble Data Assimilation

In this section we outline how the prior ensemble of model realizations is set up and how it is updated to a posterior ensemble through the assimilation of fSCA satellite retrievals using ensemble-based batch smoother schemes.

3.3.1 Ensemble Generation

The prior ensemble of model realizations is generated by independently drawing perturbation parameter values from the distributions listed in Table 4. These perturbation parameters are held constant throughout integration of the model for each water year. Two of these are multiplicative bias parameters that perturb the mass balance through the net accumulation rate

$$\mathcal{A}_{n,j} = b_{P,j} P_n - b_{M,j} M_{n,j}\,,$$

for $j \in 1 : N_e$ where N_e is the number of ensemble members. We inherently assume the model forcing to be the major source of uncertainty (De Lannoy et al., 2010; Raleigh et al., 2015). Furthermore, we assume that the error in the forcing can be modeled through **constant multiplicative biases (fixed throughout the annual integration)** in the mass balance. Consequently, the bias parameters are modeled as a positive definite lognormal random variables. This is in line with the perturbations in Girotto et al. (2014b) on the precipitation rate, but we also perturb the melt rate. Moreover, we assume that the ensemble of net accumulation rates is on average unbiased due to the applied downscaling method (Østby et al., 2017) and thus assign the two

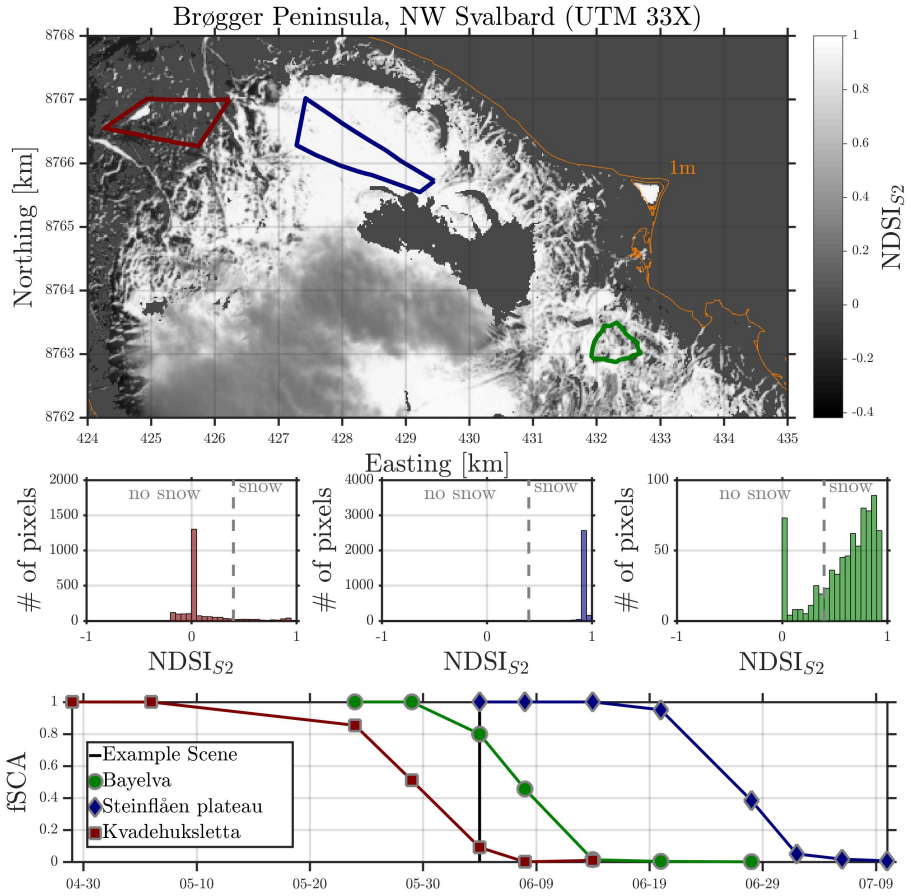

Figure 2. Top panel: Sentinel-2 NDSI estimates from an example scene (taken 04.06.2016) over the Brøgger Peninsula with Kvadehuksletta, Steinflåen plateau and Bayelva marked with red, green and blue polygons; coastline in orange. Middle panels: NDSI histograms of the same example scene (Kvadehuksletta: left, Steinflåen plateau: middle; Bayelva: right) with the threshold at NDSI$= 0.4$ marked. Bottom panel: Time series of Sentinel-2 NDSI-based fSCA retrievals for the 2016 melt season.

bias parameters a mean of unity. The precipitation rates are also perturbed by the same bias parameter in the computation of the heat advected by precipitation (Q_P) in the surface energy balance that contributes to the melt rate M_n.

In addition to the mass balance forcing, the peak subgrid coefficient of variation χ (Section 3.1.1) is a source of uncertainty. We assume a prior mean of 0.4 for χ, which corresponds to the value provided by Liston (2004) for "Arctic tundra". Moreover, χ is assumed to be double bounded between 0 and 0.8, with negative values being unphysical and the upper bound close to the maximum value in Liston (2004). Furthermore, both the initial ground heat flux at the onset of melt (Q_0) and the minimum snow albedo (α_{min}) are uncertain and we also assume that these are double bounded.

Symbol	Name	Distribution	Support	Mean	Variance	Units
χ	Coefficient of variation	Logit-normal	$(0, 0.8)$	0.4	0.01	-
Q_0	Initial ground heat flux	Logit-normal	$(0, 40)$	20	20	Wm^{-2}
α_{min}	Minimum snow albedo	Logit-normal	$(0.45, 0.55)$	0.5	0.02	-
b_P	Precipitation bias	Lognormal	$(0, \infty)$	1	0.04	-
b_M	Melt bias	Lognormal	$(0, \infty)$	1	0.01	-

Table 4. Overview of the distributions from which the prior ensemble of perturbation parameters are independently drawn.

The probability distributions of double bounded random variables are modeled as logit-normal distributions, with the logit transform for a variable x bounded between a and b given by

$$\widetilde{x} = \text{logit}_{(a,b)}(x) = \ln\left(\frac{x-a}{b-a}\right) - \ln\left(1 - \frac{x-a}{b-a}\right), \tag{15}$$

while the inverse transform is given by

$$x = \text{logit}^{-1}_{(a,b)}(\widetilde{x}) = a + (b-a) \,/\, \left(1 + e^{-\widetilde{x}}\right). \tag{16}$$

To generate a prior ensemble of a logit-normally distributed random variable, we first apply the logit transform to the mean. Then, N_e ensemble members of Gaussian white noise with a consistent variance are added and finally the inverse transform is applied. **We emphasize that through the perturbation parameters we effectively perturb the melt rate, precipitation rate and coefficient of variation. By performing a subsequent ensemble integration of the SSM we also get an ensemble of state variables that are consistent with the prior perturbation parameter ensemble.**

3.3.2 Batch Smoothers

Here, we describe the ES-MDA, the main batch smoother scheme used in the assimilation, as well as the ES and the PBS that are used for comparison. In a batch smoother all the observations, in this case all fSCA retrievals from the snow cover depletion during one melt season, are assimilated at once in a single batch (Dunne and Entekhabi, 2005), as opposed to sequentially as in a filter (Bertino et al., 2003). We follow the conventional notation in the DA literature, as laid out in Ide et al. (1997). **Let N_e, N_o, N_a, N_s, N_p and N_t denote the number of ensemble members, observations, assimilation cycles, state variables, perturbed parameters and time steps during an annual (September-August) model integration. $\mathbf{X}$ is the $(N_s \times N_t) \times N_e$ matrix containing the ensemble of states (fSCA$_{n,j}$, $D_{m,n,j}$, $\overline{D}_{n,j}$ and $\mu_{n,j}$) and $\mathbf{\Theta}$ is the $N_p \times N_e$ matrix containing the ensemble of perturbation parameters listed in Table 4. The $N_o \times 1$ observation vector $\mathbf{y}$ contains all the fSCA satellite retrievals during the ablation season (Section 3.2), $\mathbf{Y}$ is the $N_o \times N_e$ matrix containing the ensemble of perturbed fSCA satellite retrievals and $\widehat{\mathbf{Y}}$ is the $N_o \times N_e$ matrix containing the ensemble of predicted fSCA observations. Additionally, $\mathbf{H}$ is the observation operator (mapping from the state space to the observation space) and $\mathbf{R}$ is the $N_o \times N_o$ observation error covariance matrix which is a diagonal matrix containing the observation error variances (Section 3.2).**

The ensemble smoother with multiple data assimilation (ES-MDA Emerick and Reynolds, 2013) is an iterative scheme, requiring multiple ensemble model integrations and analysis steps. Collecting the perturbed and predicted observations during the ensemble integration into a batch and performing the analysis step is referred to as one assimilation cycle and we will let the current iteration number be denoted as ℓ. In such a case, the ES-MDA scheme is set up as follows, for $\ell \in 0 : N_a$ iterations:

1. Run an ensemble model integration, i.e. for $n \in 0 : (N_t - 1)$ time steps compute

$$\mathbf{X}_{n+1}^{(\ell)} = \mathcal{M}\left(\mathbf{X}_{n}^{(\ell)}, \mathbf{\Theta}^{(\ell)}\right), \tag{17}$$

where $\mathcal{M}$ is the SSM operator defined through equations (1), (2), (11) and (13).

2. If $\ell < N_a$ (*otherwise stop the algorithm here*) collect the batch of predicted observations

$$\widehat{\mathbf{Y}}^{(\ell)} = \mathbf{H}\mathbf{X}^{(\ell)}, \tag{18}$$

and perturbed observations

$$\mathbf{Y}^{(\ell)} = \mathbf{y} \otimes \mathbf{1}^{T} + \sqrt{\alpha^{(\ell)}}\mathbf{R}^{1/2}\boldsymbol{\epsilon}^{(\ell)}, \tag{19}$$

where $\otimes$ is the outer product, $\mathbf{1}$ is an $N_o \times 1$ vector of ones, the $^{\mathrm{T}}$ superscript denotes the transpose, $\alpha^{(\ell)}$ is the observation error inflation coefficient and $\boldsymbol{\epsilon}^{(\ell)}$ is a $N_o \times N_e$ matrix containing zero mean Gaussian white noise with a variance of 1.

3. Transform the perturbation parameters using analytical Gaussian anamorphosis functions ψ (Bertino et al., 2003)

$$\widetilde{\mathbf{\Theta}}^{(\ell)} = \psi\left(\mathbf{\Theta}^{(\ell)}\right), \tag{20}$$

ψ is the natural logarithm and the logit for the biases and the remaining perturbation parameters respectively.

