We would like to thank Dr. Andrew Newman for the thoughtful comments and suggestions. Based on the feedback, we have made significant changes to the manuscript. We believe that the article is now much improved and again appreciate the help from the reviewers. See below for our detailed responses to all comments.

Note that the reviewer's original comments are in regular black fonts and our responses are in *red italic fonts*.

Title: Role of forcing uncertainty and model error background characterization in snow data assimilation

Authors: S. V. Kumar, J. Dong, C. D. Peters-Lidard, D. Mocko, and B. Gomez

General comments: This study examines the impact of forcing uncertainty/errors on model simulations and the subsequent model error covariances and analysis increments in ensemble snow water equivalent (SWE) data assimilation in an idealized and real data case. They find that accounting for input forcing uncertainty improves both simulations. This is because without forcing uncertainty, the imposed model state perturbations are not large enough to create a realistic background error covariance matrix, and thus the model states receive too much weight relative to the observations, or result in a Kalman gain matrix of zero, so that the analysis increments are essentially zero at some update times.

Overall, the study is easy to read, follow, and the figures and analysis support the conclusions. I think that acknowledging input forcing uncertainty in land-surface and hydrological model data assimilation (DA) in a more realistic way is a key step to developing useful and robust automatic DA systems. I recommend acceptance after the authors address my comments.

Major comments:

1) It would be nice to see some type of plot of analysis increments for the various experiments in Figures 2 and 3. It is clear that changing the input forcing to Noah increases the magnitude of the background error as shown in Figure 3b, thus increasing the analysis increments so that those experiments better match the observations. Analysis increments are another useful way to diagnose what is happening in the system at each DA time, and would be a useful complement to Figure 3b, especially since analysis increments are not shown, yet discussed in many places.

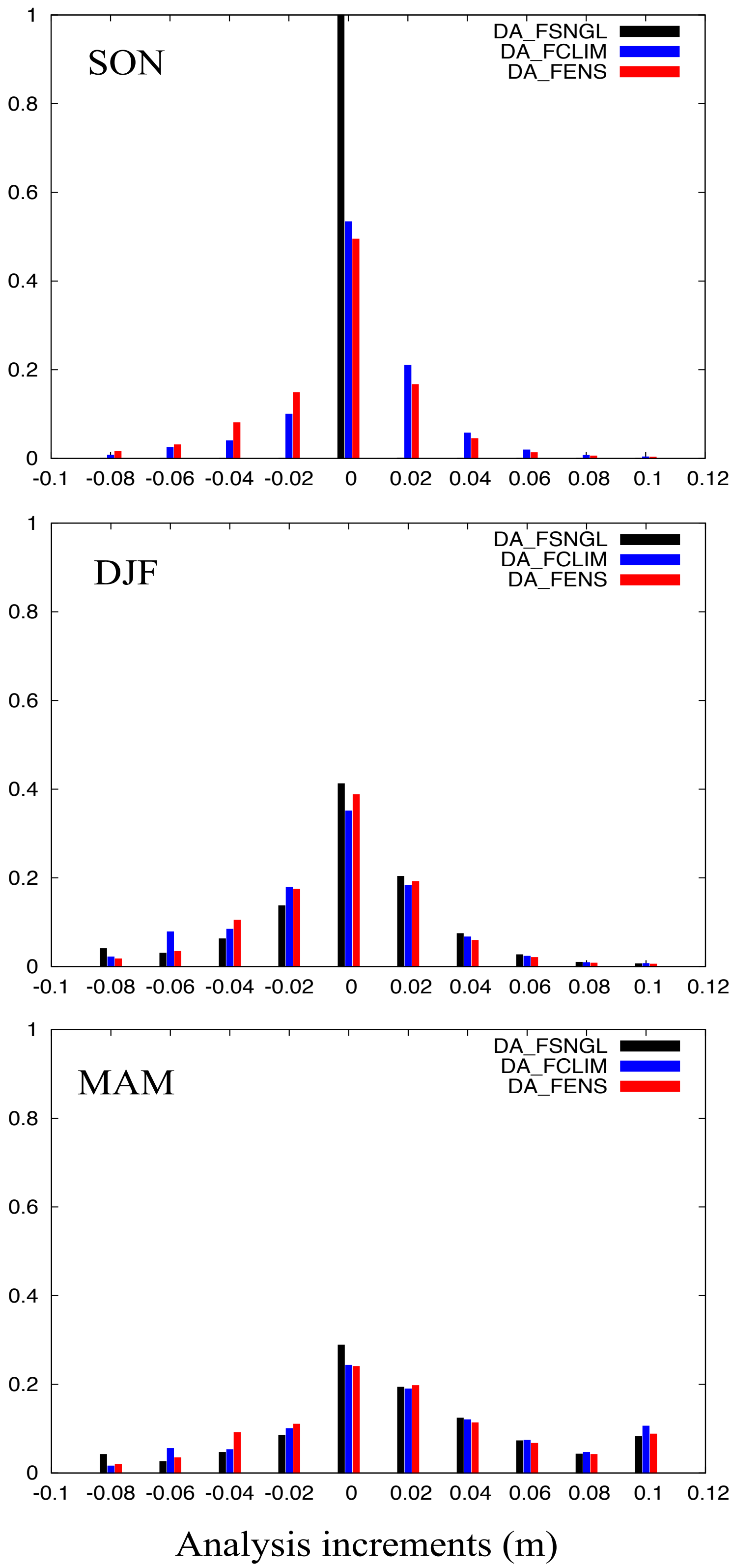
This could be particularly informative for the spatial runs in Figures 5 and 7, where the model performance has some spatial variability.

