## General Statement

*In the context of upcoming satellite missions imaging the land surface temperature (LST) with unprecedented spatial and temporal resolutions, Esteban Alonso-González and his colleagues investigate the potential of their novel, versatile data assimilation toolbox, MuSA, to assimilate such observations in order to improve energy-mass balance simulations of the snowpack. This is an exciting and promising avenue, and their effort to investigate snow surface temperature, arguably a blind spot of the snowpack modelling community, should be acknowledged, and strongly encouraged. I must add that this short paper was very pleasant to read and fits well within the scope of the Cryosphere, and want to apologize to the authors and editorial board for the delay in delivering this review.*

Author's response: We would like to thank Bertrand Cluzet for his encouraging comment, and the careful and thoughtful evaluation of our work. We also apologize for the delay in our response but we needed time to reply correctly. Indeed, the comments below led us to change the title of the paper, refine the methods, rework the figures and perform additional experiments. Overall we think this work should greatly help us submit a more convincing manuscript.

*Snow surface temperature is indeed a key variable of the snowpack surface energy balance. For instance, surface temperature observations have been long used as a boundary condition to close SNOWPACK model's energy budget under freezing conditions [1]. When the snowpack surface reaches the melting temperature, the surface energy budget can no longer be closed, but observations can be used to detect (or reject) the occurrence of surface melt, which is a valuable information per se. Finally, LST values above freezing point inform us about the absence of snow on the ground.*

*This study is an OSSE (Observing System Simulation Experiments), where the authors simulate future LST observations from the upcoming TRISHNA mission with 'degraded' model outputs from FSM. They then assimilate such observations in ensemble simulations also using FSM (therefore, this is an 'identical twin' OSSE), accounting for precipitation and air temperature uncertainties only. Their 4 study sites span a wide diversity of snowpack conditions from alpine to arctic climates. The authors investigate the influence of cloud coverage on the availability of observations by assimilating these observations under different fictive cloud coverage scenario. Their results suggest that Land Surface temperature has the potential to constrain snow water equivalent with astonishingly good performance, in particular when it comes to Landsat-like low revisit time (see e.g. Fig. 2 and l. 215-216). This seems 'too good to be true' as the authors provide little evidence to nuance their results (l. 239-243). Landsat data has been out for decades now, how could the snow data assimilation community really miss something so promising for that long, or is this result somewhat overestimated? The authors discuss this point, but the reader may remain sceptical and I see several reasons why the performance would be rather overestimated.*

Author's response: It is not possible to provide a definitive answer as to why LST assimilation has not been extensively explored in the snow science community. However, given that data assimilation (DA) in general is not widely used in this field, it is not surprising that there is ample room for new discoveries. Two years ago there was no open source

software to easily perform data assimilation with a snowpack model. In addition, an over-representation of filters vs. smoothers in snow data assimilation studies, perhaps influenced by the numerical weather prediction communities where DA is more widely used, may have hindered progress in this area. Similarly to the LST, the potential of albedo to improve snow simulations has been noted in previous studies but has not been extensively investigated (e.g. Dumont et al., 2012; Kumar et al., 2020). Our work is not the first to explore LST DA (Navari et al., 2016; Piazzi et al., 2019, 2018), but these previous studies reached very different conclusions. Piazzi et al. (2018) did not find a clear benefit from assimilating LST, whereas Navari et al. (2016) did. Our study shows that the chosen DA algorithm is probably the source of this apparent contradiction. Navari et al., (2016) used a smoother whereas Piazzi et al. (2018) used a filter. We can draw this conclusion because we used a consistent framework to evaluate both approaches.

The reviewer likely focused on Figure 2 to conclude that our results exhibit an “astonishingly good” performance. This figure represents the mean of all replicates assuming 0% cloud cover. On average, the DA works well, but may fail considering a single experiment. To highlight the uncertainty, we have modified Figure 2 to include the quantiles of the replicates. We have added the same plot for the other cloud cover scenarios as supplementary material. With the new representation it is more evident that a good performance is not guaranteed, and will be contingent on the cloud cover distribution, particularly for the Landsat-like revisit time.

Figure 3 shows the marked decline in performance as cloud cover increases, even for the 3-day revisit. The ensembles in Figure 3 represent the replicates of each experiment and not the ensemble of model realizations. Each posterior simulation in this ensemble representation includes the posterior mean of an experiment, and therefore, each is equally likely to be obtained in a real DA experiment based on the position in time of clouds and other uncertainties related to the modeling pipeline. The performance varies significantly when we move towards more realistic scenarios with 50/75% cloud cover.