4. Perform the Kalman-like analysis step in the transformed space

$$\widetilde{\mathbf{\Theta}}^{(\ell+1)} = \widetilde{\mathbf{\Theta}}^{(\ell)} + \mathcal{C}_{\widetilde{\mathbf{\Theta}}\widehat{\mathbf{Y}}}^{(\ell)}\left(\mathcal{C}_{\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}}^{(\ell)} + \alpha^{(\ell)}\mathbf{R}\right)^{-1}\left(\mathbf{Y}^{(\ell)} - \widehat{\mathbf{Y}}^{(\ell)}\right), \tag{21}$$

the transformed perturbation parameter-predicted observation and predicted observation error covariance matrices are

$$\mathcal{C}_{\widetilde{\mathbf{\Theta}}\widehat{\mathbf{Y}}}^{(\ell)} = \frac{1}{N_e}\widetilde{\mathbf{\Theta}}^{(\ell)'}\widehat{\mathbf{Y}}^{(\ell)'\mathrm{T}}, \tag{22}$$

and

$$\mathcal{C}_{\widehat{\mathbf{Y}}\widehat{\mathbf{Y}}}^{(\ell)} = \frac{1}{N_e}\widehat{\mathbf{Y}}^{(\ell)'}\widehat{\mathbf{Y}}^{(\ell)'\mathrm{T}}, \tag{23}$$

respectively, in which primes ($'$) denote anomalies (deviations from the ensemble mean).

5. Apply the appropriate inverse transforms to recover the updated perturbation parameters

$$\mathbf{\Theta}^{(\ell+1)} = \psi^{-1}\left(\widetilde{\mathbf{\Theta}}^{(\ell+1)}\right), \tag{24}$$

ψ^{-1} is the exponential and the inverse logit for the biases and the remaining perturbation parameters respectively.

The observation error inflation coefficient $\alpha^{(\ell)}$ in (21) together with the iterations sets the ES-MDA apart from the traditional ensemble smoother (ES; Van Leeuwen and Evensen, 1996). For $N_a = \alpha^{(\ell)} = 1$, the ES scheme is recovered, which was used in the probabilistic SWE reconstruction of Durand et al. (2008) and Girotto et al. (2014b). The idea behind the ES-MDA is to perform multiple smaller analysis steps as opposed to one abrupt analysis step. In the case of a non-linear model, this is expected to yield a better approximation of the true posterior (Emerick and Reynolds, 2013). A requirement for the ES-MDA to give a nearly unbiased estimate (c.f. Stordal and Elsheikh, 2015) is that the coefficients satisfy $\sum_{\ell=0}^{N_a-1} \frac{1}{\alpha^{(\ell)}} = 1$. In our case this is accomplished by setting all the coefficients as $\alpha^{(\ell)} = N_a$ and specifying N_a before any assimilation cycles are carried out. **We emphasize that the analysis step** (21) **only updates the perturbation parameters and a consistent ensemble of states is found from the subsequent ensemble model integration. The model constants listed in Table 2 remain unchanged by the analysis and the integration.** As mentioned, the perturbation parameter matrix $\widetilde{\boldsymbol{\Theta}}$ in (21) is transformed through analytical Gaussian anamorphosis (Bertino et al., 2003) to ensure that the priors are Gaussian. In this case, the Kalman-like analysis step (21) is variance minimizing for a linear model (Van Leeuwen and Evensen, 1996). The entire methodology, with ES-MDA as the DA scheme, is depicted in Figure 3.

Margulis et al. (2015) introduced the particle batch smoother (PBS) for snow data assimilation. In this scheme, each particle (i.e. ensemble member; Van Leeuwen, 2009) is given an equal prior weight of $1/N_e$. Then, after an ensemble model integration, the normalized posterior importance weights $w_j \in [0, 1]$ are diagnosed through the analysis step

$$w_j = p\left(\mathbf{y}|\widehat{\mathbf{X}}_j\right) p\left(\widehat{\mathbf{X}}_j\right) / \sum_{j=1}^{N_e} \left(p\left(\mathbf{y}|\widehat{\mathbf{X}}_j\right) p\left(\widehat{\mathbf{X}}_j\right)\right), \tag{25}$$

where $\widehat{\mathbf{X}}_j = [\mathbf{X}_j \, ; \, \boldsymbol{\Theta}_j]$ is the augmented state vector for the j-th particle and the Gaussian likelihoods are given by

$$p\left(\mathbf{y}|\widehat{\mathbf{X}}_j\right) = c_0 \exp\left[-0.5\left(\mathbf{y} - \widehat{\mathbf{Y}}_j\right)^{\mathrm{T}} \mathbf{R}^{-1} \left(\mathbf{y} - \widehat{\mathbf{Y}}_j\right)\right]. \tag{26}$$

This is a direct application of Bayes' rule in which the normalizing denominator has two important consequences. Firstly, $c_0 = 1/\sqrt{(2\pi)^{N_o} |\mathbf{R}|}$ cancels out thus avoiding errors introduced through floating point arithmetic ($(2\pi)^{N_o}$ is generally large). Secondly, the prior weights $p(\widehat{\mathbf{X}}_j)$ also cancel as they are equal for all particles. With Gaussian likelihoods, (25) becomes

$$w_j = \frac{\exp\left[-0.5\left(\mathbf{y} - \widehat{\mathbf{Y}}_j\right)^{\mathrm{T}} \mathbf{R}^{-1} \left(\mathbf{y} - \widehat{\mathbf{Y}}_j\right)\right]}{\sum_{j=1}^{N_e} \exp\left[-0.5\left(\mathbf{y} - \widehat{\mathbf{Y}}_j\right)^{\mathrm{T}} \mathbf{R}^{-1} \left(\mathbf{y} - \widehat{\mathbf{Y}}_j\right)\right]}, \tag{27}$$

where the posterior weights w_j sum to unity. The posterior ensemble still spans the range of the prior ensemble, as the analysis step only changes the relative weights of the ensemble members and not their position within the state and perturbation parameter space. Through the individual ranking of the states and perturbation parameters followed by a cumulative summation of the correspondingly sorted weights, the marginal cumulative distribution functions are readily recovered, facilitating the estimation of quantile values. Note that the PBS is equivalent to running a particle filter without re-sampling and using the prior as the importance density (see Van Leeuwen, 2009). So, the PBS is a not novel data assimilation scheme; it corresponds to the generalized likelihood uncertainty estimation method (GLUE; Beven and Binley, 1992) with a Gaussian likelihood function. Due

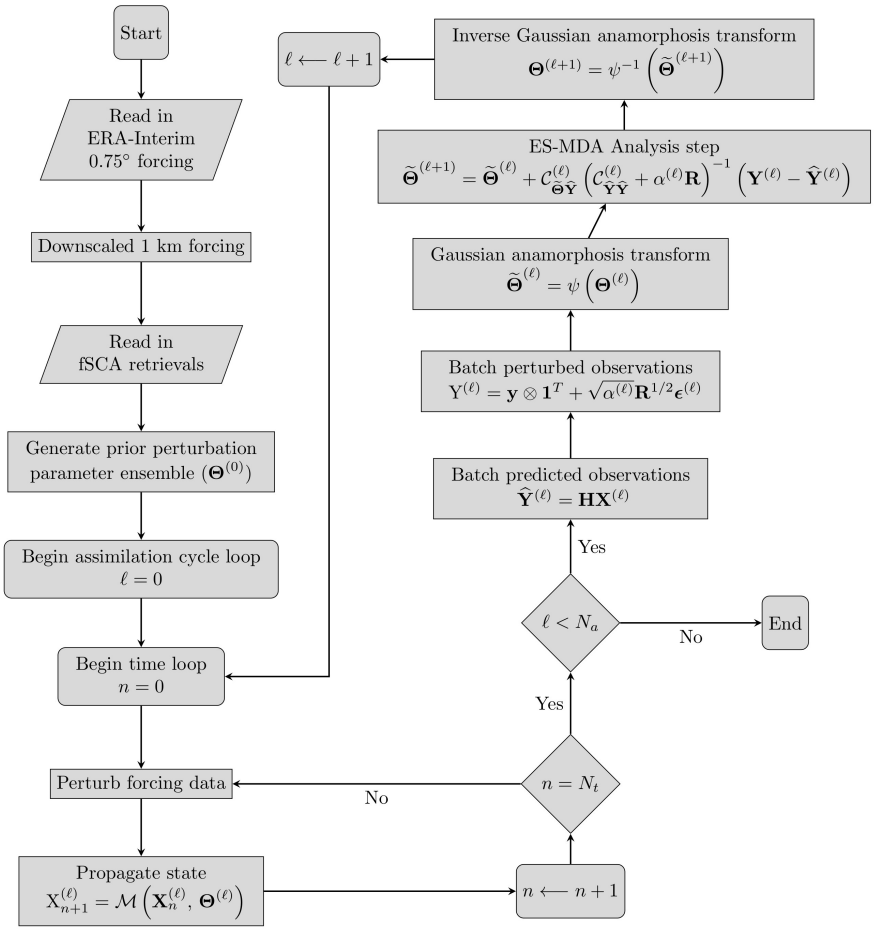

Figure 3. Flowchart depicting the methodology with the ES-MDA as the DA scheme. Symbols are defined in the text.

to the absence of re-sampling, even for medium dimensional systems with a large number of observations to be assimilated, the PBS can become degenerate with very few particles carrying the majority of the importance weights (Van Leeuwen, 2009). Nevertheless, a major advantage of the PBS is its computational efficiency, requiring only one ensemble model integration and one efficient analysis step (27). In this study, the PBS and the ES are used to benchmark the ES-MDA.

4 Results

4.1 Interannual variability and comparison to field measurements

In this section, we present results of the ES-MDA scheme with 100 ensemble members and 4 assimilation cycles (Section 3.3.2) for all the years and sites where snow surveys were conducted. Figure 4 shows the time series of the prior and posterior

fSCA (left panel) and mean SWE (right panel) estimates, as well as the assimilated fSCA satellite retrievals (left panel) and the independently observed peak mean SWE (from the snow surveys, right panel). The assimilation generally brings the posterior estimates closer to the observed fSCA and considerably constrains the spread of the ensemble compared to the prior. For some occasions, such as Bayelva 2008, Bayelva 2009 and Kvadehuksletta 2016, the timing of the snowmelt in the prior is significantly biased by as much as three weeks compared to the assimilated fSCA retrievals. Even if the prior ensemble does not encompass the retrievals, the iterative ES-MDA scheme allows the posterior to converge towards the fSCA observations (left panels), leading to much improved SWE estimates (right panels). On other occasions, such as 2015 at Bayelva and 2016 at Steinflåen plateau, the prior ensemble is a reasonable estimate and the assimilation merely constrains the spread of the ensemble and adjusts the median slightly. Both for Bayelva in 2015 and Kvadehuksletta in 2016, some of the early fSCA retrievals, which indicate a slight ablation, are completely ignored by the assimilation, as this early onset of melt is inconsistent with the model - even when biases are accounted for. However, this ablation could be real and due to processes not accounted for in the model, such as wind erosion.