It may also be interesting to examine the spatial changes in the background error at key points during the accumulation and melt season.

Thanks for the excellent suggestion. We have updated Figure 3 to include time series comparisons of the analysis increments from the DA integrations. An additional paragraph describing these plots have been included in the text in Section 3, which reads as follows:

"Comparisons of the analysis increments from DA integrations shown in panels (c) indicate the time periods where the impact of the background model error is more significant. Generally, the analysis increments from DA_FSNGL and DA_FCLIM are similar, except during the snow accumulation and melt time periods. Comparatively, larger differences in the analysis increments between the DA_FSNGL and DA_FENS integrations are observed, with more prominent differences seen during the accumulation and melt periods. During these times, larger analysis increments are observed in the DA_FCLIM and DA_FENS integrations, reflective of the ability of these configurations to respond to observations due to the improved background model error. It can also be noted that the analysis increments during the peak snow season are generally smaller in DA_FENS and DA_FCLIM integrations compared to that of DA_FSNGL, indicating the contribution of the hybrid forcing inputs for reducing the significant biases in the assimilation system."

We also examined patterns of analysis increments in the DA integrations employing AMSR2 retrievals. Generally, the analysis increments convolve the impact of multiple factors. The analysis increments include the ability/inability of the assimilation system to respond to observations and the contribution of the hybrid forcing ensemble to correcting the biases before observations are assimilated. The Figure below show the distribution of the analysis increments for the accumulation (SON), peak winter (DJF) and melt (MAM) time periods over the Great Lakes region. During the accumulation time period, the FSNGL simulation shows little variability in its distribution (inability to respond to obs), whereas during the other two time periods, the results are more mixed (though DA_FENS generally show greater span over larger analysis increments), likely due to the combined impact of different factors. Therefore, we decided not to include the comparison of analysis increments from the AMSR2 assimilation examples.



2) Are the observations aggregated up to the model resolution for Figures 5 and 7? I believe this is a key point that needs to be clarified. The authors should describe the

aggregation method, or redo the analysis if direct comparisons to the observation points were made.

In these comparisons, observations were aggregated up to the model resolution through simple averaging. We have added the following sentence in Section 4 to clarify this point:

"The available station observations are aggregated up to the model resolution through simple averaging in these comparisons."

Minor comments:

1) Model error background seems to be a non-standard phrasing of the background forecast error covariance matrix (e.g. Hamil et al. 2001, Descombes et al. 2015). I suggest re-phrasing it background model error, or background error.

Thanks the suggestion. We have updated all such references to 'background model error', including the title.

2) It is interesting to me that the article operates with snow depth rather than snow water equivalent (SWE). Could the authors expand on this choice at all? Noah seems to have SWE as a state variable and AMSR2 does have a SWE retrieval as well, so it would be possible to operate using SWE as well, which seems like a more natural state variable to work with.

