Reviewer # 1

Comment # 1.1

General Comments:

Overall, this is an excellent paper worthy of publication in GMD. The topic of snow data assimilation is of high scientific importance and providing a unifying framework for implementing such methods is to be commended and should be of high value to the community. The structure of the manuscript and presentation are generally clear. There are many specific comments listed below which aim to improve the communication of the proposed framework and aid users in its implementation and use. In particular, more details on the sample problems would aid in the reproducibility and extension of the work to other problems. In testing the code, it appears that the github repository code works, but the sample provided on zenodo has a bug. More details are provided below.

Reply:

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. References are included at the end of this document.

Comment # 1.2

Specific Comments:

The comments provide herein represent a list of relatively minor additions and/or corrections that would improve the paper.

1. *Title: The usage of "Multiscale" in the title does not seem particularly warranted. My expectation based on the title was that the implementation would be flexible enough to model snow at multiple scales (resolutions) and/or assimilate data at multiple scales (resolutions). It is not clear from the presentation whether either is the case. Or is the meaning meant to convey multiple temporal scales? The authors should consider whether the title should be changed for clarity. If there is an aspect of what you are proposing that is indeed "multiscale" you should emphasize that more for the reader's benefit.*

Reply:

We would like to clarify that MuSA can be run at a range of spatial resolutions, as shown in the manuscript from km to m scale. Assimilating multiscale observations is also possible, just that these are usually pre-processed (to the spatial geometry of the model grid) before being fed into MuSA. In terms of temporal scales of assimilation it is also quite flexible since it builds on the concept of data assimilation windows. We nonetheless agree with the reviewer that the term Multiscale can be misleading or at least too narrow and have thus opted for the term Multiple. This change more clearly highlights that the toolbox can assimilate multiple observations at multiple spatiotemporal scales using multiple data assimilation schemes. The name of our toolbox (and thus the title) has thus been changed to *The Multiple Snow Data Assimilation System* while the acronym MuSA remains unchanged.

COMMENT # 1.3

2. *Line 85: In mentioning the "posterior mean snow simulation from FSM2" it would be useful to know what variables that contains. Maybe the FSM2 variables could be shown in a Table?*

Reply:

The output of FSM2 contains the entire physical state of the snowpack as represented by this intermediate complexity model. Moreover, in MuSA it is simple to output more variables (such as fluxes) than the ones that are defined by default. This is relevant, as if other variables are selected for being output it will be possible to assimilate them directly. Some modifications are expected in the near future, this is the reason why we did not originally include it in the manuscript (it is in the Github repository). After considering the suggestion of the referee, we have included the following.

Changes:

2. Overview of the data assimilation system

...

In its current version, MuSA is able to assimilate the following variables:

- SWE (mm)
- Snow depth (m)
- Land surface temperature (K)
- Fractional snow-covered area (-)
- Albedo (-)

- Sensible heat flux to the atmosphere (W m^{-2})
- Latent heat flux to the atmosphere (W m^{-2})"

We expect to provide support for even more variables in the future.

COMMENT # 1.4

3. Line 132: In mentioning the additive/multiplicative perturbations there is no description as to whether they are perfectly independent or perfectly correlated or something in between. In other words, do all pixels get the same perturbation (i.e. from the same random number) or do they get fully independent perturbations (i.e. each sampled independently). Mentioning here or elsewhere that this neglects spatially-correlated errors/uncertainties would be appropriate. It is mentioned earlier that the model structure is fully independent, but saying whether the perturbations are as well would clarify the setup.

Reply:

Thanks for the suggestion, we have now clarified this in the text.

Changes:

2.1 Ensemble generation

...

The perturbation of the forcing is performed by drawing spatially independent random perturbation parameters from a...

COMMENT # 1.5

4. Line 413: In describing the time-invariant perturbations it may be worth mentioning what the implications of that are vs. other options (independent in time or correlated in time).

Reply:

The implications of using time-invariant perturbations has been added to the text following the reviewer's suggestion.

Changes:

3.3 Consistency

...

