General statement

This manuscript describes a study about the point-scale assimilation of land surface temperature data in an energy balance snowpack model. This is a synthetic experiment, given that there are not currently many data sets for observed land surface temperature data. The authors explain that future satellite missions should provide more data in the coming years and that it is important to be prepared to use that data. Consequently, their study aims at assessing the potential of LST to improve the representation of SWE in snowpack models. They have designed different scenarios for their synthetic data, in order to simulate the presence of clouds and also to simulate different revisiting time for the satellites taking the measures. They also compare two methods for data assimilation: the particle filter and a particle batch smoother.
In my opinion the manuscript is interesting, well-written and relevant for the readers of The Cryosphere. I have only a few minor comments that I would like the authors to address.
Author's response: We appreciate the reviewer's positive comments. Below, we provide a point-by-point response to all of the reviewer's specific and annotated comments.

Specific minor comments:
• Line 80: I am curious about OSSE, as I had never heard or read this specific appellation before. Is it a just a fancy name to refer to any form of synthetic experiment, or is it a specific type of synthetic experiment, guided by a series of principles/rules? Later in this paragraph you just refer to « synthetic experiments » . Are they interchangeable terms? Maybe uniformize?
Author's response: OSSE experiment is a widely used term in the DA community. It is a synthetic numerical experiment designed to study the impact of new types of observations on a modeling system through their assimilation. It is used by the numerical weather prediction community and others to assess the impact of new satellite observations on their modeling pipelines. Therefore, an OSSE is a type of synthetic experiment. In this work both terms can be considered interchangeable.

• Table 1: There are multiple polices in this table. Please uniformize.
Author's response: Corrected.

• Lines 113-114: There are certain methodological choices for the construction of the synthetic data sets for which I would have liked to have more detail. First, I understand that you selected the specific value for the multiplicative perturbation of precip after Beck et al. (2019), but I still would have liked to have more detail about that, and why you did not test more than one value for this multiplicative factor.
Author's response: In the new version of the paper we have further degraded the forcing, to highlight more the errors of the snow model FSM2 (see response to reviewer #1). To do so, we have included autocorrelated noise in the forcing variables. The methods section will be extended in consequence. We have not selected different values for this scaling factor, as we wanted to emulate a likely but high error value. We considered that having a 50% prior underestimation of precipitation is a sufficiently pessimistic scenario.

• *Similarly, on page 5, I would have liked to have more detail about:*

  • *Line 125-129: The trial and error process that led to the selection of mu and sigma.* Just a little bit more detail. Also, why 300 particles?

  Author's response: It is challenging to provide a justification for the use of a specific number of particles This is the reason why this justification is not often found in the DA literature (see e.g. Piazzi et al., 2019 for a sensitivity analysis). A general recommendation based on our own experience would be around 100 for most of the cases, but the more particles, the better. The computational cost increases linearly with the number of particles, while the benefit does not. The definition of the prior perturbation parameters is also subjective. There are different options. A possibility is to estimate it based on the comparison of the forcing with in situ observations. This is something that given the synthetic nature of the experiment we cannot do. Another option is to maximize the entropy, i.e. the dispersion of the ensemble. In our case, we have used probability distributions that roughly cover the expected uncertainty values.

  • *Line 139: what are the « new perturbation parameters »? Do you mean « new particles »? Or is it a more sophisticated way of preventing the collapsing of the filter?*

  Author's response: We will clarify this point in the revised manuscript. Different strategies can be found in the literature to recover ensembles from collapsing. What we do here with MuSA is only resampling the states, and generate new parameters at each analysis step (when MuSA finds an observation). At each analysis step, we calculate a normal approximation of the posterior parameters by calculating the weighted mean and standard deviation using the posterior weights. In case of a total collapse of the ensemble (i.e. the posterior standard deviation is close to zero), a fraction of the prior standard deviation is used. This updated distribution is used to create new parameters for each particle after resampling. This has two properties, the first one, that even if all the weight falls on a particle (i.e. all the particles become the same), assigning different correction parameters causes the particles to diverge again and thus the ensemble is revitalized. In addition, the normal approximation can potentially generate parameters that are better placed in the parameter space than the priors selected at the beginning of the DA process. A detailed description of the algorithm can be found in the original MuSA paper.

• *On page 6, the authors switch back and forth between the present tense and the past. I think it would be better to use the present all the time.*

Author's response: We have corrected the text in consequence following the recommendations in the annotated document. Thanks for this detailed revision.