*Identical twin OSSEs are very tricky to set up and should be designed with a lot of care for reliable conclusions to be drawn [2]. Unfortunately, the setup in this study does not convince me for a reason that requires to delve into some developments. Because (1) they do not integrate any snow model parameter perturbations, (2) do not confront their simulation with real (in-situ or satellite) observations of LST (and potentially SWE), and (3) do not discuss the literature on snow surface temperature modelling by snowpack models such as FSM, their OSSE implicitly lays on the fundamental hypothesis that FSM can accurately model both the snowpack’s surface temperature and the SWE. This ’perfect model’ assumption is a typical shortcoming of identical twin OSSE’s, that usually results in overly optimistic conclusions on assimilation performance (and other pitfalls) as thoroughly discussed in Sec VI.2.2 of [2].*

Author's response: We agree that twin OSSE experiments might give an exaggerated idea of how informative is a (pseudo) observation in DA context. We had chosen to degrade the model open loop with respect to the synthetic truth by introducing errors in the forcing variables using spatial disaggregation (similar approaches are mentioned in the reference [2] suggested by the referee). To address the reviewer concerns, we made the following changes:

- We have used a different parameterisation of FSM2 to generate the synthetic truth and run the DA experiment. In the latter case we used a simpler version of FSM2, whereby several processes are ignored and parameters removed. In particular the liquid water routing, highlighted by the referee as a key process, is now computed with a different algorithm which is much simpler than the one used for the synthetic truth. Other key variables such as the albedo or density are also computed using simpler algorithms. Changes in FSM2 parameterizations can lead to non-linear differences in the LST estimation, as shown in the Figure 5 of the FSM original paper (Essery, 2015).
- Snow model uncertainty mostly comes from the forcing, (e.g. Günther et al. (2019) who showed this specifically with FSM2). Therefore, we have increased the degradation of the forcing by adding Brownian noise in the range of values of each forcing variable in 12-hour windows (the DA windows of ERA5). We propose to include a new figure as supplementary illustrating the magnitude of this degradation (Figure 1 below). These new strong perturbations lead to degraded error metrics. In addition, with this new experimental setup, improving LST revisit has no beneficial effect anymore at one site (Gerlachovský štít). We analyze this as (i) a consequence of the shallow snow cover (ii) normalization of the RMSE to make them comparable between sites (the shallow snowpack of Gerlachovský štít makes small errors considered to be of great magnitude).