Field measurements of peak mean SWE are available for years with low (2008, 2016), medium (2013, 2015) and high (2009, 2014) values of peak mean SWE, ranging from $0.08\,\mathrm{m\,w.e.}$ (Kvadehuksletta 2016) to $0.48\,\mathrm{m\,w.e.}$ (Bayelva 2014). With the exception of two cases (Bayelva 2013 and Steinflåen plateau 2016), the assimilation brings the ensemble median closer to the observed peak mean SWE, while at the same time constraining the spread of the ensemble. We emphasize that the assimilation performs a global bias correction for peak SWE. This is especially evident for Kvadehuksletta 2016 for which the assimilation unrealistically truncates the duration of the snow season as a result of a strong correction for the positive bias. Both in 2008 and 2009 for Bayelva, ES-MDA shifts the estimates to better match field measurements of SWE (which were not assimilated), despite the prior range being far from the observations. The posterior ensemble median peak mean SWE is generally close to the independently observed peak mean SWE, but absolute relative differences up to 40% (minimum 0.5%, mean 19%) occur.

Figure 5 displays the prior, posterior and observed subgrid SWE distributions (SSDs) for the years and sites with field measurements. Again, with the exception of Bayelva 2013 and Steinflåen plateau 2016, the assimilation brings the mean of the peak SSD closer to the observations. The agreement between the posterior and observed mean value is striking for a number of years and sites, such as Bayelva in 2009 and 2014 as well as Kvadehuksletta in 2016. Furthermore, the shapes of the observed and posterior distributions agree well, e.g. for Bayelva 2008, Bayelva 2013 and Bayelva 2016. Once more, the correction from prior to posterior is largest for Bayelva 2008 and Bayelva 2009, for which the prior fSCA was furthest from the satellite retrievals. The prior ensemble SSD, apart from Bayelva 2013, is generally too positively skewed (i.e. has a long tail) compared to the observed SSD. On some occasions the match between the posterior and observed SSDs is poor, such as Steinflåen plateau 2016 and Bayelva 2015. We conclude that the analysis typically improves the fit between modeled and observed snow distributions. Some of the observed distributions, such as that for Kvadehuksletta 2016, are hard to match as they do not conform well to a lognormal distribution, most likely due to the limited number of sample points (Section 2.2).

The posterior bias parameters can be directly evaluated by comparing the bias corrected forcing to field measurements. Due to a lack of snowfall observations (see Boike et al., 2017), an evaluation of the precipitation bias parameter is not possible. However, the melt bias parameter can be evaluated by comparing the estimated snowmelt flux (which

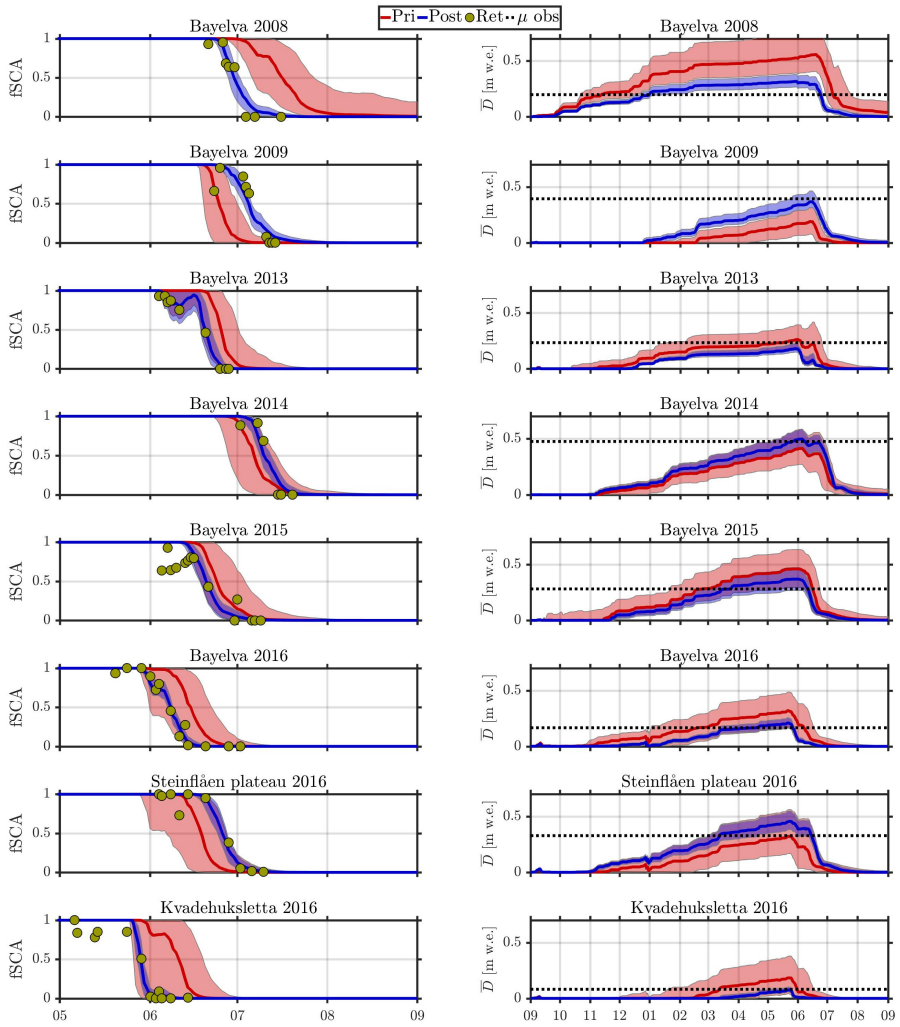

Figure 4. Time evolution of the prior (red) and ES-MDA ($N_a = 4$, $N_e = 10^2$) posterior (blue) fSCA (first column) and mean SWE ($\overline{D}$; second column); shading: 90$^{\text{th}}$ percentile range; solid lines: ensemble median; yellow dots: assimilated MODIS and Sentinel-2 fSCA retrievals; dotted black line: independently observed peak mean SWE (μ) from snow surveys (Section 2.2); x-axis: months. These results are from a single run.

is directly proportional to the perturbed melt depth) to field-based values. For June 2008, Westermann et al. (2009) estimate an average snowmelt flux of 27 $\mathrm{Wm^{-2}}$, which compares well to the ES-MDA posterior median (averaged for the same period) of 29 $\mathrm{Wm^{-2}}$, while the prior median is too low with 19 $\mathrm{Wm^{-2}}$.

In ensemble-based data assimilation the spread of the posterior ensemble should represent the uncertainty. To verify this one can compare two metrics: the residual, i.e. the instantaneous posterior RMSE of the ensemble relative to the corresponding independent field measurement, and the ensemble standard deviation (e.g. Evensen, 2009). For this

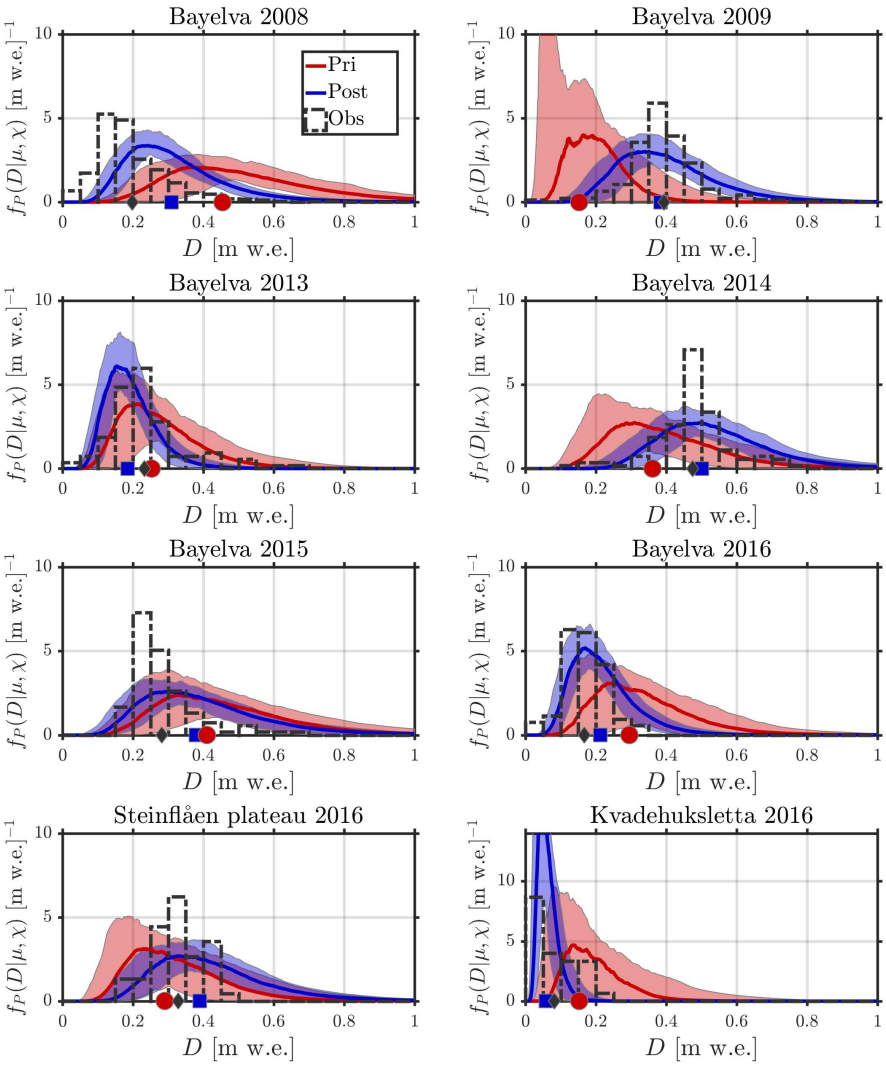

Figure 5. Prior (red), ES-MDA ($N_a = 4$, $N_e = 10^2$) posterior (blue) and the corresponding independently observed (from snow surveys; dashed black) peak subgrid SWE distributions; shaded areas: 90$^{\text{th}}$ percentile range; solid lines: ensemble median; markers: mean value. These results are from a single run.

