Responses to Referee #3

[Note: All original comments by the referee are reproduced in their entirety below in regular black text. Our responses to these comments are shown immediately after the comment in blue text.]

The paper presents a multi step approach to generate ensemble streamflow forecasts where multiple sources of uncertainties are taken into account and merged into one ensemble: forcing errors, optimized model parameters and errors (ISURF), and initial conditions errors via SWE assimilation, all merged with an EnKF. In particular, each uncertainty type (mostly model parameters and initial conditions here) is individually characterized and merged into one "big ensemble" via an EnKF. The approach is evaluated with respect to an EnKF without the individual characterization of the uncertainties (blend of undifferentiated sources of errors) and standard operational performance (removing the human forecaster's adjustment skill). The approach is validated over the American River Basin, CA.

The paper is very well written with a great and extensive literature review and justification of choices. The approach is a great contribution to the scientific community effort; characterize individual uncertainties of different sources and merge them in a single but meaningful ensemble in an effort to assess modeling uncertainties in a more accurate way.

We thank Dr. Voisin's for her comments on the value of our work.

The analysis currently assesses if the simulations are more accurate which allows a comparison with the deterministic NWS forecasts. But it does not assess if the uncertainty is better assessed. I would recommend accepting the paper with major revisions. The paper would benefit from a couple of additional explanation of the performance and limitations of the approach, and most important a schematic diagram for clarity. Also, it is presently difficult to isolate the performance of the data assimilation with the overall water-year long approach performance. The results analysis would benefit from using another measure than the dispersion only – it needs to be the dispersion with respect to the observation (is the ensemble representative?) in order to assess if the uncertainty is better assessed than a blended uncertainty (ICEA vs EnKF analysis only). See specific comments below.

We appreciate Dr. Voisin's insightful comments for us to further improve the manuscript. We detailed the responses to these comments as follows.

Specific comments

1/ add a schematic diagram that explains the chain of models and processes, and the variables being transferred (single value, or ensemble), the time scale of the analysis etc. For example: Observed precipitation, temperature, PET -> SNOW17; i)ISURF– optimized parameters and uncertainty), ii) EnKF for SWE assimilation and merging parameters uncertainties ->precip and snowmelt ensemble, PET (single value?)-

>SACMA -> ensemble streamflow forecast to be verified with respect to observed flow, for several days after the assimilation. The assimilation is performed every 7 days. Etc.
We have described the modeling and assimilation procedures in detail in Section 3.2. However, we agree with the referee that diagrams showing the processes would be further illustrative. We thus added the following figures to the manuscript.

