RC1: 'Comment on tc-2021-225', Anonymous Referee #1, 21 Sep 2021

The authors would like to thank the reviewer for this throughout review, and in particular for the question of the co-located meteorological and snow observations. Several adjustments were made to the paper in line with this, and a new figure was added. 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.

Summary of the paper

Cluzet et al. present an evaluation of the CrocO snow data assimilation system which is based on a Particle Filter technique that propagates information spatially from areas with observations (e.g., height of snow, HS) to unobserved areas. Using leave-one-out validation, the performance of this system is compared to that of the ensemble (i.e., open-loop) and operational model counterpart at hundreds of snow depth stations in the French Alps, Pyrenees, and Andorra over 10 years. Different localization schemes (rlocal vs. klocal) and radii are tested, and showed only minor changes in skill. All Particle Filter configurations showed improvements relative to the open-loop but no clear improvement over the operational model. Using the klocal scheme at 35 km for the rest of the study, the authors show it tends to mitigate the negative bias in the open-loop at certain elevations (1500-2500 m) and locations (e.g., Central-East Pyrenees). The klocal scheme shows the highest skill at locations where the open-loop has negative bias, with reduced skill when the open-loop has positive bias. The skill of the klocal PF was found to be insensitive to the density of HS observations, but the magnitude of open-loop bias decreased with increasing HS observations (a peculiar result). The study suggests there is complementarity in that areas with high errors in meteorological data (due to low station data) may be mitigated by a Particle filter with spatial propagation of snow information.

Recommendation

Overall, I find that this paper is of good quality and provides additional application and evaluation of the CrocO Particle Filter developed in recent work by the authors. It is conducted at a more coarse model scale (and less dense observation network) than previous studies (e.g., Magnusson et al 2014; Winstral et al., 2019) and is hence a unique, complementary contribution. I think there are some aspects that could be strengthened in the discussion and possibly the presentation of results, depending on what information is available about the HS stations. The issue is that the stations may benefit both the open-loop and oper (through providing meteorological data to the reanalysis) as well as the PF (through providing HS observations for assimilation). Disentangling these competing contributions to the model configurations could be useful in a revised manuscript.

Comments

The authors would like to thank the reviewer for this very acute remark. We completely agree that these observations may have an impact on the performance of the reference runs (openloop and oper), thereby reducing the potential for improvements for the PF assimilating snow observations.

Over the considered period of time, 438 observation sites provided precipitation observations to SAFRAN over this area in the snow season (November-April). These stations are mostly located at lower elevations (below 1500 m) as presented in Fig.4 of Vernay et al. (in review). Among them, 164 stations correspond to locations with snow depth measurements included in the present study. XXX other stations provide snow depth measurements but without precipitation data. We reproduced below Fig. 10 of the submitted manuscript, highlighting in green the 79 stations having an almost-daily precipitation observation frequency (150 obs on average between November and April), (see Fig. 1 here). This figure shows that the performance of the assimilation is similar in the locations with a lot of precipitation observations than in the others.

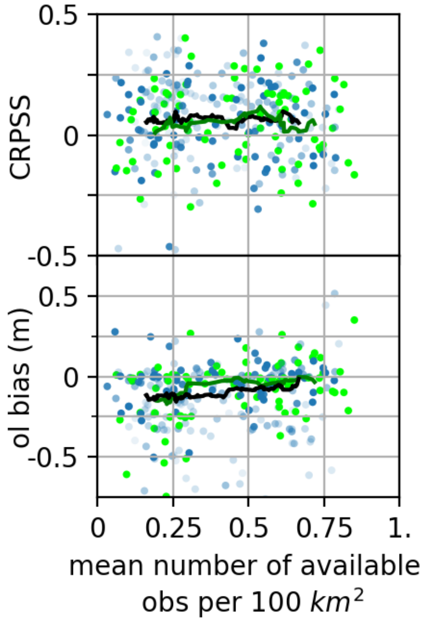


Illustration 1: R1.1: like Fig. 10, with green dots corresponding to the 79 stations with more than 1500 precipitation observations. The grey line is the average of these stations, and the black line is the average of the rest of the stations.

The reason is that SAFRAN does not assimilate precipitation observations locally, but performs a single meteorological analysis within each massif (see l. 128-136 of the submitted manuscript). Precipitation observations may therefore impact the reference runs at neighboring locations, and coincidentally the performance of the reference runs might increase with the density of the available precipitation observations.

This is the hypothesis that we formulated on l. 458-459: the increase in performance of the reference runs with the snow observations density should be explained by the fact that the snow observation density is correlated to the meteorological observation density. To confirm this, we performed a more thorough analysis of the relationship between the spatial density of snow observations and the sparial density of precipitation observations assimilated by SAFRAN. The outcome is presented in Fig. 12 of the revised manuscript, and confirms our statement l. 458-459: the density of snow observations increases with the density of precipitation observations used by SAFRAN, both in the Alps and Pyrenees.

