Response letter to CC1

Dear Nima Zafarmomen,

We thank you for your comments on our manuscript, we believe that they help improve our manuscript. In the following, the comments are posted with our response in red text.

The paper addresses an important gap: bringing a flow-dependent ensemble Kalman framework to a multi-layer snow scheme for a high-latitude regional reanalysis. The topic is timely and the modelling chain (SURFEX–ISBA explicit snow + LETKF, driven by CARRA) is potentially valuable for cryospheric and hydrologic communities.

Forcing and domain

The manuscript states that CARRA forcing is “interpolated to the model grid” but omits grid spacing for both driver and land model. Domain limits are described in text but not in coordinates; include bounding box and spatial resolution.

We add the description of grid spacing in the text, and modify Figure 2 to cover the model area based on this comment.

Ensemble generation

Only perturbing atmospheric forcing inevitably under-represents uncertainty in snow compaction, albedo metamorphism, and interception. You should quantify how ensemble spread compares to innovation statistics (e.g., spread-skill ratio) to demonstrate sufficiency of the perturbation strategy. If spread is systematically low, adding multiplicative inflation alone is insufficient; process perturbations or parameter perturbations may be needed.

This is a valid concern and we agree that only perturbing forcing will not produce a large spread of the ensemble and possibly result in an over confident background state. In the manuscript we discuss how the perturbation strategy might suffer from this limitation, and that additional perturbations of the state or model parameters should be further developed. However, we argue that only perturbing the forcing data ensures that the relationship between model variables are consistent and governed by the model physics. Introducing perturbations to state variables and model parameters would require careful implementation, to avoid spurious correlations, and potentially a need for a significant increase in ensemble members to address this, which would also increase the computational demand. For a multi-year reanalysis system, we need to balance the computational cost vs the added benefit of a particular methodology. Regarding spread-skill ratio they can provide some insight in the ensemble quality. However, due to the large error of representation of in situ snow depth measurements, even a good ensemble could appear underspread. To provide the requested assessment, good estimates of observation representation errors also need to be quantified. We discuss this issue from L370 and suggest future studies on this particular topic.

The “remapping” approach for precipitation displacement is innovative, yet Appendix A lacks diagnostic evidence that the scheme produces realistic error structures. Provide at minimum a variogram or visual comparison between perturbed and reference precipitation fields.

In the revised manuscript we suggest updating Fig. 1R to show a field from the model derived precipitation and how this is remapped . The figure now gives a visual comparison between the perturbed and reference fields.

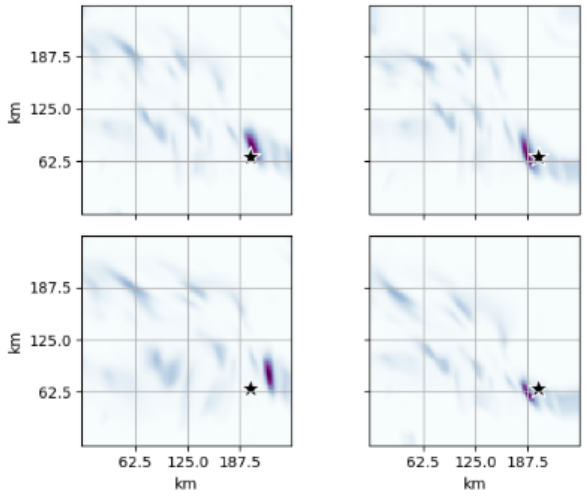

Figure 1R Precipitation field on model resolution with a small (3 member) ensemble. Upper left panel shows precipitation rate from the CARRA dataset, upper right and lower panels show ensemble realizations after remapping is applied. The star marks a reference point, for example an observation site.

Ensemble size

Ten members is very small for a 36-variable profile. You report that 20 members offered “no considerable degradation”, but give no metrics. Include a sensitivity figure (e.g., CRPS vs. ensemble size) to justify the final choice.

We agree that this is a valid concern. In the literature, this has been a recurrent topic and “small” ensembles have shown to be sufficient for land surface data assimilation. As mentioned above, long reanalysis products need to prioritize computational cost when possible. We also want to argue that while the control vector is of size 36, many of the variables are strongly correlated, as shown in the profile figure (Fig. 5) in the manuscript. Furthermore, only five (forcing) variables are perturbed and the perturbations are cross-correlated.

Increment analysis (Sect. 3.1)

Figure 4 shows domain-mean increments of several millimetres water equivalent per day—this is large. Provide histograms or spatial standard deviations to make clear whether these increments are isolated to specific subregions or pervasive. Without that context, the reader cannot judge if the LETKF is “adding missing precipitation” or merely compensating biased forcing.

We appreciate this comment and agree that this is an important detail. Figure 4 in the manuscript shows that the increments, for all variables, are largest during the spring/melting phase. Whether this is because of precipitation (missed events or bias in amount) or due to the melting process, or due to other forcing variables (eg. too warm temperature) is not easy to determine. Below we include the requested increment standard deviation maps. The maps indicate that the larger increments are found over mountain areas, and that the areas of large mean increments (Fig. 3 in the manuscript) are less pronounced in terms of standard deviation.