• *Figure 2: If I understand correctly, in your synthetic experiment, the model can only underestimate SWE (the open loop is consistently inferior to the « synthetic truth »). Is this on purpose? If so, why? I have seem many cases for which the model overestimated SWE, for instance because it overestimated the density on some occasions. I think it would be important to explain why FSM2 can only underestimate SWE.*

Author's response: Indeed, we will explain that the model can only underestimate SWE because the precipitation has been drastically reduced from its counterpart which was used to generate the synthetic truth. MuSA was able to find approximately this correction parameter (in the case of smoothers), which was the objective of the test. We do not see any reason why the conclusions would change by increasing the precipitation.

• Figure 3: it would be good to add a legend.
Author's response: We have modified this plot to include a legend.

• Figure 5: Maybe it is obvious but I don't understand why the perturbation of precip on this graph is 2 while on page 4 it is 0.5. Is it because it is inverted (2 vs 1/2)? I guess it is something like that, but it is not completely clear to me the way it is written. It would be nice if it was the same number on the figure and in the text.
Author's response: The reviewer is correct, the reason is that it is inverted. The posterior correction parameter found by the DA algorithms is close to 2 because it is a multiplicative parameter that has to compensate for the 0.5 scaling applied to degrade the forcing to recover the synthetic true forcing. But we agree it may be confusing, we will clarify this in the new version of the manuscript:

"[...] since the posterior parameter distributions approximates the actual multiplicative perturbation factor of 2 to compensate the 0.5 scaling factor used to degrade the input precipitation"

• Figure 6: the caption announces a « dashed line » but there is no dashed line.
Author's response: There was an error in the captions, we will correct it in the new version.

• Page 8: Your entire experiment if at point scale. What would be the challenges of applying this at the basin scale? It would be interesting to add some discussion about that, given that basin scale is very important for hydrology and hydrologists are also potential readers of the Cryosphere
Author's response: Most of the distributed snow DA experiments are implemented in a cell by cell basis, so we expect improvements in basin scale simulations as well. We will add the following sentence:
"It should be noted that while the experiments were conducted at a point scale, the use of remotely sensed imagery could enhance distributed simulations."

• I have noted a few typos, misplaced comas, etc in the manuscript. I include my annotated version, in which they are all encircled.
Author's response: We have corrected the typos annotated in the document (many thanks for that). Here we provide a point by point answer to the annotated comments in the document, not answered yet in the specific minor comments section.

Line147: J'aimerais bien avoir un schéma, en particulier pour illustrer tous les endroits où il y a de l'échantillonnage
Author's response: Due to the great number of replicates, it is challenging to present a schema of the cloud distribution as it changes for each of them. The clouds are randomly

distributed and differently for each of the replicates (hence the reason for different results between replicates).

Line150: Temps de calcul?

Author's response: Calculation times can vary greatly depending on the algorithm, number of particles and number of observations (In the case of the filter) and of course on the infrastructure where MuSA is launched. For reference, the PBS with 300 particles takes less than 1 minute per season and cell. Parallelisation has been implemented on a cell by cell basis, i.e. each cell is solved in one computational unit, so it is easy to make an estimate depending on the number of processors used. In the case of PF, it depends on the number of observations as the I/O cost increases considerably, but for a full season the cost can only be higher than for the PBS. An indicative estimate can be found in the MuSA original paper.

Fig1: Une mauvaise modélisation de la densité pourrait-elle mener à une surestimation de l'ÉEN?

Author's response: In this case, a biased density estimation would have an impact on the internal heat conduction, which could perhaps have an impact on the simulated LST. In the latest version we used a different density parameterisation for the synthetic truth and the degraded simulations , and the main conclusions did not change.

Fig1: Why not use the entire ensemble and compute the CRPS? Or make use of the ensemble in any way?

Author's response: The particularity of this experiment lies in the number of replicates. We have preferred to use the dispersion of the posterior means as an indicator of uncertainty, rather than the ensembles themselves for the sake of simplicity.

Fig4: Ce serait bien d'avoir aussi des barres d'incertitude pour le open loop

Author's response: This figure shows the uncertainty between replicates (replicates were used to account for the stochasticity in the DA process), not between particles. As the perturbation models are the same, the variation of the open loop between replicates is not meaningful. Moreover, what is shown is the deterministic open loop (ie, the simulation without any perturbation), since what we are trying to show is the benefit of LST-DA with respect to a classical simulation pipeline, so it remains constant between replicates.