Thanks for raising this point. There are a couple of reasons for using snow depth as the retrieval variable instead of SWE. In most passive microwave retrieval algorithms (Chang et al. 1987, Kelly et al. 2003, Kelly 2009) compute snow depth first and then derive SWE by using a climatological snow density. The basic retrieval product, in other words, is snow depth. In addition, since most in-situ observations of snow are also available as a depth measurement, the use of snow depth enables a more straightforward evaluation. We have modified the text in Section 3 (first paragraph) as follows:

"We employ snow depth as the measurement variable as most passive microwave retrieval algorithms (Chang et al. (1987); Kelly et al. (2003); Kelly (2009)) compute snow depth first and derive the snow water equivalent (SWE) through a climatological snow density (Brown and Braaten (1998); Krenke (1998, updated 2004)) assumption. In addition, most in-situ observations of snow are also available as depth measurements, allowing a more straightforward evaluation of the results from the model and DA integrations."

3) Line 216: Why is the AMSR2 standard error assumed to be 50 mm when Kachi et al. (2014) cite the standard error as 20 cm (200 mm)? Is the Kachi et al. (2013) citation in the manuscript giving a different standard error than the update?

The reviewer correctly points out that in Kachi et al. (2013), the AMSR2 retrievals satisfy the 20 cm error expectation based on their evaluation against GSOD measurements. The results in that paper also indicate that there is considerable spread in the evaluation of AMSR2 retrievals. In addition, the evaluation was limited to a single water year (2012-2013). In the paper, we use a higher value of standard error, based on the snow DA literature, which generally indicate low skill for passive microwave snow depth retrievals. The higher error standard deviation assumed here is consistent with prior snow DA studies (Liu et al. 2013, Liu et al. 2015, Kumar et al. 2014). We have added the following acknowledgement within the article:

"Note that we use a higher value of observation error standard deviation than that reported by Kachi et al. (2013), based on the previous snow DA studies (Liu et al. (2013, 2015); Kumar et al. (2014, 2015)) that generally assume low skill for passive microwave snow depth retrievals."

4) Line 218: Was the model resolution of 25km selected to match the approximately 30-km footprint of AMSR2? If so, it would be good to state that.

We use a model resolution of 25km, as two key near-real time global modeling environments, the NASA Global Land Data Assimilation System (GLDAS) and the U.S. Air Force 557th Weather Wing operational land data assimilation system) that use LIS are conducted at approximately 25 km resolution.

5) Line 280, change stronger to larger. The authors may want to check the entire paper for instances of this.

Thanks for the suggestion. All such instances have been corrected.

6) Lines 306-308. The two sentences starting with "Though underestimated" and ending with "data assimilation updates" are confusing to me. What are authors trying to describe here?

We agree that these sentences are confusing. The entire paragraph has been rewritten a follows:

"In general, the DA integrations (DA_FSNGL, DA_FCLIM and DA_FENS), have comparable performance at both these locations and they mostly follow the snow evolution patterns in the AMSR2 data. Note that though AMSR2 observations capture the seasonality of snow observations, they show significant underestimation compared to in-situ observations of snow depth. The influence of undersampling the model error background can be observed in the early part of the snow season at location C and during late season at location D, where the DA_FSNGL integrations fail to match the snow events captured by AMSR2. During the peak snow time periods, however, the undersampling of model error background in OL_FSNGL is less of a problem over this domain, as the non-zero model snow states provide an adequate background for

subsequent data assimilation updates. Thus, the evaluation of the snow DA integrations at these two regions …"

7) Figures 5 and 7: The authors may want to consider having a gray color that spans zero as small error differences are likely not significant. The figure is nice as it is with the lighter shades near zero; this is merely a suggestion to look into.

Thanks for the suggestion. We have redone Figures 5 and 7 to have an improved color scheme with a gray color spanning small error magnitudes.

8) The authors may be interested in the article Huang et al. (2016) that is in press in HESS. This article uses an ensemble of forcing data to generate and ensemble of internally consistent (with the forcing traces) initial model states uncertainty for EnKF SWE assimilation. They examine the impact of the relative weighting of the model and observational error covariance matrices. They also find similar results to those stated on lines 365-368 as well, if the open loop simulation has high quality forcing, DA is less beneficial. I am not suggesting the authors need to cite this paper, as I am a co-author on it; it just seems to be very relevant to the study reviewed here and some of their discussion points.

Thanks for the suggestion about this appropriate reference. We have modified the text in Section 5 as follows:

"If the single forcing dataset being used is of high skill, then the added benefit of using the forcing ensemble is likely to be less, consistent with the results of more recent studies to employ an ensemble of forcing data for generating an ensemble of internally consistent model uncertainty representation for applications such as DA (Newman et al. (2015); Huang et al. (2017)). Overall, the results in this article indicate …."