comparison we define the relative residual as the ratio of the residual to the standard deviation. Ideally this ratio should have a value of 1 which indicates that the two metrics are equal so that the posterior ensemble spread accurately captures the estimation uncertainty. For the fSCA, peak mean SWE and peak subgrid coefficient of variation the average (over all available field measurements) relative residuals were 2.22, 1.53 and 1.66 respectively, so the posterior ensemble underestimates the uncertainty. This effect has been extensively described by Evensen (2009), it arises in part because

	Prior			ES-MDA			ES			PBS		
Symbol	Bias	RMSE	R^2	Bias	RMSE	R^2	Bias	RMSE	R^2	Bias	RMSE	R^2
fSCA (106)	0.21	0.02	0.80	0.03	0.01	0.97	0.03	0.01	0.97	0.03	0.01	0.97
μ (8) [m w.e.]	0.13	0.21	0.01	0.06	0.09	0.77	0.06	0.12	0.76	0.06	0.08	0.76
χ (8)	0.01	0.16	0.00	-0.02	0.13	0.37	0.02	0.14	0.33	-0.03	0.13	0.06

Table 5. Summary of evaluation metrics, i.e. bias, RMSE and square correlation coefficient (R^2), for the fSCA, peak SWE (μ) and peak subgrid coefficient of variation (χ). These metrics are based on comparisons to all the field measurements presented in Section 2.2 with the number of observations for the comparisons in brackets next to the corresponding symbols. All the metrics are averaged over 100 independent runs each with 100 ensemble members. The ES-MDA was run with $N_a = 4$ assimilation cycles.

of model structural errors related to neglected physical processes (Section 3.1). Still, the assimilation is generally able to simultaneously (but not to the same extent) reduce the spread and the error in the ensemble (Figure 4).

4.2 Evaluation of data assimilation schemes

In addition to the ES-MDA scheme, we evaluate the PBS and ES (Section 3.3.2) with regards to field measurements, using an ensemble size of 100 members for all schemes. Three error metrics are summarized in Table 5: the bias (mean error), RMSE and the square correlation coefficient. For the fSCA, all the schemes achieve a major improvement relative to the prior with an almost tenfold reduction in bias, a halving of RMSE and an almost perfect correlation to the field measurements of fSCA (Section 2.2). For the peak mean SWE (μ), the PBS performs best in terms of RMSE and bias, followed closely by ES-MDA which, in turn, has the highest correlation coefficient. With regards to the peak subgrid coefficient of variation (χ), ES-MDA performs best across all the metrics, tying with ES for (absolute) bias and the PBS for RMSE. As considerably more field measurements are available for fSCA than for μ and χ, the evaluation for fSCA must be considered more robust. **Figure 6 displays scatter plots for a single run for the prior and all the DA schemes relative to the field measurements.**

Observed, prior and posterior peak mean SWE and peak subgrid coefficient of variation for different years/sites are shown in Figure 7. As discussed in Section 4.1, the assimilation moves the posterior peak mean SWE estimates closer to the observed peak mean SWE in most cases when compared to the prior. However, clear performance differences are found between the different schemes for a number of situations. In 2008, the PBS is not able to correct for as much of the bias in the peak mean SWE compared to the ES-MDA and the ES. For the remaining years, the performance of the schemes in terms of estimating peak mean SWE is similar, but the spread of the ES is by far the largest, followed by the PBS and the ES-MDA. The PBS ensemble shows indications of degeneracy for some years (e.g. 2008 and 2009) where the median coincides with either the 5$^{\text{th}}$ or 95$^{\text{th}}$ percentile value. This indicates that the majority of the weight in the PBS is carried by just a few ensemble members. For the coefficient of variation, the 90$^{\text{th}}$ percentile range of the ES-MDA posterior ensemble typically encompassed

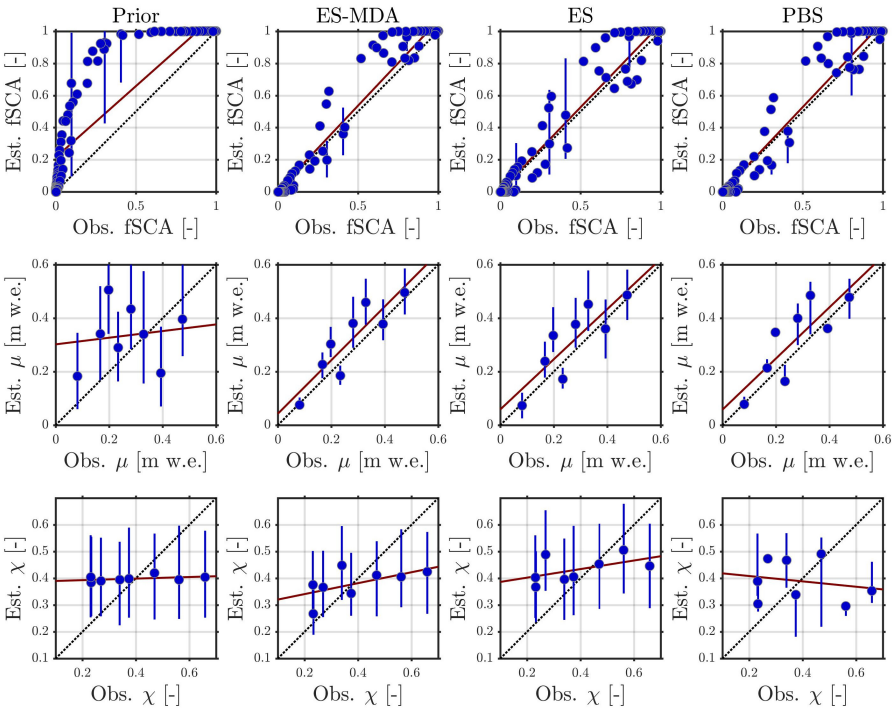

Figure 6. Scatter plots of the prior median (left panels) as well as ES-MDA ($N_a = 4$; left-middle panels), ES (right-middle panels) and PBS (right panels) posterior median estimates of fSCA (top panels), peak mean SWE μ (middle panels) and peak subgrid coefficient of variation χ (lower panels) versus the observations (field measurements); error bars: 90$^{\text{th}}$ percentile range; all DA schemes were run with 10^2 ensemble members; one to one: dotted black line; linear best fit: solid red line. These results are from a single run.

the observed value (with two exceptions), while this was not true to the same extent for the ES (three exceptions) and the PBS (five exceptions). These performance differences explain the higher correlation coefficient for the coefficient of variation for the ES-MDA scheme (Table 5). The PBS also shows signs of degeneracy (e.g. Bayelva 2009, complete degeneracy) for the coefficient of variation estimation. On some occasions (e.g. Bayelva in 2008, 2009 and 2014), the posterior ensemble median is effectively pulled closer to the observed coefficient of variation when compared to the prior. On the same occasions the ensemble spread is slightly constrained. Compared to the peak mean SWE, it is much harder to constrain estimates of the coefficient of variation regardless of scheme, although it is possible to shift the ensemble in the right direction.

We gauged the sensitivity of the three batch smoother schemes with respect to ensemble size and the number of assimilation cycles by considering the fractional improvement (FI) in RMSE that was achieved through the analysis step based on all available field measurements (Figure 8). On the one hand, the PBS requires an ensemble size of 1000 to converge to a stable FI of around 75%, 20% and 60% for the fSCA, peak subgrid coefficient of variation, and peak mean SWE respectively. On the other hand, ES-MDA with 4 assimilation cycles converges at already 100 ensemble members at similar FIs to the PBS. The ES performs worst regardless of ensemble size, with FIs of around 70%, 10% and 55% even with 10^5 ensemble members

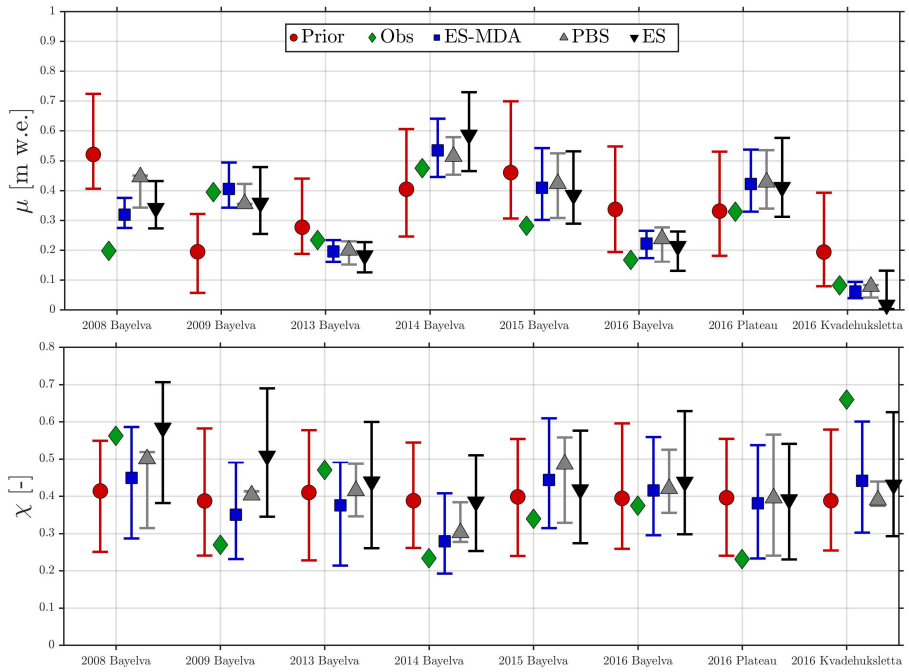

Figure 7. Prior median, observed, ES-MDA (with $N_a = 4$) posterior median, PBS posterior median and ES posterior peak mean SWE μ (upper panel) and peak subgrid coefficient of variation χ (lower panel); error bars: 90$^{\text{th}}$ percentile range; all DA schemes were run with 10^2 ensemble members. These results are from a single run.

requiring 100 ensemble members for convergent results. **For all schemes the available validation data suggests that the greatest improvements are achieved for fSCA, followed by peak mean SWE, while by far the lowest improvements are found for the peak subgrid coefficient of variation**. With 100 ensemble members, the ES-MDA converges to a stable performance at 4 assimilation cycles, i.e. there is no marked improvement of FI for more cycles (Figure 8, bottom right).