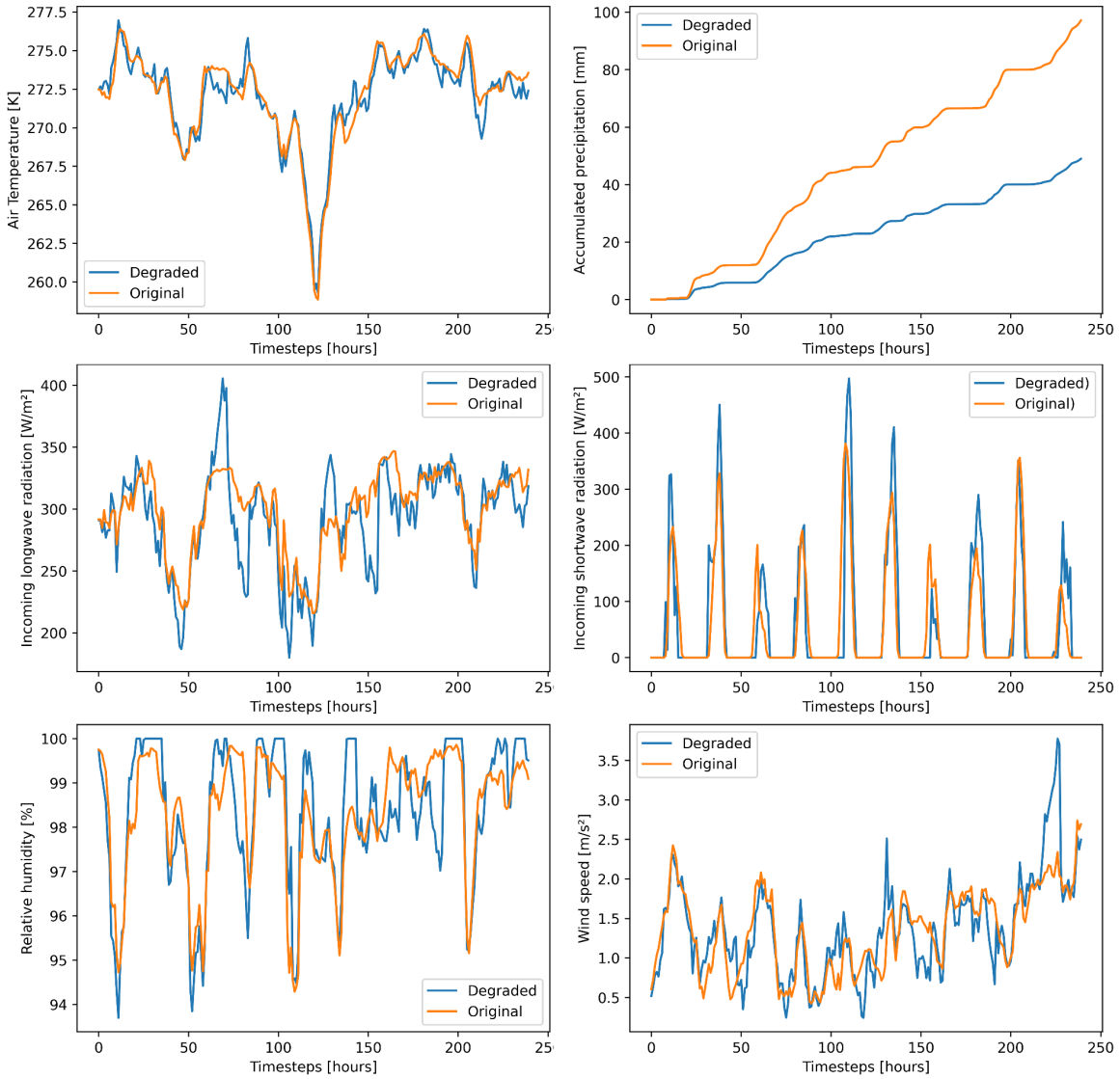


Figure 1: Comparison between synthetic true and degraded forcing

*In fact, some elements in the current literature show that models similar to FSM [3], [4] (or more sophisticated [5], [6]) can exhibit rather strong (negative) surface temperature biases with respect to in-situ measurements, despite representing bulk snow properties such as the snow water equivalent (SWE) with a satisfying accuracy. As a user of FSM, I know that very accurate SWE simulations can be obtained with more than 2-3K of negative surface temperature bias (I am pleased to provide the authors with evidence for that).*

*As an example, a physical process that could lead to an overestimated impact of surface temperature assimilation on SWE is the liquid water routing through the snowpack. Qualitatively, the snow melts near the surface, and is routed towards the bottom, implying percolation, retention, and refreezing (in particular), before it runs off (SWE decrease). FSM represents water routing with a bucket approach, which is a very simplified representation of the truth (see e.g. [7]). Because only one version of FSM is used in the current setup, there*

*is a common 'transfer function' from surface melt (surface temperature) to runoff (SWE ablation), between the synthetic truth and the ensemble members from the assimilation run. A good surface melt (surface temperature) gets artificially mapped into the good runoff (SWE decrease). But the truth is for sure blurrier, and there are plenty of plausible runoff scenarii associated with a given surface melt scenario: in real life, assimilating surface temperature might result in looser constraints on the SWE.*

*Therefore, in the current setup which does not allow the data assimilation algorithm to update FSM parameters (and therefore adjust liquid water routing, or tackle error compensations with albedo, surface density, turbulent fluxes within FSM), data assimilation of real (and accurate) observations, will probably perform much worse than the presented results. As an example, assuming a negative surface temperature bias from FSM, correcting surface temperatures (by increasing them) by way of data assimilation will very likely degrade SWE modelling performance. This is arguably a bigger potential problem, than the question of cloud influence on observation availability, and should therefore be assessed (not necessarily addressed) first.*

Author's response: We acknowledge the reviewer's concerns regarding the use of Land Surface Temperature (LST) assimilation in snow simulations. In fact, our results corroborate this suspicion, as the filter-based assimilation of the LST has a limited impact on model performance (Fig 6 of the manuscript). However, we found a clear benefit of assimilating the LST using a smoother. In other words, LST observations bring little information if taken sequentially, but LST is efficient when considering the full seasonal cycle. We interpret this as the consequence of two simple but strong pieces of information brought by the LST, as highlighted by the reviewer: (i) the LST is 0°C during the melt period (ii) the LST can only be above 0°C if the SWE is zero (see Figure 2 of the current document). Such level of information does not require an accurate representation of the snow surface temperature.

This physical fact provides information about the length of the melting period and the total length of the season, which are two crucial moments of the snow season that allow for simulating the maximum accumulation explicitly. The length of the season and melting period are the first-order information when assimilating snow data, which cannot be exploited with a filter as suggested by the reviewer. We argue that LST improves the simulations through a mechanism similar to the fractional snow cover area (FSCA), but by providing information over longer periods since the melting period starts before the FSCA shows values different from binary information, especially for deep snowpacks.