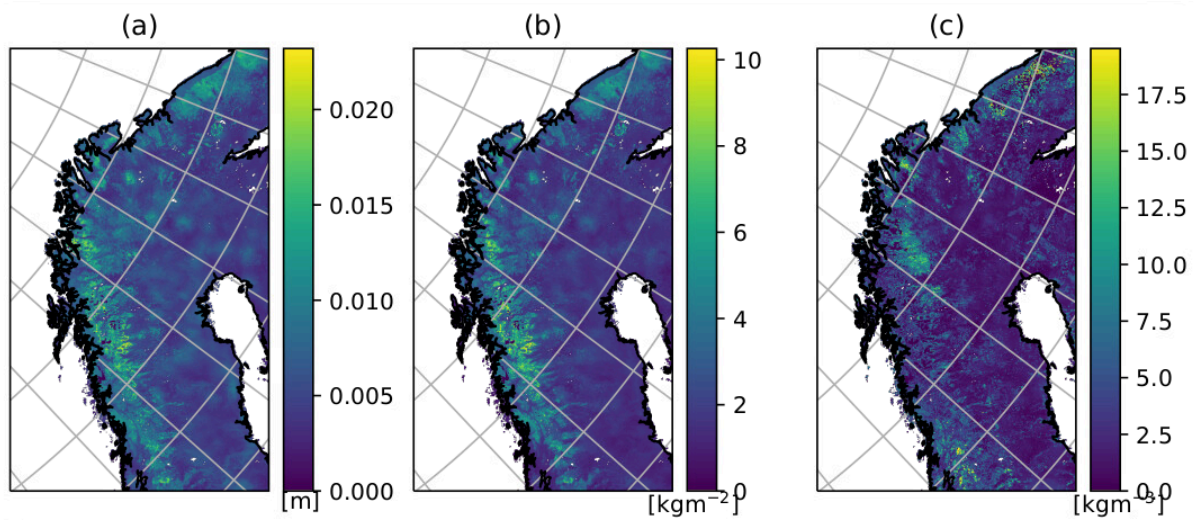

Figure 2R: Standard deviation of increments for a) snow depth, b) snow water equivalent and c) snow density.

Skill metrics

– CRPS and MAE are reported, but no sampling uncertainty is provided. Bootstrap confidence intervals would show whether the apparent improvements are statistically robust.

– Station splits (OBS-ONLY-Pv1, etc.) prove useful, yet the sample sizes differ dramatically. Present RMSE normalised by climatological variance to avoid overweighting dense station clusters.

The metrics presented in Table 3 were recomputed using bootstrap resampling. The resulting 95% confidence intervals (2.5th–97.5th percentiles) indicated statistically significant differences between the compared values, as the intervals did not overlap.

SWE validation

Only six pillow sites are available, but you can still compute Kling-Gupta or Nash–Sutcliffe across time to give hydrologists a sense of hydro-logical skill. Also, the negative bias at one degraded site coincides with orographic precipitation maxima; examine whether forcing under-catch is the root cause.

The figure below presents Kling-Gupta Efficiency scores for each snow pillow station. These results reflect the same relative performance patterns already illustrated in Figure 10 of the manuscript. While the figure may be of interest to some readers, we have chosen not to include the Kling-Gupta values in the revised paper, as they only provide limited additional insight beyond the metrics already discussed.

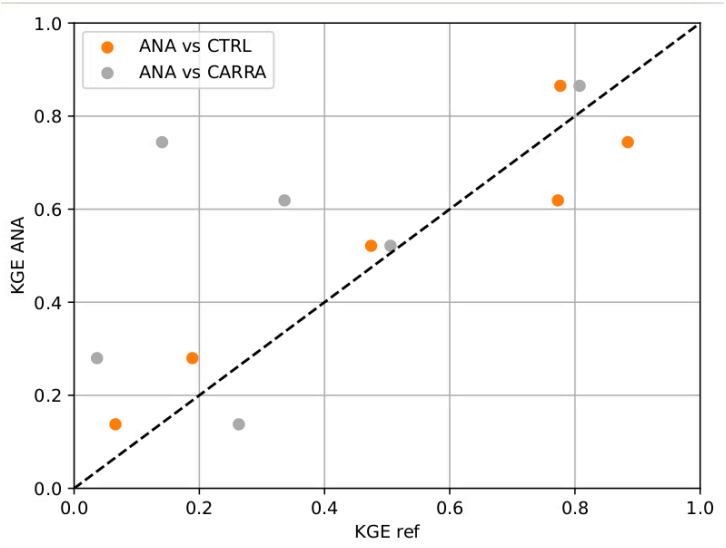

Figure 3R: Kling-Gupta scores for CARRA-Land-Pv1 (ANA) vs the reference datasets (CARRA and CTRL)

I strongly recommend that the authors expand their discussion, as data assimilation is not only applicable to snow schemes but is also widely used in other areas such as streamflow prediction. I also recommend citing the following papers: Optimising ensemble streamflow predictions with bias correction and data assimilation techniques; Assimilation of Sentinel-Based Leaf Area Index for Modeling Surface-Ground Water Interactions in Irrigation Districts