4.3 Effects of observation error and assimilation frequency

The effects of observation error and assimilation frequency are studied by running the ES-MDA ($N_e = 10^2$, $N_a = 4$) and assimilating first only MODIS and then both MODIS and Sentinel-2 retrievals for the 2016 snow season at all study sites. As discussed in Section 3.2, the Sentinel-2 fSCA retrievals are based on higher resolution optical reflectance retrievals. As such they are expected to be less prone to representativeness error and thus observation error since the area in which the snow surveys were conducted is more accurately covered by the retrievals. Furthermore, the Sentinel-2 scenes used for fSCA retrievals were manually checked to be cloud free, which was not the case for the MODIS scenes. Table 6 summarizes various performance metrics for the two different runs. For the peak mean SWE depth (μ), there is no difference when including Sentinel-2 fSCA retrievals in the assimilation. For the coefficient of variation (χ), however, there is an increase in FI for both the bias and the RMSE, as well as an increase in the square correlation coefficient. Sentinel-2 fSCA retrievals with lower observation error help

to further constrain the shape of the snow depletion curve which explains the improvement in the χ estimation. **We emphasize that this evaluation is based on the only 3 available field measurements of μ and χ in 2016 (from the snow surveys), so that these preliminary results need to be consolidated by future studies with more validation data.**

		MODIS				MODIS+S2			
Symbol	# of obs	FI Bias	FI RMSE	R^2_{prior}	R^2_{post}	FI Bias	FI RMSE	R^2_{prior}	R^2_{post}
μ [m w.e.]	3	0.61	0.62	0.67	1.00	0.61	0.62	0.69	1.00
χ	3	-0.77	0.11	0.00	0.41	-0.60	0.15	0.00	0.48

Table 6. Summary of evaluation metrics, i.e. fractional improvement in bias and RMSE as well as prior and posterior square correlation coefficient (R^2), using the ES-MDA ($N_e = 10^2$, $N_a = 4$) for peak mean SWE (μ) and coefficient of variation (χ) when assimilating only MODIS as well as assimilating both MODIS and Sentinel-2 observations. These metrics are based on a comparison to all the snow surveys conducted in 2016 (see Table 1) and are averaged over 100 independent runs each with 100 ensemble members.

5 Discussion

5.1 Interannual variability and comparison to field measurements

For all considered years and sites, the ES-MDA scheme both brings the ensemble median fSCA closer to the observed fSCA and significantly constrains the spread of the ensemble (Figure 4). Thus, the posterior effectively fills the gaps in the remotely sensed fSCA time series using a physically based snow model which is bias-corrected through the assimilation, while at the same time accounting for uncertainties in the retrievals. In addition, the ES-MDA is generally able to correct the prior estimates of the peak mean SWE towards the independently observed values, which is essentially achieved through a bias correction of the model forcing. Although the downscaled forcing is biased, it is a more reliable input than forcing data obtained directly from coarse-scale reanalyses (Østby et al., 2017). For example, the lapse rate correction on temperature in the downscaling (c.f. Fiddes and Gruber, 2014) influences the snowfall and melt rates at the more elevated Steinflåen plateau. This effect is not captured in the reanalysis product (Dee et al., 2011) for which the elevation of the nearest grid point is near sea level.

An inherent equifinality problem (see Beven, 2006) exists in SWE reconstruction since different perturbation parameter sets can provide similar results. For example, if the prior fSCA melts out earlier than the observations this could be due to the prior precipitation having a negative bias, the prior melt having a positive bias or a combination of these two. The opposite would be true if the prior fSCA melts out too late. It is not possible to resolve this equifinality problem with observations of fSCA alone. A key assumption in deterministic SWE reconstruction is that the melt flux is more constrained than the precipitation so that uncertainty in the melt is ignored (Slater et al., 2013). We perturb both the

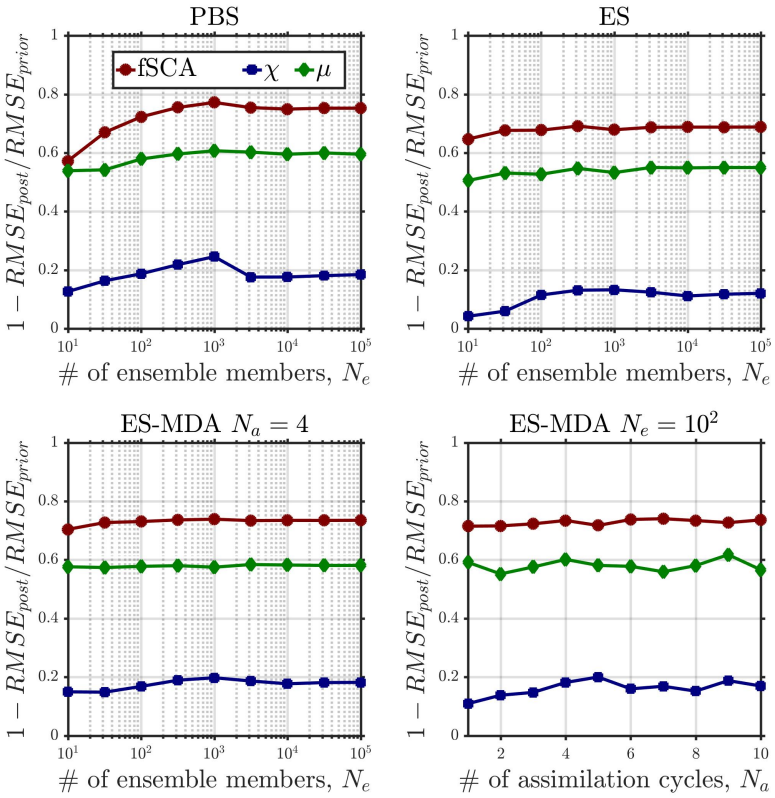

Figure 8. Fractional improvement in RMSE through the analysis step (1 being perfect and 0 no effect) as a function of the number of ensemble members for the fSCA, peak mean SWE μ and coefficient of variation χ; top left: particle batch smoother, PBS; top right; ensemble smoother, ES; bottom left; ensemble smoother with multiple data assimilation, ES-MDA; bottom right: FI as a function of assimilation cycles in the ES-MDA. The FI for $N_e \leq 100$ are averaged over 100 independent ensemble model integrations. Errors were computed based on comparisons to all the corresponding field measurements presented in Section 2.2.

precipitation and the melt, although the latter is assigned a lower uncertainty (Table 4). Through the assimilation we obtain snowmelts that are consistent with the observed snow cover depletion. The close match of the posterior peak mean SWE estimates to the independent field measurements (Figure 7) suggests that the assimilation yields consistent accumulations and that the inherent equifinality problem is of minor consequence.

Figure 5 shows that for most years the prior median is a poor estimate of the observed peak mean SWE. This indicates that a deterministic (no assimilation and unperturbed) run is not a good representation of the true state. In addition to biases in the precipitation and melt forcing, crucial processes for peak SWE, such as deposition, are not included in the simple snow model. Furthermore, the subgrid variability of the SWE is typically overestimated in the prior, with the prior distributions typically being too skewed. To circumvent these issues, a more sophisticated model (e.g. ALPINE3D; Lehning et al., 2006)

accounting for wind drift could be employed, and the climatological snow distribution pattern (Sturm and Wagner, 2010) could help formulate the prior peak subgrid coefficient of variation distribution.

The posterior distributions are on the other hand much closer to the observed distributions for most of the years and sites considered. This suggests that there is sufficient information contained in the remotely sensed snow cover depletion to constrain the peak SSD estimates. On some occasions, especially for Bayelva 2015, the posterior SSD is far from the observed SSD both in shape and in mean. However, the posterior estimate is still slightly better than the prior, indicating that the assimilation has a positive effect on the outcome. A similar marginal performance is found for Steinåen plateau in 2016, but the number of SWE point observations (see Table 1) is not sufficient to reliably constrain the shape of the observed distribution in this case.

5.2 Evaluation of data assimilation schemes

Compared to the ES and the PBS, which were used in previous studies (e.g. Durand et al., 2008; Girotto et al., 2014b; Margulis et al., 2015), the ES-MDA exceeds or at least nearly matches in performance for all the evaluation metrics considered, i.e. bias, RMSE and correlation coefficient for fSCA, peak mean SWE and peak subgrid coefficient of variation. The performance gain over the ES is explained by the iterative nature of the ES-MDA, performing a sequence of smaller corrections in the analysis steps as opposed to one abrupt correction (Emerick and Reynolds, 2013; Stordal and Elsheikh, 2015). Particularly in the case of a non-linear model, as is the case for the SSM, this process of simulated annealing (c.f. Stordal and Elsheikh, 2015) in the ES-MDA leads to a better approximation of the posterior than a single analysis step.

At least with a low number of ensemble-members the ES-MDA also outperforms the PBS. A possible reason for this is that the PBS posterior ensemble spans the same range as the prior ensemble and only changes the relative weights of the ensemble members in the analysis. **Thus, if the prior ensemble is so biased that it does not encompass the observations, the PBS is incapable of correcting the posterior towards the observations outside the bounds of the prior.** In such a case, the region with high likelihood is very small and not necessarily close to the observations. A good example is the 2008 season at Bayelva (c.f. Figure 4 and Figure 7) for which the prior is far away from the observed fSCA. Consequently, the PBS is unable to shift the ensemble outside the prior range as opposed to both the ES and the ES-MDA. In several years, the PBS also shows signs of degeneracy, i.e. a large part of the weight is carried by a very low number of particles. As the PBS is essentially a particle filter without re-sampling (Van Leeuwen, 2009), the weights can quickly converge on just a few particles in high likelihood regions. Therefore, the remaining particles become redundant even for low-dimensional systems with a relatively large number of observations such as the one considered here.

The sensitivity analysis for the ensemble size is consistent with higher dimensional models. The ES-MDA requires relatively few ensemble members for convergence, similar to the EnKF (Evensen, 2009), while the PBS requires a larger ensemble for convergence as with the PF (Van Leeuwen, 2009). The number of assimilation cycles required for convergence of the ES-MDA (4 cycles) is also in line with previous studies (Emerick and Reynolds, 2013). While the PBS and the ES-MDA have not yet been compared with respect to improvement in RMSE, our findings are in agreement with previous studies for both the PBS versus the ES (Margulis et al., 2015) and the ES-MDA versus the ES (Emerick and Reynolds, 2013).