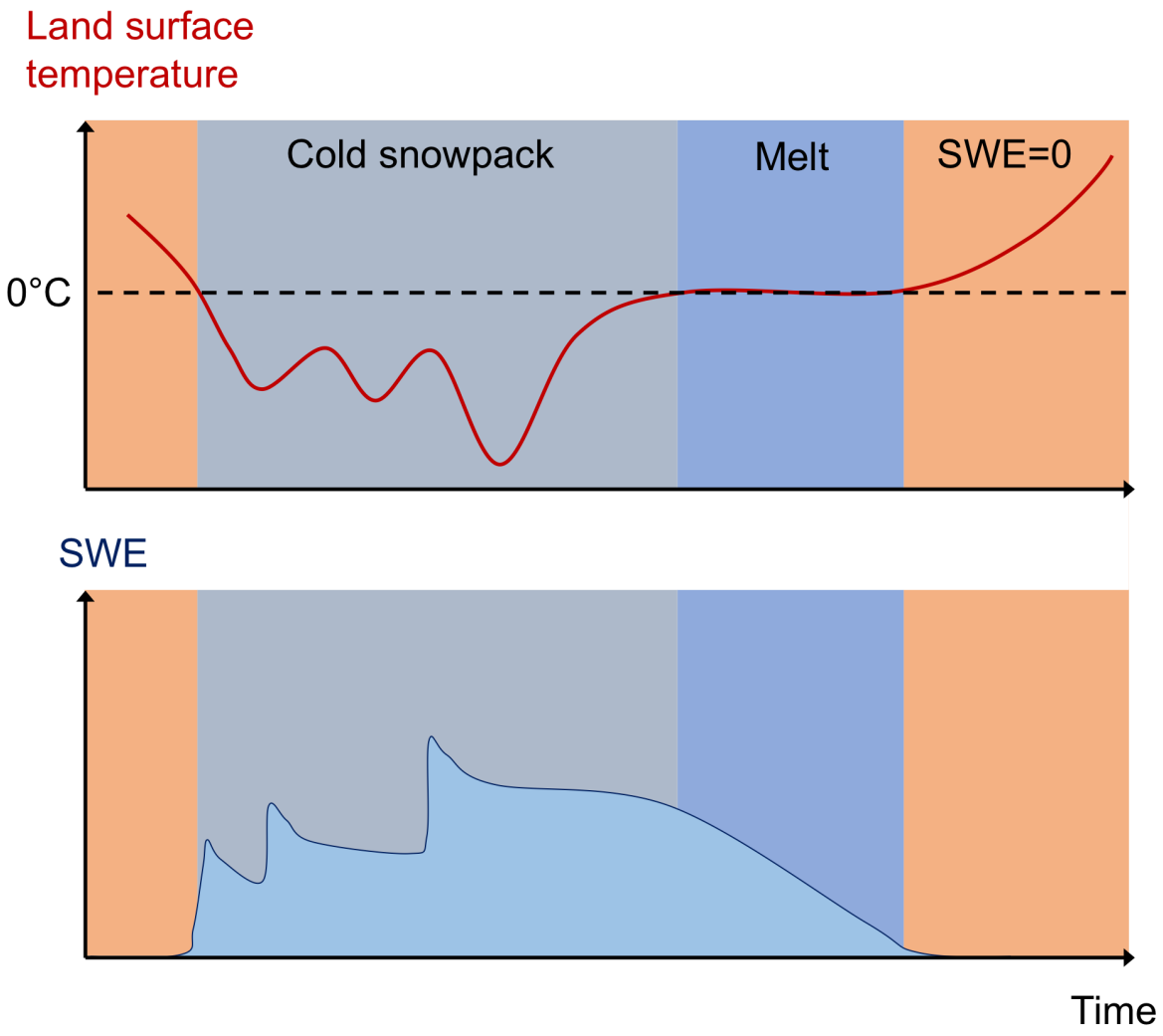


Figure 2: Schematic representation of the seasonal behavior of the LST.

*I am confident that the authors can address the above comments, but nevertheless, I think that they are substantial enough to require a major rework of the study: the conclusions of this study rely on strong hypotheses that are not substantiated enough, and significantly reduce their applicable range. Furthermore, these hypotheses could be partially avoided, but at the cost of a substantial redesign of the study. I am please to suggest pathways to address these major comments. I would further suggest that the authors add a detailed discussion on remaining shortcomings of their approach as listed below and in the annotated manuscript.*

*A premise for this OSSE experiment is therefore to (a) compare FSM outputs with in-situ or satellite surface temperature observations. I am confident that this can easily be achieved, thanks to the amazing versatility of MuSA, that would allow them to run it at the FluxAlp station [9], or any of the sites within the ESM snowMIP setup (I think that data is publicly available) [10], for example. In the absence of biases, or any potential 'mismatch', (b) in-situ or real Landsat observations [9] could even be assimilated as a proof of concept, as a very nice addition to the current study. (c) In the presence of biases, the current study should be*

*substantially redesigned, probably by integrating snow model parameter perturbations, in order to allow the snow model to 'learn' from its surface temperature errors, and trying to fit more into the identical twin OSSE guidelines provided in [2].*

*(d) An additional easy (thanks to FSM modularity) and minimal (although still quite imperfect) requirement for this OSSE would be to use a different FSM configuration for the generation of the synthetic truth (as acknowledged l.239-243) to reduce the inbreeding of the current identical twin setup.*

*Below, and throughout the annotated manuscript, the authors will find other minor (but substantial) comments that would also need to be addressed in a revised version of the manuscript.*

Author's response: Although our work is not based on actual LST observations, we believe it already provides useful insights to the snow community as explained above. In fact our results precisely call for a dedicated study on the benefit of assimilating actual remote sensing LST data, which was not an obvious prospect given the available literature on this topic.

In addition, our main intention in this article is to explore the potential of future remote sensing products, which is only feasible through a synthetic experiment. In particular, our study was designed to evaluate the added-value of an increased revisit time in different climatic contexts. To better reflect this intention and avoid misleading the reader, we propose to modify the title of our study to:

*'Exploring the potential of thermal infrared remote sensing to improve a snowpack model through an observing system simulation experiment'*

We will also modify the discussion to highlight the need for another study using in situ data to consolidate the results of this OSSE.