~~Thereby~~As noted by (1), when these perturbation parameters are interpreted as errors this can be viewed as a limiting case of perfect time-correlation where the errors become constant biases. The lack of dynamics in the perturbation parameters is thus akin to assuming that errors in the forcing are constant in time within a particular water year. This simplifying assumption imposes a longer memory in the system than with jitter and facilitates the propagation of information backwards in time using smoothers. Given this transition density, the full prior for the entire DAW can be factorized as follows…

COMMENT # 1.6

5. *Line 654: I think a couple of sentences describing the mechanics of the iterative nature of the method and why it outperforms other methods would be warranted.*

Reply:

The mechanics of the iterative nature of the MDA method are now described in Section 3.5.

Changes:

3.5 Ensemble Kalman methods

…

In this iterative scheme the prior ensemble moves gradually to the posterior ensemble through a tempering procedure (see 2, and references therein). The iterations thus mitigate the impact of the linearity assumption inherent in ensemble Kalman methods (3), typically leading to marked improvements compared to the ES for nonlinear models without the risk of degeneracy associated with particle methods as the curse of dimensionality rears its head (4; 5).

COMMENT # 1.7

6. *Lines 696-703: More explanation of what it meant by "inflated observation errors" and how it fits into the method would be helpful to the reader. An elaboration on the note about the "multiple data assimilation approach does not actually violate" would also be helpful. Since this particular method is less standard than others, I assume most readers would benefit from more detail here.*

Reply:

We have now thoroughly explained what is meant by inflated observation errors. More generally, we have also elaborated on the idea behind the MDA schemes.

Changes:

3.5 Ensemble Kalman methods

...

Recall that for $N_a = 1$ we recover the non-iterative stochastic EnKF and ES, while for $N_a > 1$ we are using iterative versions of these schemes that involve multiple data assimilation with inflated observation errors. ~~This~~ The term "multiple data assimilation" refers to the assimilation of the same data multiple times rather than an assimilation of different types of data (joint assimilation). We speak of inflated observation errors since the role of the coefficients $\alpha^{(\ell)}$ is to inflate the observation error covariance $\mathbf{R}$ in the Kalman gain $\mathbf{K}^{\ell}$ as well as the observation error term $\varepsilon_{\alpha}^{(\ell)}$. This inflation is tantamount to tempering the likelihood as discussed by ~~(2; 6)~~ (2) and (6) which explains why these iterative schemes perform better than their non-iterative counterparts for non-linear models in that they involve a more gradual transition from the prior to the posterior. Despite what the name might suggest, this multiple data assimilation approach does not actually violate the consistency of Bayesian inference by using the data more than once due to the way the observation error inflation is constructed, particularly due to the constraint that $\cancel{\sum_{\ell=1}^{N_a]} = 1}$ $\sum_{\ell=1}^{N_a} 1/\alpha^{(\ell)} = 1$. Simply stated, this constraint ensures consistent results with a linear model since by construction we get the same result by assimilating the data once with the original uninflated ($\alpha = 1$) observation errors as assimilating the same data multiple times with an inflated ($\alpha > 1$) observation errors. With a nonlinear model, practice has shown that these iterations of the ensemble Kalman analysis ensure that the approximate posterior is closer to the true posterior than if a more conventional single uninflated iteration is used (3). It is possible to satisfy ~~this constraint~~ the constraint on the $\alpha^{(\ell)}$ both with uniform and non-uniform inflation coefficients (c.f. 1). For simplicity, following (7), we currently opt for former as a default in MuSA by setting $\alpha^{(\ell)} = N_a$ ~~set $\alpha^{(\ell)} = N_a$~~ for all ℓ by default while allowing for ~~the latter~~ non-uniform coefficients as an option.