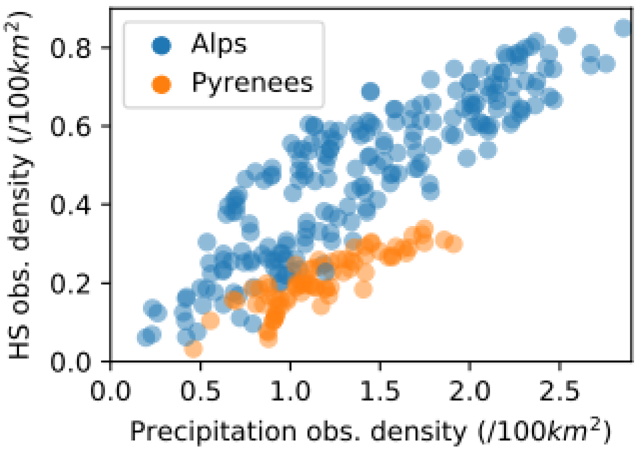
Several changes were performed to reflect this and answer the reviewer's remark in a concise way:

a. more details were included on the precipitation observations, the following sentences being added in Sec. 2.2.2 (see track-changes):

Over the considered period of time, 438 observation sites provided precipitation observations to SAFRAN between November and April. These stations are mostly located at lower elevations (below 1500 m) as presented in Fig.4 of \citet{vernay2021meteorological}. Among them, 164 of these sites correspond to locations with snow depth observations included in the present study.

b. Fig. 12 was added and Sec. 5.5 amended to confirm our hypothesis on the correlation between the HS and precipitation observation densities.

The density of in-situ observations has been pointed out as a critical parameter for the success of data assimilation \citep{largeron2020towards}. \citet{winstral2019bias} managed to strongly reduce modelling errors with a high observation density, (about 1 observation site every 100 \unit{km^2}). Because of natural variability, they considered detection of systematic errors may be more difficult with a lower density. Our study case explores a wide range of observation density (Fig. \ref{fig:res_CRPSS_bias_density_scatter}), from about 0.1 to 0.8 observations every 100 \unit{km^2} (accounting for the availability of observations). Yet, as mentioned in Secs. \ref{sec:res_assim} and \ref{sec:disc_ref_perf}, the assimilation performance relative to the open-loop does not decrease with a lower observation density. It may be due to the fact that the assimilation is efficient only for strong open-loop negative biases (Fig. \ref{fig:res_CRPSS_scatter}b), which seems the highest where the station density is the lowest (Fig. \ref{fig:res_CRPSS_bias_density_scatter}b). In other words: the assimilation can not outperform the open-loop in the most densely observed areas (e.g. in the Northern Alps, where the observation density is similar to the studies of \citet{magnusson2014assimilation} and \citet{winstral2019bias}) because the open-loop performance is already high there. This behaviour is explained by the fact that the HS observation density is correlated with the density of precipitation observations used by SAFRAN to analyse the meteorological forcings (see Fig. \ref{fig:corr_HS_precip} and Sec. \ref{sec:forc})). Both (at the exception of the Nivôse and EDF nivo stations for the HS observations) are actually related to human implantation in the valleys and the presence of ski resorts. A higher weather station density for SAFRAN is likely to result in more accurate meteorological forcings, thus reducing the bias of the reference runs, which finally leaves less room for improvement by the assimilation.\\



1.A result highlighted in the paper is that the open loop has lower performance in areas with fewer stations and decreased bias as station density increases (L. 299-301 and Fig. 10), but this seems coincidental, as the open loop does not integrate HS observations. Is this more suggestive of reduced quality in the reanalysis forcing data, given the lower availability of surface stations? Alternatively, I wonder how many of the HS stations include meteorological observations that are integrated into the reanalysis data (see lines 92-93)? Overall, the authors should carefully consider the two sources of data offered by stations (meteorology and HS). It might seem that a lack of stations would influence both the open loop and DA for that reason. The authors seem to understand this later in the paper (L. 458-459), so a consistent viewpoint should be included through the paper. Is it possible to identify which HS stations have met data that are being integrated into the reanalysis forcing data? If so, it may be necessary to differentiate the results based on whether the HS stations have met data that are contributing to the forcing data or not.

This point is answered above.

2.The leave-one-out validation (L. 63-65, 166-169) should be clarified whether this is done spatially (e.g., one station is removed) or temporally (e.g., individual HS data are removed at a station). I am still unsure after re-reading these parts.