***Minor comments***

1. *As the authors say, LST informs on the snow surface temperature as well as on snow absence (when LST>273.15K). In the smoothing mode, because information can be propagated backwards in time, it is hard to tell apart which one of these pieces of information have the most impact, although the poor performance of the filter points towards a significant impact of the melt out date info on the smoother performance (as discussed in l. 209-214). Companion DA experiments discarding LST observations past the melt out date (or stopped in the core of the winter season) could help tell this apart more rigorously. This would be essential for real-time applications where there is no information available about the future melt-out date.*

   Author's response: We have tested a filter algorithm which is typically the strategy followed in real time operational applications. Therefore we think that the current setup already shows that such applications would not strongly benefit from LST assimilation. However, we agree with the Reviewer that this setup is not sufficient to

explain how LST assimilation is driving the SWE error reduction. However, using only the values when LST>273.15K would reduce the number of observations, which could have effects on performance, not necessarily related to the issue to be addressed here. To avoid this problem we have set up a new experiment in which we assimilate the fractional snow cover (FSCA) instead of the LST. This experiment is a way to test the information brought by the LST in comparison to just assimilating the snow cover duration, which is also measured by FSCA, keeping the same number of observations for both the LST-DA and FSCA-DA. For this purpose, we assimilated synthetical observations of the FSCA, to which we added noise in the range described by the literature (RMSE 0.17, Aalstad et al. (2020). The FSCA synthetical observations were generated using a different snow depletion curve than the one used in the MuSA parameterisation. The results suggest that the FSCA assimilation does not reduce the SWE errors as much as the LST assimilation (Figure 3 of the current document, in the case of 50% cloud cover). In addition, FSCA seems to be more sensitive to the influence of the cloud distribution given the greater dispersion of the NRMSE values. We propose to include this new analysis in the paper to better discuss the benefit of the LST, with a new figure as a support for the discussion.

2. *This paper provides no evidence of the actual behaviour of LST and SWE timeseries (there are for SWE, but only the mean, and for different cloud coverage scenarios, not for a single run). Adding example plots with openloop/synthetic truth/ assimilation runs including ensemble spread would be insightful.*

   Author's response: LST is the assimilated variable. Hence such plots would only show that the posterior LST is closer to the synthetic truth. We do not believe that including such plots would add much to the conclusions but we would be happy to include them in the future if the editor or the reviewer consider it would help support our conclusions.

3. *No information is given on the physical configuration of FSM used in this study. In particular, it would be important to know whether the albedo is parameterized as a function of surface temperature.*

   Author's response: We will include a new table (Table 1) with the information on both the configuration used to generate the synthetic true and the simplified version used for assimilation.