COMMENT # 1.8

7. *Line 741: Can you provide some justification of the choice of four (4) for the number of assimilation cycles?*

Reply:

In general, the number of iterations of MDA N_a is a compromise between computational cost, since each iteration adds N_e simulations, and accuracy, since the performance (in terms of approximating the true posterior) improves as N_a increases. For the latter, however, the improvement is usually asymptotic. As such there is usually an optimal choice of N_a that balances the quality of the posterior ensemble approximation against the cost of the iterative ensemble of model runs. Experience (e.g. 8; 3), has shown that $N_a = 4$ tends to be a satisfactory choice. However, with any particular model this is usually worth exploring in a sensitivity analysis. Since this was not the main topic of this paper, such an exploration was not pursued herein. In principle, N_a can is set by the user and should be viewed as a tuneable hyperparameter by taking the aforementioned compromise into account.

Changes:

4.1 Single cell and distributed assimilation of drone-based snow depth retrievals

...

In the iterative versions of the ensemble Kalman based approaches we fixed the number of assimilation cycles to $N_a = 4$ as a compromise between computational cost and performance. The former is directly proportional to N_a while the latter converges to an optimum as N_a increases. This choice of N_a is also in line with sensitivity analyses performed elsewhere (7; 8; 3)...

COMMENT # 1.9

8. Section 4: I would urge that more consistent (and maybe simpler) language be used throughout to refer to the three experiments being done so that readers can follow more easily. It doesn't seem that "drone data" or "satellite data" are as relevant to the first two experiments compared to the first being a spatially-distributed (snow depth) data assimilation experiment and the second being a point-scale joint (FSCA+LST) data assimilation experiment. For the benchmark case, it is not clearly defined what experiment is actually being done. Is it snow depth or LST+FSCA assimilation? There is a mention of "single cell" which may imply it is the same setup as the second experiment, but this is not clear as currently presented.

Reply:

In an effort to homogenize the section titles and make them clearer as suggested by the reviewer, we have changed the section titles. In particular, Section 4.1&5.1 are now named **Single cell and distributed assimilation of drone-based snow depth retrievals** and Section 4.2&5.2 are now named **Joint assimilation of satellite-based LST and FSCA retrievals** In the text, we have also emphasized that the benchmark was conducted using drone data in the Izas catchment.

Changes:

4.3 Computational benchmarks

...

The comparison was performed using 100, 200 and 300 particles and four iterations for the iterative ensemble Kalman approaches in a single cell, assimilating drone snow depth retrievals at a random location in the Izas catchment. The reported values of the benchmarks are the average of 10 MuSA runs, and includes the FSM2 compilation time (≃2 seconds using the GNU Fortran compiler 10.3.0 in the aforementioned local machine) which is negligible compared with the whole run.

COMMENT # 1.10

9. *Section 4: In an effort to make the sample experiments more reproducible for the readers, I would suggest tabulating any key parameter differences (beyond default values) in the config.py and/or constants.py input files that are specific to each experiment being done. It would also be helpful to connect the individual experiments to the theory provided earlier in the paper, i.e. description of the states, measurement, etc. In particular, if transforms are used with respect to the measurements (as referred to on Lines 359-370), it would be useful to see the form of the those transforms in the experimental setup in Section 4.*

Reply:

Thanks for the suggestions. We have added a config.py file to the Zenodo repository with different configuration suggestions. It should be noted that the constants.py file is the same for all experiments and is stored on github. In its current versions the transformations are rather simple, and therefore although we considered adding a new figure to visualize these we thought that might be a bit excessive. Instead, as an example, we have generated this simple example in Python using Google Colab that you can experiment with online:

Simple example of Gaussian anamorphosis (hyperlink)

COMMENT # 1.11

10. *Section 4: Perhaps in each case you can explain what the measurement model is for that experiment, i.e., is it just an internal model state (snowdepth, LST?) or a prescribed diagnostic relationship (FSCA?). In cases where it is a prescribed diagnostic relationship, how is that handled within the framework? I imagine that the current FSCA is built-in to FSM2, but what if an alternative representation was desired. Would that be handled via modification of the FSM2 snow model, or via another method.*

Reply:

At the moment, all the variables that can potentially be assimilated are calculated as state variables in FSM2. In the specific case of FSCA, if a different representation is desired, we would recommend implementing it directly in FSM2 as this variable interacts with several components of the surface energy balance in the model. In any case, it is very easy to include new variables derived from FSM2 outputs (e.g. statistics such as snow cover frequency, season length, etc.), and this could be done in MuSA without having to modify the FSM2 code, or with minimal modifications to it in case other internal states of the model are needed.

