**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.

**Specific Comments:**

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

- **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.

- 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?

- 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.

- 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).

- 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.
- 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.

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

- 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.

- 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 those transforms in the experimental setup in Section 4.
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.

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)?

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?

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.

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

- 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.

- 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.

- 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.

- 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?
- 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”?

- 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.

**Technical Corrections:**

This is not an exhaustive list of typos., but ones that jumped out:

- In Figure 1 there is a typo., where “weigths” should instead be “weights”.
- Line 205: Typo. in the phrase “the are usually”.
- Line 383: Typo. in the phrase “we will let denote anamorphosed”.
- Line 391: “SM2” should be “FSM2”.
- Line 491: “converege” should be “converge”.
- Line 790: “smothers” should be “smoothers”.