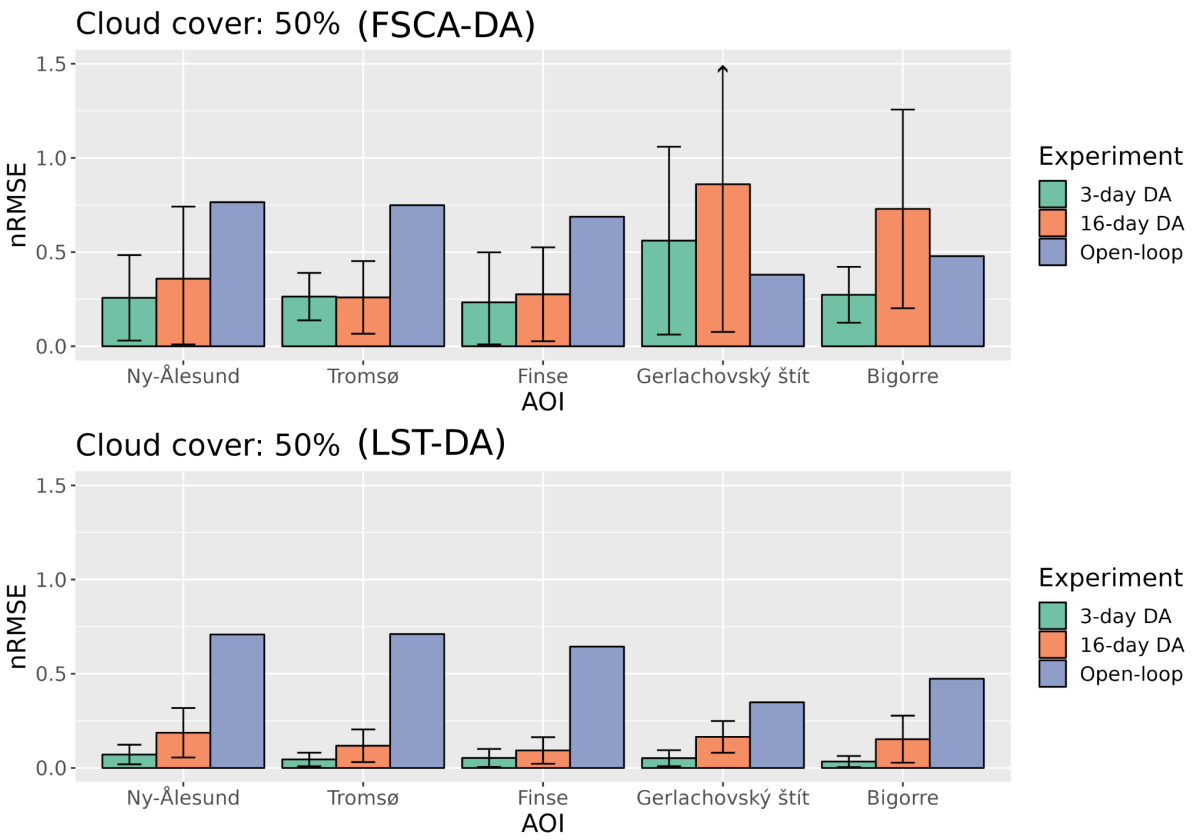


Figure 3: Comparison of the assimilation performance of LST and FSCA

Table 1: FSM2 configuration chosen (and configuration number) to generate the synthetic truth and simulations.

| Process | Syntetic true FSM2 | MuSA FSM2 |
|---|---|---|
| Snow albedo | [2] Diagnosed from snow age | [1] Diagnosed from LST |
| Snow thermal conductivity | [1] Stimated from density | [0] Constant thermal conductivity |
| Snow density | [2] Viscous compaction | [0] Constant density |
| Turbulent exchange | [1] Stability from Richardson number | [0] Neutral stability |
| Snow hydraulics | [2] Gravitational drainage | [1] Bucket storage |
| Snow cover fraction | [2] Asymptotic function | [1] Linear function |

4. *An undiscussed question is the potential (tiny) mismatch between what is modelled (the skin temperature of the snowpack, assuming a flat infinite surface) and what the satellite will see : the thermal infrared emission of a potentially rugged snow surface within a mixed pixel. I expect the difference to be tiny, but adding a discussion sentence on this would be appreciated.*

Author's response: We agree that the question of mixed pixel remains unexplored in our study. We propose to expand the discussion on that matter:

"Also, it is necessary to take into account that the surface temperature observation in complex terrain may differ from the simulated temperature, due to intra-pixel variability. But we expect these issues to be greatly mitigated, due to the expected increase in resolution,'

5. *The Snow surface temperature assimilation being rather scarce, please consider citing [11] which also assimilated SST although in a multivariate setting.*

Author's response: We thank the reviewer for this suggestion. We will add this second reference in the Introduction.

**Technical notes**

- The title does not reflect the content of the paper. 'OSSE', or 'synthetic', or 'towards the assimilation of…' or 'feasibility' should reflect the fact that no real data is being assimilated.

Author's response: We have modified the title as follows, to highlight the fact that this is a synthetic experiment.

*'Exploring the potential of thermal infrared remote sensing to improve a snowpack model through an observing system simulation experiment*

**SPECIFIC COMMENTS in the document**

*Line8: SWE can be replaced by ' snow cover' for the sake of clarity/accessibility, and this acronym should be introduced.*
Author's response: Modified as follows:
"The assimilation of data from Earth observation satellites into numerical models is considered as the path forward to estimate snow cover distribution in mountain catchments, providing accurate information of the mountainous snow water equivalent (SWE)"

*Line15: biased? lower resolution? please be more specific*
Author's response: We consider this level of detail for the abstract to be excessive. We prefer to leave it as it is for the sake of simplicity.

*Line18: with respect to cloud cover. Arguably many other aspects of DA algorithm robustness are not explored here (e.g. snow model or observation biases, snow model errors)*
Author's response: Included.

*Line21: nRMSE on which variable?*
Author's response: Corrected.

*Line24: performance?*
Author's response: Accepted.

*Line25: backward in time? (to be more accurate)*
Author's response: We prefer to keep the current formulation. Although the beginning is probably the least informative part of the season, the information also spreads into the future.

*Line32: monitoring? (or see following comment)*
Author's response: The generation of quality snow data has many applications in hydrology, ecology and economics. We prefer not to confine the text to the operational applications.

*Line32: I would suggest: the current snowpack conditions…*
Author's response: see previous comment.

*Line37: Awkward sentence logic. Consider replacing 'difficult' by 'challenging' and reformulating into 'maintain dense enough ground based...'*
Author's response: accepted.

*Line37: References are needed here.*
Author's response: We will include the following references in the new version; (Fayad et al., 2017; Condom et al., 2020).

*Line43: 'primarily constrained', or mention also snowpack modelling uncertainties themselves, which are less prominent, but not negligible (see Fig 9. of the cited paper). Also that paper is mostly talking about integrated variables such as the SWE, whereas variables such as the Snow Surface Temperature surface temperature may arguably be more sensitive to model parameterization (surface albedo, density and turbulent fluxes).*
Author's response: Accepted primarily.

*Line49: A citation to De Lannoy et al., 2012 (doi:10.1029/2011WR010588) is needed here and in the previous sentence*
Author's response: Reference included.