The purpose of the leave-one out approach is to assess the propagation of information in a localised setting. In this approach, any available local observation is excluded from the assimilation, so that we can evaluate what is the impact of nearby observations on this location.l. 63-65 were amended to clarify this point:

To assess the potential transfer of information, we opt for a leave-one-out approach \citep[e.g.][]{slater2006snow}, whereby the assimilation is performed considering neighbouring observations, but discarding any local observation. The assimilation performance can be then evaluated independently from these local observations. If such potential transfer could be demonstrated, ...

3.The study highlights a range of elevations (1500-2500 m) where is a strong negative bias in the open loop and oper simulations. The role of gauge undercatch is discussed as the most likely cause of this bias (L. 351-353). However, I do not find this convincing, as gauge undercatch is likely a universal problem for measuring precipitation and snowfall in mountain environments. Is there another factor that might explain the bias in this specific elevation zone? Are the wind speeds higher, and hence larger undercatch errors at these elevations? Or is this suggestive of an oddity in the HS station data, for example, more ski area observations in these zones, which may be biased toward higher snow depths?

Here, higher winds at higher elevations was implicitly the factor behind higher under-catch at higher elevations. To our knowledge, the driving factor for the localisation of ski resorts is the general exposition (preferentially north), elevation,and economic potential, and not potential local higher snow amounts (there are also resorts in snow scarce areas). And observation sites are flat and away from any human influence such as grooming of snow-making. The corresponding sentence l 351-353 was amended to reflect that higher winds may be the driving factor.

...and depending on the considered region. The strong negative biases at those altitudes may be explained by higher wind speeds, causing an underestimation of solid precipitation amounts in gauges \citep{kochendorfer2017quantification}, and consequently in SAFRAN, as evidenced by \citep{queno2016snowpack} during strong precipitation events.

4.In the discussion, the authors discuss variability of snow redistribution by wind in the context of error compensation with the DA system (L. 423-432). One aspect that is missing here is the impact of scale – both the process scale and the model scale – and this is something that should be identified and discussed. The resolution of the model in this study is quite large, such that a process like wind redistribution cannot be represented explicitly, and ultimately the variability must be handled in other manners (e.g., as a sub-grid parameterization, see Clark et al., 2011 for example; or improving the model resolution to a finer scale). Please comment more on this issue.

This point is quite technical. Indeed, as stated in l 132-134, the meteorological analyses are determined by topographic classes at the scale of SAFRAN massifs, but are then downscaled to the topographic conditions and terrain shading of the stations where the simulations are performed. They are conceptually very close to point simulations forced with local meteorological observations, so that observations and simulations are the representation are directly identifiable. However, and as pointed out by the reviewer, our aim in l. 423-432 was to stress out that unrepresented processes in the model (mainly wind drift and intra-massif variability), and the limited representativeness of snow observations at the plot scale may explain discrepancies between modelled and observed variables, and induce error compensation by the assimilation algorithm. The corresponding paragraph from Sec 5.4 was rewritten to reflect this:

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.\\

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}.

General comments

- The argument that weather stations provide better estimates of surface meteorology than NWP may not be a widely supported viewpoint, for instance see Lundquist et al. 2019. Please comment.

This comment refers to an introduction paragraph l. 30-37. We agree that the sentence 'Information from weather stations located in the mountains provide better estimates of surface meteorology than Numerical Weather Prediction (NWP) models' is wrong, particularly at the light of Lundquist et al., (2019). In particular, NWP models are much more able to capture the spatial variability of meteorological conditions than in-situ observation networks. We want to reformulate this sentence because our point is rather to mention the complementarity between these sources of information:

In that context, additional sources of information are needed to mitigate snowpack modelling uncertainty in the mountains. Observations from weather stations located in the mountains can be used to correct Numerical Weather Prediction (NWP) model outputs. Dedicated downscaling and analysis schemes such as SAFRAN \citep{durand1993} or RhiresD interpolation in Switzerland \citep{frei1998precipitation} can be used to efficiently reduce the large errors of the NWP models in the mountains, in particular by the assimilation of local precipitation observations.

- There are many instances where papers that are "in prep" or "in review" are cited, for instance Vernay et al., Deschamps-Berger et al., etc. Please check with the TC journal guidelines on whether these types of papers (not yet published) are permitted to be cited.

We checked on this webpage: https://www.the-cryosphere.net/submission.html#references. The "in prep"and "in review" papers are now "in review" or published.

Technical corrections

- Line 13: Add "of" after "strategy". OK

- Lines 53: Add "the" before "Météo-France". OK

- Line 71: Replace "Does" with "Can". Also, delete "manage to". OK

- Line 84: Add "Alps" after "Northern". OK

- Line 104: Replace "less" with "fewer". OK

- Line 137-142: Can you include a brief summary of the key elements of the ensemble generation technique? This would be useful for someone who has not read Cluzet et al. (2020, 2021) or who needs a reminder about the methods.