A major downside of the ES-MDA is the computational cost. The ES-MDA requires $N_a + 1$ ensemble model integrations and N_a analysis steps, where N_a (typically ≥ 2) is the number of assimilation cycles (Section 3.3.2). On the other hand, the ES requires only two ensemble model integrations and a single analysis step, while the PBS only needs one ensemble model integration and a single analysis step. Based on a sensitivity analysis (Section 4.2) we set $N_a = 4$, so the computational cost of the ES-MDA is higher than for the other schemes. For more complex models, such as Crocus (Vionnet et al., 2012) and SNOWPACK (Bartelt and Lehning, 2002), the ES-MDA could prove to be prohibitively expensive. However, an adaptive version of the ES-MDA (Le et al., 2016) could be employed instead in which the inflation coefficients are calculated on the fly based on a cost function and the iterations stop once the algorithm converges. This scheme could significantly reduce the computational costs for applications of the ES-MDA, as it is equivalent to the ES in years for which the prior encompasses the fSCA retrievals, requiring only one iteration. In years where the prior is far from the observations to be assimilated, on the other hand, multiple data assimilation steps are performed. Furthermore, both the snow model (which has no interaction between neighboring grid cells) and the ES-MDA algorithm can be parallelized using high performance computing.

5.3 Effects of observation error and assimilation frequency

All the DA methods have problems constraining the spread of the peak subgrid coefficient of variation (χ, see Figure 7) although they can pull the median in the right direction. A likely reason is the limited information available in the remotely sensed snow cover depletion, with either too sparse or too uncertain fSCA retrievals. It is worth considering fSCA retrievals from even more sensors such as Landsat and PROBA-V, which increases the chances of obtaining more cloud-free scenes. With more scenes available, it may be possible to better constrain the posterior χ ensemble: even with just a few additional retrievals from Sentinel-2 the performance was improved with respect to χ estimation across all evaluation metrics. This also points towards the benefits of including higher resolution fSCA retrievals from the Landsat and Sentinel-2 satellites which will be more representative and thus accurate. The effective MODIS footprint is inhomogeneous and differs markedly from the nominal 500 m pixel resolution when the view angle deviates from nadir (Peng et al., 2015). So, even for gridded applications, there is a considerable representativeness error in MODIS fSCA, although this is reduced when several pixels are aggregated.

5.4 Outlook

Several extensions to the presented ensemble-based data assimilation framework could be considered. The first is to change the grid scale of the framework from the order of $1\,\mathrm{km}$ to larger or smaller scales. For the latter, it would be possible to assimilate only Landsat and Sentinel-2 based fSCA retrievals and operate at a grid scale on the order of $100\,\mathrm{m}$ in line with the work of e.g. Girotto et al. (2014a). For the former, one would aggregate the satellite retrievals even further and perform the assimilation at a grid scale on the order of $10\,\mathrm{km}$ or larger. This implementation could be problematic as the uniform snowmelt assumption in the SDC (Liston, 2004) may no longer be justified across such large grid cells.

Furthermore, the method could be applied to a larger domain in spatially distributed mode (i.e. multiple grid cells). In this case, fSCA assimilation could be complemented by the assimilation of GRACE TWS and/or PM SWE retrievals which can

also improve SWE estimates during the entire build-up, not only at peak SWE. Both TWS and PM retrievals could constrain the large-scale areal average SWE estimate within the domain and thus further bias-correct the multiple grid-scale peak mean SWE estimates. However, GRACE TWS retrievals feature a very coarse resolution (around $100\,\mathrm{km}$) so that they would only be useful in conjunction with fSCA retrievals for very large scale applications. On the other hand, the use of higher resolution PM SWE retrievals (order $25\,\mathrm{km}$) in the assimilation has shown particular promise (e.g. De Lannoy et al., 2012; Li et al., 2017). At the same time, PM SWE retrievals are not accurate in complex topography and forested areas as well as for wet and deep snowpacks (Foster et al., 2005), which might limit the applicability of such multisensor assimilation approaches.

The major problem in the assimilation of fSCA retrievals is the occurrence of clouds which causes extended gaps in time series obtained from optical sensors. As discussed, using fSCA retrievals from even more sensors could help to fill in the gaps in the remotely sensed snow cover depletion and further constrain the peak subgrid coefficient of variation (χ). The use of additional higher resolution fSCA retrievals with lower representativeness error (and thus observation error) could also prove especially beneficial for constraining χ.

To reduce the computational costs of the ES-MDA, the adaptive ES-MDA (Le et al., 2016) should be considered. Furthermore, the bias corrected ES-MDA outlined in Stordal and Elsheikh (2015) may be worth pursuing for future applications, especially when applied to bigger domains with possibly even larger misfits between the prior and the observations. Using a more complex snow model such as Crocus (Vionnet et al., 2012) or SNOWPACK (Bartelt and Lehning, 2002) may not only improve the modeled melt rates, but also offer the possibility to assimilate snow grain size retrievals (c.f. Painter et al., 2009), as noted by Durand et al. (2008). In addition, the method could be applied in a fully coupled land-atmosphere model. The intermediate complexity atmospheric research model (ICAR; Gutmann et al., 2016) shows particular promise in terms of an atmospheric model that can efficiently and iteratively be run in ensemble mode, as required for applications of ES-MDA. In principle, one could run ICAR in ensemble mode coupled to a land surface model with an adequately complex snow scheme and assimilate fSCA (and possibly PM and TWS) retrievals with the ES-MDA to deliver a snow reanalysis.

As snow is a crucial driver for many terrestrial and atmospheric processes, the presented framework has the potential to improve process modeling in a range of disciplines, especially since the spatial resolution is considerably higher than in passive-microwave derived SWE data sets. For example, the subgrid variability of permafrost temperatures is closely tied to that of SWE depth (e.g. Gisnås et al., 2016), which has major implications for permafrost mapping (e.g. Westermann et al., 2015b, 2017). Similarly, snow cover information is an important component of many ecological models (e.g. Kohler and Aanes, 2004), and peak SWE is intimately linked to stream flow which is crucial for hydrology and water resource management (Andreadis and Lettenmaier, 2006; Barnett et al., 2005). Finally, knowledge of the snow distribution and snowmelt is of interest for tourism given its importance for e.g. skiing, hiking and backcountry travel.

6 Conclusions

In this study, we use the ensemble smoother with multiple data assimilation (ES-MDA) scheme to estimate peak SWE distributions at the kilometer scale from time series of remotely sensed fractional snow covered area (fSCA) from MODIS and

Sentinel-2. The ES-MDA is combined with analytical Gaussian anamorphosis to update perturbation parameters that were either lower or double bounded in physical space. The data assimilation is applied to a simple snow model based on the surface energy balance coupled to a probabilistic snow depletion curve. The scheme is driven by statistically downscaled ERA-Interim reanalysis data. As such, both the model forcing and the satellite retrievals are globally available.

The results are compared to field measurements of peak SWE distributions and fSCA from Arctic sites near Ny-Ålesund (79°N, Svalbard, Norway) so that the performance can be evaluated with respect to the estimated fSCA, peak mean SWE and peak subgrid coefficient of variation. From this study, the following conclusions can be drawn:

- At the kilometer scale, the ES-MDA is able to successfully assimilate fSCA retrievals into the simple snow model and estimate the peak subgrid SWE distribution prior to the snowmelt.
- A physically-based interpolation of the remotely sensed fSCA time series is obtained that takes into account uncertainties in both the model and the retrievals.
- **For the peak mean SWE, the ES-MDA features an average RMSE of 0.09 $\mathrm{m\ w.e.}$ compared to field measurements**.
- For the peak subgrid coefficient of variation that controls the width and skewness of the distribution, the ES-MDA usually manages to pull the posterior median in the right direction, but the spread of the ensemble was difficult to constrain.
- By including higher resolution fSCA retrievals from Sentinel-2, the posterior peak subgrid coefficient of variation ensemble can be better constrained. This highlights the potential benefits of assimilating additional higher resolution fSCA retrievals from sensors on board the Landsat and Sentinel-2 satellites in future work.
- In line with previous studies, the ES-MDA converges with as low as 100 ensemble members and 4 assimilation cycles.
- With this ES-MDA configuration, the fractional improvement in RMSE from prior to posterior is around 75%, 60% and 20% for the fSCA, peak mean SWE and peak subgrid coefficient of variation.
- The ES-MDA exceeds or at least nearly match the performance of the particle batch smoother and the ensemble smoother for all evaluation metrics considered.

As the scheme exploits high and medium resolution satellite images from optical sensors, it is capable of estimating snow distribution at considerably higher spatial resolutions than traditional SWE products, e.g. based on passive microwave retrievals. On the other hand, the scheme can only recover the peak subgrid SWE distribution prior to the onset of melt, as opposed to providing information on the seasonal evolution of the snow distribution, so that it can rather complement than replace existing SWE retrieval algorithms. However, the method could become a part of satellite-era hydrometeorological reanalysis schemes with a wide range of applications.

7 Data and code availability

Data and code are made available upon request from the corresponding author.

Author contributions. KA performed this study as part of his PhD project, co-supervised by SW and LB. KA wrote the code and performed the analysis. SW conceived the original idea and helped with downloading satellite data. SW performed the field measurements, with KA assisting in the final year. LB assisted with the implementation of the data assimilation. TVS performed the statistical downscaling. JB provided data from and information about the Bayelva station. The manuscript was first written by KA and then edited by all the co-authors.

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

Acknowledgements. This work was funded by SatPerm (239918; Research Council of Norway), in collaboration with EmblA (56801; Nordforsk) and ESA GlobPermafrost (www.globpermafrost.info). SW acknowledges additional support by COUP (2449037/E10; JPI Climate; Research Council of Norway) and PermaNor (255331/E10; Research Council of Norway). TVS acknowledges funding from ESCYMO (224024; Research Council of Norway). This study is a contribution to the strategic research area LATICE at the University of Oslo.

References

Aas, K. S., Gisnås, K., Westermann, S., and Berntsen, T. K.: A Tiling Approach to Represent Subgrid Snow Variability in Coupled Land Surface-Atmosphere Models, Journal of Hydrometeorology, 18, 49–63, doi:10.1175/JHM-D-16-0026.1, 2017.

Andreadis, K. M. and Lettenmaier, D. P.: Assimilating remotely sensed snow observations into a macroscale hydrology model, Advances in Water Resources, 29, 872–886, doi:10.1016/j.advwatres.2005.08.004, 2006.

Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential Impacts of a warming climate on water availability in snow-dominated regions, Nature, 438, 303–309, doi:10.1038/nature04141, 2005.

Bartelt, P. and Lehning, M.: A physical SNOWPACK model for the Swiss avalance warning: Part I: numerical model, Cold Regions Science and Technology, 35, 123–145, doi:10.1016/S0165-232X(02)00074-5, 2002.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, International Statistical Review, 71, 223–241, doi:10.1111/j.1751-5823.2003.tb00194.x, 2003.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. and Binley, A.: The Future of Distributed Models: Model Calibration and Uncertainty Prediction, Hydrological Processes, 6, 279–298, doi:10.1002/hyp.3360060305, 1992.