*Line51: Evidence on this would be appreciated. I'm not saying that this is wrong, but to my knowledge there is little evidence in the literature that getting an accurate surface temperature is a prerequisite for a good energy-mass balance of the snowpack. The thing is that the vertical temperature gradient is usually pretty strong close to the surface (and the snow be light): a model can be several kelvins off in terms of surface temperature and still get the total energy budget within a few percents.*
Author's response: LST controls longwave emission, as well as turbulent and sensible heat fluxes. Moreover, the energy balance equation is precisely solved through the surface temperature, which is unknown in FSM2. This calculation indicates whether there is surface melt or not, which occurs at 273.15 K, so there cannot be a large error at least during the melting period. Therefore, we assume that it is an important variable in modeling the snowpack.

*Line54: detecting*
Author's response: Accepted.

*Line54: Surface melting event, as opposed to 'total melt' or ablation which is arguably more essential hydrologically. When working with real observations, liquid water routing and refreezing will become critical to replicate SWE observations.*
Author's response: Included surface melting.

*Line59: True, however in real-case situations, assimilation of LST in snow-free or patchy snow conditions may become arduous. Snowpack models such as FSM are not designed to accurately model the LST (bad vertical discretization in particular), and therefore may exhibit errors or biases that might lead the DA to degrade the performance. Such a phenomenon can not be evidenced in this OSSE approach, since the synthetic truth is the model itself.*
Author's response: The uncertainty induced by patchy conditions is probably more limited at the high resolution we are considering here (Landsat, Trishna) than with MODIS-like resolutions, as the time period of patchy conditions at high resolution will be shorter. In any case, the parameterisation would be similar to the current one used for the calculation of FSCA, which, despite its uncertainties, allow improving simulations through data assimilation.

*Line61: No! This was also an OSSE, not an actual proof with real data. 'Could potentially improve' is a much more appropriate statement.*
Author's response: Accepted.

*Line62: synthetic IST*
Author's response: Accepted.

*Line63: typo*
Author's response: Corrected.

*Line63: references?*
Author's response: Included.

*Line67: This is somewhat contradictory with the good performance that you obtain with the 16-day revisit time (Fig. 2).*
Author's response: The fact that it is an average may cause the figure to be misinterpreted. We have included the standard deviation to clarify this point.

*Line72: worldwide? At which latitude?*
Author's response: We have specified as follows:

“TRISHNA is expected to provide surface temperature measurements at 60 m spatial resolutions every 3 days at the equator, with an increasing revisiting time towards the poles”

*Line79: temporal frequency of observations*
Author's response: Included resolution.

*Line85: No, this paper was assimilating real data, not synthetic ones.*
Author's response: It is true that this experiment is not an OSSE per se. But we included it in the references since the precipitation was synthetically degraded. We have removed the reference.

*Line89: people might not be familiar with this jargon, which is worth introducing*
Author's response: Included explanation.

*Line93: swap these terms. Indications about the countries might help*
Author's response: Swapped, and included countries.

*Table1: near sea level?*
*Table1: high elevation?*
Author's response: Table1: We have corrected this terms in the text.

*Line99: resolution?*
Author's response: Included.

*Line101: used*
Author's response: Accepted.

*Line102: This formulation is way too simplistic. Instrumental noise is only a fraction of the whole 'error budget' of the observation. Retrieval errors, assumptions on snow emissivity, snow surface properties, and amospheric properties, cloud classification, all of these contribute to observation error and should be acknowledged*
Author's response: We have changed instrumental by observational.

*Line103: I would recommend to include more scenarios of observation error (bias and noise values) in this study, as the expected LST error is quite loosely constrained and may vary in time/space: this would allow the authors to conclude on 'acceptable' retrieval error levels for the data assimilation to work well, which would be very insightful.*
Author's response: Although the error model is the same, the errors change for each observation, so they vary both in space (each experimental area is repeated 100 times) and in time. Adding different error models would add many degrees of freedom to the experiment, making it difficult to interpret, and increasing its computational cost tremendously. Finally, as snow scientists we have to work with the expected errors as the specifications of the satellites will not be modified for our specific needs.

*Line105: I could not find any estimation of the LST products expected performance for Trishna in this reference.*
Author's response: We have removed this reference.

*Line108: ... on data availability. Cloud cover will also induce additional errors due to difficult snow/cloud discrimination. Especially, if I understand Lagouarde et al., since clouds will be discrininated using bands at coarser resolution than the TIR bands.*

Author's response: Included.

*Line111: It is hard to assess how much this step actually 'degrades' the forcing. Plotting some statistics, or timeseries of key meteorologic variables would help grasping how much this is the case.*
Author's response: We have further degraded the forcing and included an example plot for a 10 days period.

*Line116: remove this word in that case*
*Line116: OSSE is when model outputs are used as a fake truth, in this paper they were* assimilating real (Pl'eiades data)
Author's response: We have removed synthetic, but the forcing was artificially degraded in this experiment which is what we wanted to highlight.

*Line121: earlier on you also mentioned using the PF*
Author's response: Correct, added.

*Line125: Why only temperature and precipitation? Snowpack modelling errors are also due to SW/LW errors, not to mention Wind speed and snow model parameters themselves*
Author's response: There are two reasons for this. The first is that PF and especially PBS are very prone to degeneracy. Keeping a limited number of dimensions helps with this problem. But also, not correcting for errors in the rest of the forcing introduces errors in the simulations that allow at least partial consideration of the structural error of the model. In any case, correcting only the variables that are considered most uncertain is a standard procedure. We will add this explanation in the new version.