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.

Thanks for pointing this out. We introduced some more elements. The following sentence was added:

For each member, perturbations are auto-correlated in time following an auto-regressive process and are spatially homogeneous.

- Line 151: Replace "radius" with "radii". OK

- Line 152: Should be "size" instead of "sizes". OK

- Line 156: Suggest replacing "remind" with "note". OK

- Line 157: Should be "rlocal". OK

- Line 166: Should be "requires use of independent data". OK

- Line 202: Add "of" at the end of this line. OK

- Line 203: Remove "If". OK

- Line 218: Replace "Despite" with "Although". OK

- Line 228: A final sentence is needed here to offer an explanation for interpreting the new formulation of the skill score.

The following sentence was added:

These properties are important to visually compare and average improvements (positive CRPSS) and degradations (negative CRPSS) of the CRPS.

- Line 236-237: This argument is not reliable, as the 2012 year looked to be snowier but the bias was lower (0.11 vs. 0.16). Please remove example or provide better justification/explanation.

This proposition was removed as it is not essential and would require unnecessary additional explanations. Thanks for pointing this out.

- Lines 239-240 and Table 1: Suggest adding a column with SS so the reader does not have to do the simple math.

The corresponding column was added to Table 1.

- Line 242: Add "negative" before "biases". OK

- Line 250: Replace "radius" with "radii". OK

- Line 292: Remove "Then". OK

- Line 295: Please revise the phrasing "pointed as", this is unclear to me. OK

'pointed as' was replaced by 'identified as'.

- Line 309: Please clarify which simulation is being referred to here when discussing the "simulation bias".

We referred to all the simulations. Rephrased into:

The most interesting feature here, is that the biases of all the simulations are increasing

- Line 316: Replace “inferior to” with “less than”. OK

- Line 335: The phrasing is awkward here and I suggest rewording “enables to satisfactorily account for”.

Reworded into:

The open-loop run is reliably accounting for its modelling uncertainties and errors, since its SS is slightly below unity over the ten years.

- Line 402: Replace “Despite” with “Although”. OK

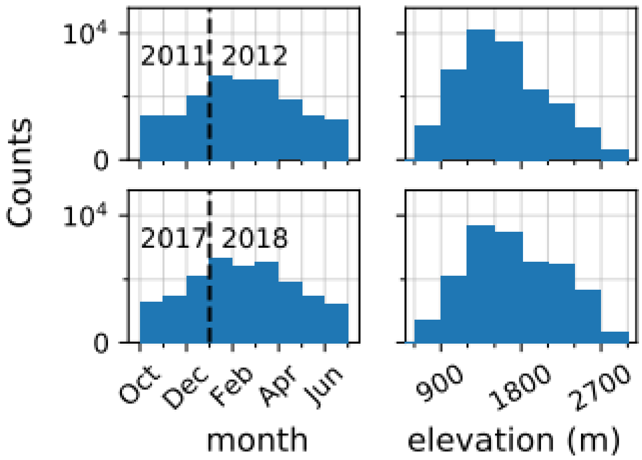
- Line 423: Delete “Then,”. OK

- Line 431: Delete “of course”. OK

TABLE AND FIGURE COMMENTS

- Figure 2 could be simplified into two panels: 2009-2015 and 2016-2018, since 2016 seems to be the most major change in the station network in the study period. Otherwise, the histograms appear to be similar. OK

Thanks for this suggestion. Figs 2 and 3 were merged into one single figure, and winters 2011-2012 and 2017-2018 were selected for display:



REFERENCES

Clark, M. P., Hendrikx, J., Slater, A. G., Kavetski, D., Anderson, B., Cullen, N. J., et al. (2011). Representing spatial variability of snow water equivalent in hydrologic and land-surface models: A review. *Water Resources Research*, *47*(7), W07539. https://doi.org/10.1029/2011WR010745

Lundquist, J., Hughes, M., Gutmann, E., & Kapnick, S. (2019). Our Skill in Modeling Mountain Rain and Snow is Bypassing the Skill of Our Observational Networks. *Bulletin of the American Meteorological Society*, *100*(12), 2473–2490. https://doi.org/10.1175/BAMS-D-19-0001.1

Vernay, M., Lafaysse, M., Monteiro, D., Hagenmuller, P., Nheili, R., Samacoïts, R., Verfaillie, D., and Morin, S.: The S2M meteorological and snow cover reanalysis over the French mountainous areas, description and evaluation (1958–2020), Earth Syst. Sci. Data Discuss. [preprint], https://doi.org/10.5194/essd-2021-249, in review, 2021.