Changes:

4. Data and experimental setup

...

~~First, we~~ Note that all the variables that we assimilate in these experiments are state variables in FSM2. We ...

COMMENT # 1.12

11. *Section 5: I found the organization of Section 4 flowing into Section 5.1 hard to follow. As mentioned above, I would suggest using the same language to refer to the three experiments throughout to help in this regard. It wasn't clear to my why the benchmark (single cell) and distributed snow depth results were presented together in Section 5.1. They are described as two different experiments in Section 4 and so I think it would be easier to follow if they were treated as such in Section 5. Lumping them together in 5.1 seems a bit disjointed. Or maybe "single cell" here does not refer to the benchmark case (although "single cell" is used in that context too)?*

Reply:

As previously mentioned, and following the suggestion of the reviewer, we have changed the names of these sections. "Single cells" simply refers to the fact that MuSA has been launched at a specific single grid cell in the Izas domain. For the drone-based snow depth retrievals the single cell is at a the resolution of 5 m and used to compare algorithms (Sections 4.1&5.1) as well as computational cost (Sections 4.3&5.3). For the joint assimilation of satellite-based FSCA and LST retrievals the single cell is at a much coarser resolution of 1 km as described in (Sections 4.2&5.2). This should be clearer now that we have renamed the sections.

COMMENT # 1.13

12. *With respect to Table 1, it is not clear what the reference data being used to compute RMSE is. It implies snow depth, but the description of what data was assimilated in the benchmark experiments is unclear (see comment above). The notation used for each scheme is also not defined. Perhaps define PF-c, PF-r, in caption?*

Reply:

This has been clarified as outlined below. In addition, we have recalculated the metrics in Table 1, since upon reviewing the code we found a bug in the RMSE calculation. In any case, the relative performance of each algorithm has not changed, so the conclusions remain the same.

Changes:

Table 1 caption
~~Evaluation metrics~~ RMSE for the reference run (Ref), particle filter with bootstrap resampling (PF-b), particle filter with redraw resampling (PF-r), ensemble Kalman filter (EnKF), ensemble Kalman filter with multiple data assimilation (EnKF-MDA), particle batch smoother (PBS), ensemble smoother (ES), and ensemble smoother with multiple data assimilation (ES-MDA). These errors were computed using the assimilated drone-based snow depth observations as the truth and using the posterior ensemble mean ~~based on~~ as the ~~assimilated observations~~estimate from the respective DA schemes. All the schemes were run with $N_e = 200$ particles and the MDA schemes used $N_a = 4$ iterations...

...

5.1 Single cell and distributed assimilation of drone-based snow depth retrievals
The results show how the performance of the different data assimilation algorithms differs, even with the same initial conditions and experimental setup~~(Figure 3 and Figure 4, single cell test comparison~~, when the posterior ensembles are compared against the assimilated snow depth observations (Table 1, Figures 3&4)...

COMMENT # 1.14

13. *In the context of Figure 3, it would be helpful to explain the meaning of "MDA" when only snow depth is being assimilated. I believe this method actually differs in this case due to its iterative nature rather than multi-data? This comes into play later where different notation is used to refer to iterative versions of method. Perhaps you can harmonize how you refer to iterative methods across the manuscript.*

Reply:

As pointed out in an earlier reply, "Multiple data assimilation" (MDA) is synonymous with iterative in this context and refers to the fact that the these ensemble Kalman with MDA methods are composed of a number of gradual transitions rather than a single abrupt movement of the ensemble. As such, they assimilate the same data multiple times but with a constrained inflation of the observation error to avoid a circular analysis (double dipping) as explained in response to Comment #1.7. This MDA method is applicable regardless of whether one or several variables are assimilated at the same time (joint assimilation). The nature of the assimilated observations does not influence the method per se, so the fact that only snow depth is being assimilated does not have any bearing on the use of MDA.
As suggested by the reviewer, we have now homogenised the nomenclature throughout the text and figures by using multiple data assimilation (MDA) instead of Iterative (I).