*Line153: yes, but it would be important to tell whether the pixel averaging or the precip scaling has the biggest influence*
Author's response: It is not straightforward to quantify the importance of each. It will depend on the topography and climatic conditions. But we do not see the importance of analyze this in detail.

*Line 154: consider reminding that this is a synthetic true (for the sake of clarity at the beginning of the results)*
Author's response: included.

*Line 174: 0.5*
Author's response: It is 2. Explained as follows to avoid confusions:

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

*Line 199: time*
Author's response: Added.

*Line 199: not necessarily. A systematic bias in modelled IST, would cause saw-tooth patterns too. Do you evidence this yourself? Is this a statement from Piazzi? More details would be interesting here.*

Author's response: Even if it's because of a systematic bias, these patterns will probably not appear with a smoother-like algorithm. But it is true that this statement is probably too speculative, as without access to their experiment it is not possible to compare. We have removed it in this updated version of the manuscript.

*Line 208: wintertime*

Author's response: Corrected.

*Line 210: smoother*

Author's response: Corrected.

*Line211: Which figure?*

Author's response: Included 'Figure 6'.

*Line 242: Why does it simplify the interpretation of the results of your work? From an external perspective, this inbreeding could make the results more dubious.*

Author's response: Because it moves the problem to a proper simulation (and observation) of the surface temperature. But the results suggest that from the assimilation side, it can work using smoothers. We have removed it anyway as now very different parameterization is used.

*Line 249: and resolution?*

Author's response: added.

*Line 272: please provide a tag for the musa version*

Author's response: There is no specific tag for the version used here, so we just included the version number. The modifications mentioned here are in the FSM code, but is also included in the MuSA version.

*Fig2: Midi de Bigorre should read Bigorre.*

Author's response: Corrected.

*Fig3: which location?*

Author's response: Added.

*Fig3: average*

Author's response: Corrected.

*Fig6: wrong caption*

Author's response: Corrected.

# References


Aalstad, K., Westermann, S., and Bertino, L.: Evaluating satellite retrieved fractional snow-covered area at a high-Arctic site using terrestrial photography, Remote Sens. Environ., 239, 111618, https://doi.org/10.1016/j.rse.2019.111618, 2020.

Condom, T., Martínez, R., Pabón, J. D., Costa, F., Pineda, L., Nieto, J. J., López, F., and Villacis, M.: Climatological and Hydrological Observations for the South American Andes: In situ Stations, Satellite, and Reanalysis Data Sets, Front. Earth Sci., 8, 2020.

Dumont, M., Durand, Y., Arnaud, Y., and Six, D.: Variational assimilation of albedo in a snowpack model and reconstruction of the spatial mass-balance distribution of an alpine glacier, J. Glaciol., 58, 151–164, https://doi.org/10.3189/2012JoG11J163, 2012.

Essery, R.: A factorial snowpack model (FSM 1.0), Geosci. Model Dev., 8, 3867–3876, https://doi.org/10.5194/gmd-8-3867-2015, 2015.

Fayad, A., Gascoin, S., Faour, G., López-Moreno, J. I., Drapeau, L., Page, M. L., and Escadafal, R.: Snow hydrology in Mediterranean mountain regions: A review, J. Hydrol., 551, 374–396, https://doi.org/10.1016/j.jhydrol.2017.05.063, 2017.

Günther, D., Marke, T., Essery, R., and Strasser, U.: Uncertainties in Snowpack Simulations—Assessing the Impact of Model Structure, Parameter Choice, and Forcing Data Error on Point-Scale Energy Balance Snow Model Performance, Water Resour. Res., 55, 2779–2800, https://doi.org/10.1029/2018WR023403, 2019.

Kumar, S., Mocko, D., Vuyovich, C., and Peters-Lidard, C.: Impact of surface albedo assimilation on snow estimation, Remote Sens., 12, https://doi.org/10.3390/rs12040645, 2020.

Navari, M., Margulis, S. A., Bateni, S. M., Tedesco, M., Alexander, P., and Fettweis, X.: Feasibility of improving a priori regional climate model estimates of Greenland ice sheet surface mass loss through assimilation of measured ice surface temperatures, The Cryosphere, 10, 103–120, https://doi.org/10.5194/tc-10-103-2016, 2016.

Piazzi, G., Thirel, G., Campo, L., and Gabellani, S.: A particle filter scheme for multivariate data assimilation into a point-scale snowpack model in an Alpine environment, The Cryosphere, 12, 2287–2306, https://doi.org/10.5194/tc-12-2287-2018, 2018.

Piazzi, G., Campo, L., Gabellani, S., Castelli, F., Cremonese, E., Cella, U. M. di, Stevenin, H., and Ratto, S. M.: An Enkf-Based Scheme for Snow Multivariable Data Assimilation at an Alpine Site, J. Hydrol. Hydromech., 67, 4–19, https://doi.org/10.2478/johh-2018-0013, 2019.

# 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.