RC3: 'Comment on tc-2021-225', Anonymous Referee #3, 30 Sep 2021

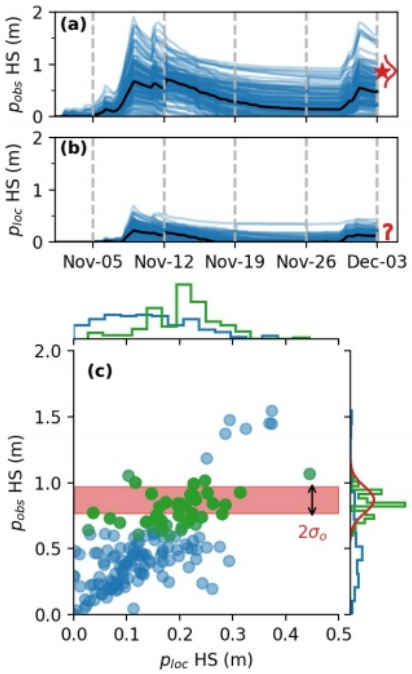
The authors would like to thank the reviewer for his acute review asking for more clarity in the description of the procedure, globally in the same line as RC2. Sec. 2 was considerable adjusted to fit with these requirements, and we believe that new Fig. 3 will really help understand our assimilation methods. As special care was taken to the quality of the figures. Note that some slight changes were also made in the manuscript in order to improve its clarity, and are visible in the track changes. In the following, the reviewer initial comments are written in black, our answer in blue and the corrections in the paper are highlighted in red.

The paper demonstrates that assimilation of in-situ sown depth observations does not provide significant RMSE improvements over the open-loop and operational simulations. It reduces bias in snow depth estimates and outperforms the open-loop simulations in specific elevation bands in locations with a lower observation density.

The methodology and the presentation of results are clear, and I agree that this approach seems to be relevant for the estimations of snow depth and SWE. For all these reasons, I think the paper should be published with a minor revision.

The paper is interesting and provides insight into analyzing ensemble data assimilation approaches. However, there is too much emphasis on the Continuous Ranked Probability Score (CRPS). It would be more insightful to show snow depth maps or scatter plots of model simulations (open loop, oper, DA vs. observations).

Thanks for this thorough remark on our methodology and presentation of results. Indeed, focusing on the CRPS is a deliberate choice, as it is much more suited to the evaluation of ensembles than RMSE or bias, since it accounts for the actual distribution of the ensembles, and offers a way to aggregate scores both spatially and temporally, giving robustness to the evaluations. RMSE and bias are given as a means to compare the ensemble median to the operational (deterministic) run globally (Tab. 1 and Fig. 6). Snow depth maps cannot be plotted, as simulations are only ran at the stations, and not in a grided setting. We also argue that scatter plots would lack the ability to synthesize information, while requiring the need to reduce the ensembles to their median. However, we acknowledge that illustrations on the behaviour of the ensemble (in terms of spread, and temporal variations) would be appreciated, and therefore we think that Fig. 3 of the revised manuscript provides some insight on the ensemble behavior.



L137-140: Please provide perturbation statistics (additive/multiplicative, mean, standard deviations/correlation coefficients, spatial, and temporal correlations) for each forcing).

We agree that the lack of details was detrimental to the clarity of the study. However, we believe that explaining the full perturbation procedure, already described in the provided reference, would help focusing on the meaningful details. We rather added some outstanding details. In particular we point out that perturbations are spatially homogeneous (see below)

An ensemble of forcings was generated by applying stochastic perturbations in the same spirit as \citet{charrois2016} but with slight corrections in the implementation of the perturbations compared with \citet{cluzet2020towards,cluzet2021croco} as described in \citet{deschamps2021assimilation}. For each member, perturbations are auto-correlated in time following an auto-regressive process and are spatially homogeneous. The perturbation parameters were taken from \citet{charrois2016}. Precipitation parameters were adjusted (i.e. multiplicative noise with auto correlation time $\tau= 1500 h$, and dispersion $\sigma=0.5$) in order to obtain a spread-skill close to 1 for the open-loop run (see Sec. \ref{sec:res_ref}). We used these perturbed analyses as input for the snowpack simulations at the stations.

L165: a summary about the updating step and how the PF updates the snow profile would improve the clarity of the setup section.

The authors would like to thank the reviewer for pointing our the lack of clarity and details of the corresponding Sec. 2.2.3. This section was thoroughly rewritten. Furthermore, a new Sec. 2.2.4 was added to illustrate the behavior of the localised POF in a simple example. We belive that this will really improve the understandably of the manuscript.