COMMENT # 1.15

14. *Figure 3. Refer to which experiment this corresponds to. And is this a particular cell? Is it the one shown in Figure 2?*

Reply:

Yes, it is in the particular cell highlighted in Figure 2. We have clarified this by adding the following in Figures 3&4:

Changes:

Figure 3&4 captions
. . . in the single cell of the Izas catchment highlighted in Figure 2. . .

COMMENT # 1.16

15. *Line 797: The reader would benefit from more description of how the prior forcing perturbations are generated in this context and how the posterior emerges from that. Can you clarify whether prior was identical across space and why patterns in the posterior emerge. Is there anything to be learned from the posterior uncertainty of these, i.e. is one more certain than the other (i.e. precip. vs. temperature? And why are the posterior patterns between the two fields seemingly so highly correlated. More discussion either here or in Section 6 would benefit the reader.*

Reply:

This is discussed below, where we hypothesise that the spatial patterns emerge from the fact that in the reference simulation there is no implicit representation of wind redistribution. That redistribution is implicitly induced in the simulations by assimilating the drone snow depth maps, allowing MuSA to generate these spatial patterns. It is difficult to estimate, and outside the scope of this paper, whether one parameter is more uncertain than the other. It is likely that there is some equifinality as well. This problem could be addressed by jointly assimilating more variables related to the energy balance such as LST and albedo. In any case, such a study would require specific work to be fully relevant, in the same way that we try not to dwell on the intercomparison between algorithms as these are topics that deserve much more attention than is given here. Here we try to limit ourselves to describing the capabilities of MuSA by means of examples that may be of interest, and hopefully a source of inspiration, to future users.

COMMENT # 1.17

16. *Discussion associated with Figure 6. Indicate that the fields in Figure 6 are the posterior mean. Units should be associated with temperature. Could more discussion be provided to hypothesize why the patterns are what they show.*

Reply:

We have added the units to the caption, thanks for the suggestion. As for the spatial patterns, in the manuscript (see **6. Discussion**) we hypothesise that it is a consequence of the wind redistribution patterns that are not explicitly represented in FSM2 but can be captured implicitly by perturbing the forcing. For example, the precipitation bias perturbation parameter effectively accounts for both biases in the large scale precipitation field as well as the local effects of wind redistribution of snow that occurs mainly during the accumulation season. But to broaden the discussion in this sense would be somewhat speculative and, we believe, outside the scope of this paper.

COMMENT # 1.18

17. *Figure 7: There are inconsistencies (and typos.) between the use of what should be "LST" in the caption and "SST" in the figure. Is SST meant to be "snow surface temperature". If that is preferred, SST should be used throughout instead of LST. The acronym "IKS" should be defined in the caption.*

Reply:

We have corrected these acronyms such that they are consistent throughout the manuscript, they are now always LST and ES-MDA. Thank you for pointing this out.

COMMENT # 1.19

18. *Figure 8: Acronyms need to be defined in the caption and reconciled with earlier ones. How does the Ensemble Smoother – MDA compare to any of these? Is it the same as I-ES?*

Reply:

These acronyms have now been homogenised in the text and figures, thank you.

COMMENT # 1.20

19. *Line 869: It is not clear what is meant by: "The assimilation of the FSCA provides information when FSCA saturates at 1, . . . ". Should this read ". . . does not provide information"?*

Reply:

The first use of FSCA in this sentence was a typo, it should have said LST. It has now been corrected.

Changes:

6. Discussion

. . .

The assimilation of ~~the FSCA provides~~ LST has the potential to provide additional information when FSCA saturates at 1, ~~i.e.~~ for example during most of the ~~snow season in snow dominated areas, or~~ accumulation season and during the polar night in the absence of ~~light~~sunlight.