Blöschl, G.: Scaling issues in snow hydrology, Hydrological Processes, 13, 2149–2175, doi:10.1002/(SICI)1099-1085(199910)13:14/15<2149::AID-HYP847>3.0.CO;2-8, 1999.

Boike, J., Roth, K., and Ippisch, O.: Seasonal snow cover on frozen ground: Energy balance calculations of a permafrost site near Ny-Ålesund, Spitsbergen, Journal of Geophysical Research, 108, 1–11, doi:10.1029/2001JD000939, 2003.

Boike, J., Ippisch, O., Overduin, P. P., Hagedorn, B., and Roth, K.: Water, heat and solute dynamics of a mud boil, Spitsbergen, Geomorphology, 95, 61–73, doi:10.1016/j.geomorph.2006.07.033, 2008.

Boike, J., Juszak, I., Lange, S., Chadburn, S., Burke, E., Overduin, P., Roth, K., Ippisch, O., Bornemann, N., Stern, L., Gouttevin, I., Hauber, E., and Westermann, S.: A 20-year record (1998-2017) of permafrost, active layer, and meteorological conditions at a High Arctic permafrost research site (Bayelva, Spitsbergen): an opportunity to validate remote sensing data and land surface, snow and permafrost models, Earth System Science Data Discussions, pp. 1–86, doi:10.5194/essd-2017-100, 2017.

Bruland, O., Sand, K., and Killingtveit, Å.: Snow Distribution at a High Arctic Site at Svalbard, Hydrology Research, 32, 1–12, 2001.

Cannone, N., Augusti, A., Malfasi, F., Pallozzi, E., Calfapietra, C., and Brugnoli, E.: The interaction of biotic and abiotic factors at multiple spatial scales affects the variability of CO_2 fluxes in polar environments, Polar Biology, 39, 1581–1596, doi:10.1007/s00300-015-1883-9, 2016.

Charrois, L., Cosme, E., Dumont, M., Lafaysse, M., Morin, S., Libois, Q., and Picard, G.: On the assimilation of optical reflectances and snow depth observations into a detailed snowpack model, The Cryosphere, 10, 1021–1038, doi:10.5194/tc-10-1021-2016, 2016.

Clark, M. P., Slater, A. G., Barret, A. P., Hay, L. E., McCabe, G. J., Rajagopalan, B., and Leavesley, G. H.: Assimilation of snow covered area information into hydrologic and land-surface models, Advances in Water Resources, 29, 1209–1221, doi:10.1016/j.advwatres.2005.10.001, 2006.

Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., Kerr, T., Hreinsson, E. O., and Woods, R. A.: Representing spatial variability of snow water equivalent in hydrologic and land-surface models: A review, Water Resources Research, 47, 1–23, doi:10.1029/2011WR010745, 2011.

Cline, D. W., Bales, R. C., and Dozier, J.: Estimating the spatial distribution of snow in mountain basins using remote sensing and energy balance modeling, Water Resources Research, 34, 1275–1285, doi:10.1029/97WR03755, 1998.

Cortés, G. and Margulis, S.: Impacts of El Niño and La Niña on interannual snow accumulation in the Andes: Results from a high-resolution 31 year reanalysis, Geophysical Research Letters, 44, 6859–6867, doi:10.1002/2017GL073826, 2017.

Cortés, G., Girotto, M., and Margulis, S. A.: Analysis of sub-pixel snow and ice extent over the extratropical Andes using spectral unmixing and historical Landsat imagery, Remote Sensing of Environment, 141, 64–78, doi:10.1016/j.rse.2013.10.023, 2014.

Cortés, G., Girotto, M., and Margulis, S.: Snow process estimation over the extratropical Andes using a data assimilation framework integrating MERRA data and Landsat imagery, Water Resources Research, 52, 2582–2600, doi:10.1002/2015WR018376, 2016.

De Lannoy, G. J. M., Rechile, R. H., Houser, P. R., Arsenault, K. R., Verhoest, N. E. C., and Pauwels, V. R. N.: Satellite-scale snow water equivalent assimilation into a high-resolution land surface model, Journal of Hydrometeorology, 11, 352–369, doi:10.1175/2009JHM1192.1, 2010.

De Lannoy, G. J. M., Reichle, R. H., Arsenault, K. R., Houser, P. R., Kumar, S., Verhoest, N. E. C., and Pauwels, V. R. N.: Multiscale assimilation of Advance Microwave Scanning Radiometer-EOS snow water equivalent and Moderate Resolution Imaging Spectroradiometer snow cover fraction observations in northern Colorado, Water Resources Rresearch, 48, 1–17, doi:10.1029/2011WR010588, 2012.

Dee, D. et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, doi:10.1002/qj.828, 2011.

Dozier, J., Bair, E. H., and Davis, R. E.: Estimating the spatial distribution of snow water equivalent in the world's mountains, Wiley Interdisciplinary Reviews: Water, doi:10.1002/wat2.1140, 2016.

Drusch, M. et al.: Sentinel-2:ESA's Optical High-Resolution Mission for GMES Operational Services, Remote Sensing of Environment, 120, 25–36, doi:10.1016/j.rse.2011.11.026, 2012.

Dunne, S. and Entekhabi, D.: An ensemble-based reanalysis approach to land data assimilation, Water Resources Research, 41, W02013, doi:10.1029/2004WR003449, 2005.

Durand, M., Molotoch, N. P., and Margulis, S. A.: A Bayesian approach to snow water equivalent reconstruction, Journal of Geophysical Research, 113, D20117, doi:10.1029/2008JD009894, 2008.

Dutra, E., Balsamo, B., Viterbo, P., Miranda, P., Beljaars, A., Schär, C., and Elder, K.: An Improved Snow Scheme for the ECMWF Land Surface Model: Description and Offline Validation, Journal of Hydrometeorology, 11, 899–916, doi:10.1175/2010JHM1249.1, 2010.

Emerick, A. A. and Reynolds, A. C.: Ensemble smoother with multiple data assimilation, Computers & Geosciences, 55, 3–15, doi:10.1016/j.cageo.2012.03.011, 2013.

Esau, I., Argentini, S., Przybylak, R., Repina, I., and Sjöblom, A.: Svalbard Meteorology, Advances in Meteorology, 2012, 1–4, doi:10.1155/2012/818473, 2012.

Evensen, G.: Data Assimilation: The Ensemble Kalman Filter, Springer-Verlag Berlin Heidelberg, doi:10.1007/978-3-642-03711-5, 2009.

Fiddes, J. and Gruber, S.: TopoSCALE v.1.0: downscaling gridded climate data in complex terrain, Geoscientific Model Development, 7, 387–405, doi:10.5194/gmd-7-387-2014, 2014.

Førland, J. E., Benestad, R., Hanssen-Bauer, I., Haugen, J. E., and Skaugen, T. E.: Temperature and Precipitation Development at Svalbard 1900-2100, Advances in Meteorology, 2011, 1–14, doi:10.1155/2011/893790, 2011.

Foster, J. L., Sun, C., Walker, J. P., Kelly, R., Chang, A., Dong, J., and Powell, H.: Quantifying the uncertainty in passive microwave snow water equivalent observations, Remote Sensing of Environment, 94, 187–203, doi:10.1016/j.rse.2004.09.012, 2005.

Girotto, M., Cortés, G., Margulis, S. A., and Durand, M.: Examining spatial and temporal variability in snow water equivalent using a 27 year reanalysis: Kern River watershed, Sierra Nevada, Water Resources Research, 50, 6713–6734, doi:10.1002/2014WR015346, 2014a.

Girotto, M., Margulis, S. A., and Durand, M.: Probabilistic SWE reanalysis as a generalization of deterministic SWE reconstruction techniques, Hydrological Processes, 28, 3875–3895, doi:10.1002/hyp.9887, 2014b.

Gisnås, K., Westermann, S., Schuler, T., Litherland, T., Isaksen, K., Boike, J., and Etzelmüller, B.: A statistical approach to represent small-scale variability of permafrost temperatures due to snow cover, The Cryosphere, 8, 2063–2074, doi:10.5194/tc-8-2063-2014, 2014.

Gisnås, K., Westermann, S., Schuler, T. V., Melvold, K., and Etzelmüller, B.: Small-scale variation of snow in a regional permafrost model, The Cryosphere, 10, 1201–1215, doi:10.5194/tc-10-1201-2016, 2016.

Gutmann, E., Barstad, I., Clark, M., Arnold, J., and Rasmussen, R.: The Intermediate Complexity Atmospheric Research Model (ICAR), Journal of Hydrometeorology, 17, 957–973, doi:10.1175/JHM-D-15-0155.1, 2016.

Hall, D. K. and Riggs, G. A.: MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6, doi:10.5067/MODIS/MOD10A1.006, Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center, 2016a.

Hall, D. K. and Riggs, G. A.: MODIS/Aqua Snow Cover Daily L3 Global 500m Grid, Version 6, doi:10.5067/MODIS/MOD10A1.006, Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center, 2016b.

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E., and Bayr, K. J.: MODIS snow-cover products, Remote Sensing of Environment, 83, 181–194, doi:10.1016/S0034-4257(02)00095-0, 2002.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A.: Unified Notation for Data Assimilation: Operational, Sequential and Variational, Journal of the Meteorological Society of Japan, 75, 181–189, 1997.

Kępski, D., Luks, B., Migała, K., Wawrzyniak, T., Westermann, S., and Wojtuń, B.: Terrestrial Remote Sensing of Snowmelt in a Diverse High-Arctic Tundra Environment Using Time-Lapse Imagery, Remote Sensing, 9, 1–22, doi:10.3390/rs9070733, 2017.

Kohler, J. and Aanes, R.: Effect of Winter Snow and Ground-Icing on a Svalbard Reindeer Population: Results of a Simple Snowpack Model, Arctic, Antarctic and Alpine Research, 36, 333–341, doi:10.1657/1523-0430(2004)036[0333:EOWSAG]2.0.CO;2, 2004.

Kolberg, S. A. and Gottschalk, L.: Updating of snow depletion curve with remote sensing data, Hydrological Processes, 20, 2363–2380, doi:10.1002/hyp.6060, 2006.

Le, D. H., Emerick, A. A., and Reynolds, A. C.: An Adaptive Ensemble Smoother With Multiple Data Assimilation for Assisted History Matching, Society of Petroleum Engineers Journal, 21, 2195–2207, doi:10.2118/173214-PA, 2016.

Lehning, M., Völksch, I., Gustafsson, D., Nguyen, T. A., Stähli, M., and Zappa, M.: ALPINE3D: a detailed model of mountain surface processes and its application to snow hydrology, Hydrological Processes, 20, 2111–2128, doi:10.1002/hyp.6204, 2006.