The Particle Filter used in this work is based on the version described in \citet{cluzet2021croco}. Only a brief description of the procedure is given here. The ensemble is updated sequentially with the PF on each assimilation date and propagated forward until the following assimilation date. The PF is localised: each point receives a different analysis. Based on the comparison of neighbouring simulations of HS with their corresponding HS observations, the PF selects a sample of the best ensemble members. The idea is that if a particle is performing well against nearby observations, it should also be efficient locally \citep{farchi2018comparison}. Different localisation radius are tested in this study ranging from 17 \unit{km} to 300 \unit{km}. Note that when a particle is selected by the PF, the full local state vector is copied: the local physical consistency of the variables is preserved.\\

Particle Filter degeneracy (see Sec. \ref{sec:intro}) may arise even with a reduced local domain size, and approaches to increase the PF tolerance may be required to overcome it. The localisation is complemented here by two different strategies described in \ citet{cluzet2021croco}, inflation and k-localisation, leading to the 'rlocal' and 'klocal' algorithms, respectively. If the initial analysis is degenerated (i.e. the effective sample size N_{eff} is inferior to a target N_{eff}^*), the rlocal and klocal iteratively modify the assimilation settings to make it more tolerant, so that the PF analysis reaches a sample size of N_{eff}^*. The rlocal algorithm performs an inflation of observation errors inspired by \citet{larue2018assimilation}. The klocal algorithm discards observations coming from locations exhibiting the lower ensemble correlations with the considered location. It is important to note that inside a localisation radius, the rlocal method assimilates all available observation stations whereas the klocal method only selects a subset of observations from locations where the ensemble members are sufficiently correlated with the simulation members of the considered point.\\

L229: Please provide some information about the resolution of simulations.

How are the bias and RMSE computed (i.e., point vs. gridded simulations)? What is the uncertainty of this comparison?

The authors can elaborate more on the representative error. Gridded simulations are compared with ground-based point measurements. There is a scale gap in this comparison. It will benefit the paper if the authors discuss this source of uncertainty.

The authors acknowledge that more details were necessary in the description of the simulation setup. The simulations are not gridded, but performed at point locations colocated with the measurements: there is no scale mismatch between simulations and observations, and therefore no uncertainty in the comparison. We agree that there remains representativeness errors in the observations, because of snowpack varibility at the plot scale scale.

We also clarified the model configuration by stating more clearly that SAFRAN analyses are downscaled at the stations in Sec. 2.2.2:

Meteorological forcings are taken from SAFRAN reanalysis over the Alps and Pyrenees. SAFRAN \citep{durand1993} is a surface meteorological analysis system adjusting backgrounds from NWP model ARPEGE \citep{courtier1991} with local meteorological observations (air temperature, pressure, precipitation, humidity) within so-called massifs of about 1000 \unit{km^2} (see Fig. \ ref{fig:meth_obs_density_massif}) and further downscaled to the stations of our study. [...] SAFRAN analysis is issued separately for each massif in a semi-distributed geometry, i.e within 300 \unit{m} elevation bands, aspect and slopes, the main topographic parameters controlling the snow cover evolution. This analysis is subsequently downscaled into the specific topographic conditions (i.e. elevation, slope, aspect and local topographic mask) of the simulated station \citep{vionnet2016}.

We also adapted the discussion in Sec 5.4 was rewritten to reflect the consideration on potential scale mismatch and observation representativeness:

Nevertheless, obtaining a perfect spread-skill may be a challenging goal for our assimilation system. Indeed, under dispersion is a common issue in the NWP \citep[e.g.][]{bellier2017sample} and snow cover modelling communities \ citep{lafaysse2017multiphysical,nousu2019statistical}, and can be explained here by several factors. On the one hand, despite meteorological forcings are downscaled to the stations (see Sec. \ref{sec:forc}, so that there is no scale mismatch), the ensemble modelling chain does not account for two important processes affecting the observations. The variability of the meteorological conditions inside SAFRAN massifs is limited to topographic parameters (including local masks) so that two distant stations with the same topography will receive the exact same forcing, and the snow redistribution by wind is not represented \citep{vionnet2018,mott2018seasonal}. On the other hand, the representativeness of observations is limited by plot-scale variability.
\\
Data assimilation is known to partly compensate for such mismatches via error compensation \ citep{klinker1992diagnosis,rodwell2007using,wong2020model}. For example, an ablation event in one observation can be compensated in the Particle Filter by selecting some members with a lower precipitation factor or a compaction scheme with a higher settling \ citep{deschamps2021assimilation}. This compensation immediately results in lower errors, but implicitly, the model does a wrong assumption, which results in being over confident, thus with a lower spread. The only way to mitigate for this over confidence is to account for any relevant physical phenomenon, which is a desirable goal, but a real challenge when it comes to snowdrift by wind, local meteorology and plot-scale variability. This goal is to date out of reach at the temporal and spatial scale of this study.\\

L300: What is the reason for presenting figure 10b?

This figure is presented and commented in l. 298-301 of the submitted manuscript. The intention is to show that there is a relation between the openloop bias and the density of observations.

If possible, please improve the quality of the figures.

Thanks for this remark. More than half of the figures were carefully rehandled (Figs. 1, 2, 4, 5, 7, 8 and 11).