COMMENT # 1.21

20. *Code and data availability: It seems that the MuSA code from the original github repository vs. the version provided on zenodo are different. In particular, when run on a mac, the github version worked, while the zenodo version did not. It appears to center on differences in the code, where the latter crashed out when finding the OS to be 'darwin' (macOS) instead of 'linux'. I suggest making sure to reconcile the two so that the one posted on zenodo works. It would also be helpful for reproducing the results to 1) tabulate key parameters specific to each experiment (as suggested above) and 2) providing the actual input files for each experiment with the code distribution. This would make it much easier to reproduce the results from the paper and extend the framework to other cases rather than having to interpret which* *parameters to change.*

Reply:

This error comes as a surprises to us, and we have not been able to find where it comes from. It is true that MuSA checks the operating system, and is not to be used if darwin (macOS) or linux is not identified. Note that for Windows users we have now tried MuSA successfully using the Windows Subsystem for Linux (WSL). The OS checking function is identical in the Zenodo and Github version, which makes sense as the copying is done automatically since both repositories are connected. In any case, small differences between Zenodo and Github will always occur, this is expected behaviour. The goal of using Zenodo is to be able to provide a given release with a unique DOI. But all recent activity on Github that isn't part of a release will be out of sync with Zenodo, this is unavoidable. We have added the LST and FSCA observations from MODIS to the repository as suggested, as well as a tabulated configuration file.

COMMENT # 1.22

Technical Corrections:
This is not an exhaustive list of typos, but ones that jumped out:
1. *In Figure 1 there is a typo., where "weigths" should instead be "weights".*

Reply:

Corrected, thanks.

COMMENT # 1.23

2. *Line 205: Typo. in the phrase "the are usually".*

Reply:

We have corrected the typo and restructured the sentence for clarity.

Changes:

3.1 Bayesian inference
. . . ~~As~~ These are usually probability density rather than mass functions as we tend to deal with continuous variables in DA~~, the are usually probability density rather than mass functions~~

COMMENT # 1.24

3. *Line 383: Typo. in the phrase "we will let denote anamorphosed".*

Reply:

This has been corrected, a **u** was missing

Changes:

3.3. Consistency

...

As such, we will let **u** denote anamorphosed parameters that have undergone a forward transform to the unbounded space.

COMMENT # 1.25

4. *Line 391: "SM2" should be "FSM2".*

Reply:

Corrected.

COMMENT # 1.26

5. *Line 491: "converege" should be "converge".*

Reply:

Corrected.

COMMENT # 1.27

6. *Line 790: "smothers" should be "smoothers".*

Reply:

Corrected.

REFERENCES

[1] G. Evensen, "Accounting for model errors in iterative ensemble smoothers," *Computational Geosciences*, vol. 23, p. 761–775, 2019.

[2] A. S. Stordal and A. H. Elsheikh, "Iterative ensemble smoothers in the annealed importance sampling framework," *Advances in Water Resources*, vol. 86, p. 231–239, 2015.

[3] G. Evensen, "Analysis of iterative ensemble smoothers for solving inverse problems," *Computational Geosciences*, vol. 22, p. 885–908, 2018.

[4] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, "Obstacles to High-Dimensional Particle Filtering," *Monthly Weather Review*, vol. 136, p. 4629–4640, 2008.

[5] N. Pirk, K. Aalstad, S. Westermann, A. Vatne, A. van Hove, L. Tallaksen, M. Cassiani, and G. Katul, "Inferring surface energy fluxes using drone data assimilation in large eddy simulations," *Atmospheric Measurement Techniques Discussions*, 2022.

[6] P. J. van Leeuwen, H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, "Particle filters for high-dimensional geoscience applications: A review," *Quarterly Journal of the Royal Meteorological Society*, vol. 145, p. 2335–2365, 2019.

[7] A. A. Emerick and A. C. Reynolds, "Ensemble smoother with multiple data assimilation," *Computers & Geosciences*, vol. 55, p. 3–15, 2013.

[8] K. Aalstad, S. Westermann, T. V. Schuler, J. Boike, and L. Bertino, "Ensemble-based assimilation of fractional snow-covered area satellite retrievals to estimate the snow distribution at Arctic sites," *The Cryosphere*, vol. 12, p. 247–270, 2018.