Li, D., Durand, M., and Margulis, S. A.: Estimating snow water equivalent in a Sierra Nevada watershed via spaceborne radiance data assimilation, Water Resources Research, 53, 647–671, doi:10.1002/2016WR018878, 2017.

Liston, G. E.: Interrelationships among Snow Distribution, Snowmelt, and Snow Cover Depletion: Implications for Atmospheric, Hydrologic, and Ecologic Modeling, Journal of Applied Meteorology, 38, 1474–1487, 1999.

Liston, G. E.: Representing Subgrid Snow Cover Heterogeneities in Regional and Global Models, Journal of Climate, 17, 1381–1397, doi:10.1175/1520-0442(2004)017<1381:RSSCHI>2.0.CO;2, 2004.

López-Moreno, J. I., Boike, J., Sanchez-Lorenzo, A., and Pomeroy, J. W.: Impact of climate warming on snow processes in Ny-Ålesund, a polar maritime site at Svalbard, Global and Planetary Change, 146, 10–21, doi:10.1016/j.gloplacha.2016.09.006, 2016.

Luce, C. H. and Tarboton, D. G.: The application of depletion curves for parameterization of subgrid variability of snow, Hydrological Processes, 18, 1409–1422, doi:10.1002/hyp.1420, 2004.

Lüers, J., Westermann, S., Piel, K., and Boike, J.: Annual CO_2 budget and seasonal CO_2 exchange signals at a high Arctic permafrost site on Spitsbergen, Svalbard archipelago, Biogeosciences, 11, 6307–6322, doi:10.5194/bg-11-6307-2014, 2014.

Magnusson, J., Winstral, A., Stordal, A. S., Essery, R., and Jonas, T.: Improving physically based snow simulations by assimilating snow depths using the particle filter, Water Resources Research, 53, 1–19, doi:10.1002/2016WR019092, 2017.

Margulis, S. A., Girotto, M., Cortés, G., and Durand, M.: A Particle Batch Smoother Approach to Snow Water Equivalent Estimation, Journal of Hydrometeorology, 16, 1752–1772, doi:10.1175/JHM-D-14-0177.1, 2015.

Margulis, S. A., Cortés, G., Girotto, M., and Durand, M.: A Landsat-Era Sierra Nevada Snow Reanalysis (1985-2015), Journal of Hydrometeorology, 17, 1203–1221, doi:10.1175/JHM-D-15-0177.1, 2016.

Martinec, J. and Rango, A.: Areal Distribution of Snow Water Equivalent Evaluated by Snow Cover Monitoring, Water Resources Research, 17, 1480–1488, doi:10.1029/WR017i005p01480, 1981.

Molotch, N. P. and Margulis, S. A.: Estimating the distribution of snow water equivalent using remotely sensed snow cover data and a spatially distributed snowmelt model: A multi-resolution, multi-sensor comparison, Advances in Water Resources, 31, 1503–1514, doi:10.1016/j.advwatres.2008.07.017, 2008.

Niu, G. Y., Seo, K. W., Yang, Z. L., Wilson, C., Su, H.and Chen, J., and Rodell, M.: Retrieving snow mass from GRACE terrestrial water storage change with a land surface model, Geophysical Research Letters, 34, 1–5, doi:10.1029/2007GL030413, 2007.

Nowak, A. and Hodson, A.: Hydrological response of a High-Arctic catchment to changing climate over the past 35 years: a case study of Bayelva watershed, Svalbard, Polar Research, 32, 1–17, doi:10.3402/polar.v32i0.19691, 2013.

NPI: Norwegian Polar Intsitute DEM, doi:10.21334/npolar.2014.dce53a47, 2014.

Østby, T. I., Schuler, T. V., Hagen, J. O., Hock, R., Kohler, J., and Reijmer, C. H.: Diagnosing the decline in climatic mass balance of glaciers in Svalbard over 1957-2014., The Cryosphere, 40, 191–215, doi:10.5194/tc-11-191-2017, 2017.

Painter, T. H., Rittger, K., McKenzie, C., Slaughter, P., Davis, R. E., and Dozier, J.: Retrieval of subpixel snow covered area, grain size and, and albedo from MODIS, Remote Sensing of Environment, 113, 868–879, doi:10.1016/j.rse.2009.01.001, 2009.

Pedersen, C. and Winther, J.: Intercomparison and validation of snow albedo parameterization schemes in climate models, Climate Dynamics, 25, 351–362, doi:10.1007/s00382-005-0037-0, 2005.

Peng, J., Liu, Q., Wang, L., Liu, Q., Fan, W., Lu, M., and Wen, J.: Characterizing the Pixel Footprint of Satellite Albedo Products Derived from MODIS Reflectance in the Heihe River Basin, China, Remote Sensing, 7, 6886–6907, doi:10.3390/rs70606886, 2015.

Raleigh, M. S., Lundquist, J. D., and Clark, M. P.: Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework, Hydrology and Earth System Sciences, 19, 3153–3179, doi:10.5194/hess-19-3153-2015, 2015.

Roth, K. and Boike, J.: Quantifying the thermal dynamics of a permafrost site near Ny-Ålesund, Svalbard, Water Resources Research, 37, 2901–2914, doi:10.1029/2000WR000163, 2001.

Salomonson, V. V. and Appel, I.: Estimating fractional snow cover from MODIS using the normalized difference snow index, Remote Sensing of Environment, 89, 351–360, doi:10.1016/j.rse.2003.10.016, 2004.

Slater, A. G. and Clark, M. P.: Snow Data Assimilation via an Ensemble Kalman Filter, Journal of Hydrometeorology, 7, 478–493, 2006.

Slater, A. G., Barrett, A. P., Clark, M. P., Lundquist, J. D., and Raleigh, M. S.: Uncertainty in seasonal snow reconstruction: Relative impacts of model forcing and image availability, Advances in Water Resources, 55, 165–177, doi:10.1016/j.advwatres.2012.07.006, 2013.

Smith, R. B. and Barstad, I.: A Linear Theory of Orographic Precipitation, Journal of the Atmospheric Sciences, 61, 1377–1391, doi:10.1175/1520-0469(2004)061<1377:ALTOOP>2.0.CO;2, 2004.

Stordal, A. S. and Elsheikh, A. H.: Iterative ensemble smoothers in the annealed importance sampling framework, Advances in Water Resources, 86, 231–239, doi:10.1016/j.advwatres.2015.09.030, 2015.

Sturm, M. and Wagner, A. M.: Using repeated patterns in snow distribution modeling: An Arctic Example, Water Resources Research, 46, 1–15, doi:10.1029/2010WR009434, 2010.

Su, H., Yang, Z. L., Dickinson, R. E., Wilson, C. R., and Niu, G. Y.: Multisensor snow data assimilation at the continental scale: The value of Gravity Recovery and Climate Experiment terrestrial water storage information, Journal of Geophysical Research, 115, doi:10.1029/2009JD013035, 2010.

Tarboton, D. G. and Luce, C. H.: Utah Energy Balance Snow Accumulation and Melt Model, Computer model techincal description and users guide, 1996.

Van Leeuwen, P. J.: Particle Filtering in Geophysical Systems, Monthly Weather Review, pp. 4089–4114, doi:10.1175/2009MWR2835.1, 2009.

Van Leeuwen, P. J. and Evensen, G.: Data assimilation and inverse methods in terms of a probabilistic formulation, Monthly Weather Review, 124, 2898–2913, doi:10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2, 1996.

Vionnet, V., Brun, E., Morin, S., Boone, A., Faroux, S., Le Moigne, P., Martin, E., and Willemet, J. M.: The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2, Geoscientific Model Development, 5, 773–791, doi:10.5194/gmd-5-773-2012, 2012.

Westermann, S., Lüers, J., Langer, M., Piel, K., and Boike, J.: The annual surface energy budget of a high-arctic permafrost site on Svalbard, Norway, The Cryosphere, 3, 245–263, doi:10.5194/tc-3-245-2009, 2009.

Westermann, S., Boike, J., Langer, M., Schuler, T. V., and Etzelmüller, B.: Modeling the impact of wintertime rain events on the thermal regime of permafrost, The Cryosphere, 5, 945–959, doi:10.5194/tc-5-945-2011, 2011a.

Westermann, S., Langer, M., and Boike, J.: Spatial and temporal variations of summer surface temperatures of high-arctic tundra on Svalbard - Implications for MODIS LST based permafrost monitoring, Remote Sensing of Environment, 115, 908–922, doi:10.1016/j.rse.2010.11.018, 2011b.

Westermann, S., Langer, M., and Boike, J.: Systematic bias of winter-time land surface temperatures inferred from MODIS at a site on Svalbard, Norway, Remote Sensing of Environment, 118, 162–167, doi:10.1016/j.rse.2011.10.025, 2012.

Westermann, S., Boike, J., Guglielmin, M., Gisnås, K., and Etzmüller, B.: Snow melt monitoring near Ny-Ålesund, Svalbard, using Automatic Camera Systems, doi:10.1594/PANGAEA.846617, *Department of Geosciences, University of Oslo*, 2015a.

Westermann, S., Østby, T. I., Gisnås, K., Schuler, T. V., and Etzelmüller, B.: A ground temperature map of the North Atlantic permafrost region based on remote sensing and reanalysis data, The Cryosphere, 9, 1303–1319, doi:10.5194/tc-9-1303-2015, 2015b.

Westermann, S., Langer, M., Boike, J., Heikenfeld, M., Peter, M., Eztelmüller, B., and Krinner, G.: Simulating the thermal regime and thaw processes of ice-rich permafrost ground with the land-surface model CryoGrid 3, Geoscientific Model Development, 9, 523–546, doi:10.5194/gmdd-8-6931-2015, 2016.

Westermann, S., Peter, M., Langer, M., Schwamborn, G., Schirrmeister, L., Etzelmüller, B., and Boike, J.: Transient modeling of the ground thermal conditions using satellite data in the Lena River delta, Siberia, The Cryosphere, 11, 1441–1463, doi:10.5194/tc-11-1441-2017, 2017.

Winther, J. G., Godtliebsen, F., Gerland, S., and Isachsen, P. E.: Surface albedo in Ny-Ålesund, Svalbard: variability and trends during 1981-1997, Global and Planetary Change, 32, 127–139, doi:10.1016/S0921-8181(01)00103-5, 2002.

You, J., Tarboton, D., and Luce, C.: Modeling the snow surface temperature with a one-layer energy balance snowmelt model, Hydrology and Earth System Sciences, 18, 5061–5076, doi:10.5194/hess-18-5061-2014, 2014.