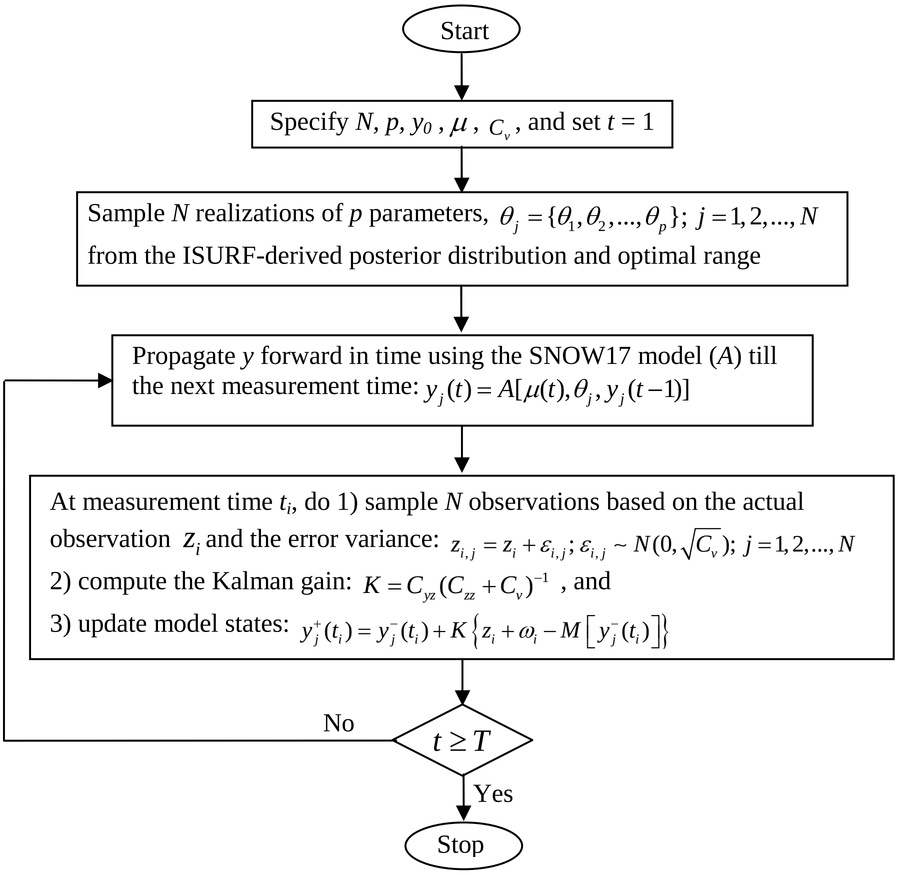

Start
Specify N, p, y_0 , μ , C_v , and set t = 1
Sample N realizations of p parameters, $\theta_j = \{\theta_1, \theta_2, ..., \theta_p\}$; $j = 1, 2, ..., N$
from the ISURF-derived posterior distribution and optimal range
Propagate y forward in time using the SNOW17 model (A) till the next measurement time: $y_j(t) = A[\mu(t), \theta_j, y_j(t-1)]$
At measurement time t_i, do 1) sample N observations based on the actual observation z_i and the error variance: $z_{i,j} = z_i + \varepsilon_{i,j}$; $\varepsilon_{i,j} \sim N(0, \sqrt{C_v})$; $j = 1, 2, ..., N$
2) compute the Kalman gain: $K = C_{yz}(C_{zz} + C_v)^{-1}$, and
3) update model states: $y_j^+(t_i) = y_j^-(t_i) + K\left\{z_i + \omega_i - M\left[y_j^-(t_i)\right]\right\}$
No
$t \geq T$
Yes
Stop

Fig. 2. Flowchart of the EnKF applied in ICEA to recursively update SNOW17 model states with the uncertainty of sensitive model parameters considered. N and p indicate the ensemble size and the number of sensitive parameters, respectively; y_0 and μ represent model initial condition and forcing, respectively; z_i and M designate the observation and measurement operator, respectively; C_v , C_{yz}, and C_{zz} denote the variance of observation error, the covariance between model states and observations, and the variance of observations, respectively.

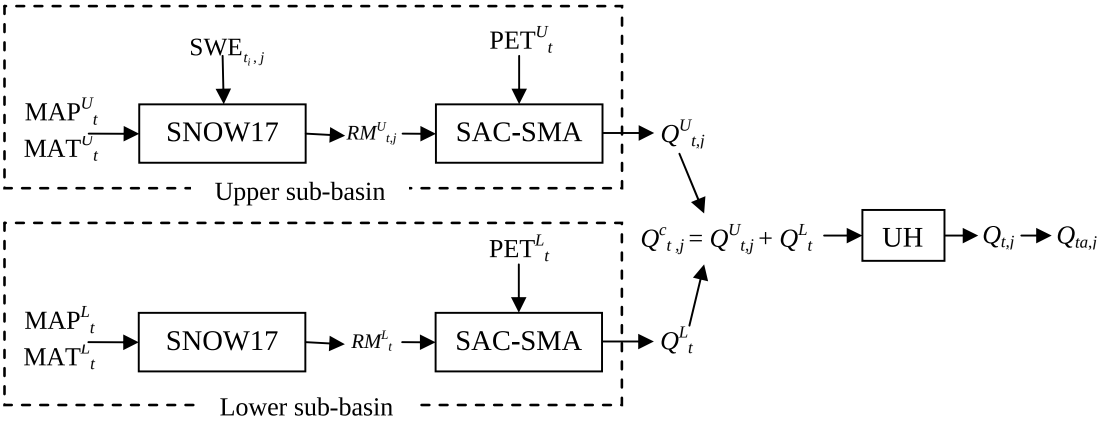


Figure 3. Flowchart of modeling and assimilation (of areal SWE) procedures in generating ensemble daily streamflow. *RM* and *Q* represent rain plus snowmelt and streamflow, respectively; *UH* stands for unit hydrograph; *t* indicates time step (t=1,2,...,T; 6 hourly) ; *ta* is the daily time step aggregated from 6-hourly step; t_i reprsents the measurement time when areal SWE is assimilated (refer to Figure 2 for detailed assimilation procedure); *j* is the ensemble number (j = 1,2,…,N); subscripts *U* and *L* denote variables for upper and lower sub-basins, respectively; subscript *c* designates the combined variable.

2/The forcing uncertainty is accessed via a proxy by SNOW17 parameters. It seems to me that if SNOW17 was driven by an ensemble weather forecast, the SCF is adjusting the forcing in a way that the uncertainty information is decreased, unless the ensemble weather forecast was calibrated with respect to a specific observed meteorological dataset knowing the SCF to be used. As such, I am not sure how much of the forcing uncertainty is really taken into account. Similarly, it means that in a full probabilistic approach where ensemble weather forecasts were to be used, the approach does not allows yet to merge forcing uncertainty with model parameterization and initial conditions uncertainties. This being said, this approach allows generating an ensemble merging initial conditions and parameters uncertainties with an estimate of forcing uncertainty. As far as I know, this is a first on how forcing uncertainty can be merged in an ensemble with parameters uncertainties and initial conditions errors.

In this study, the SNOW17 is not driven by ensemble weather forecasts. Instead, it is driven by deterministic historical observations of precipitation and temperature (Lines 9-11, Page 7716). The current version of ICEA (presented in this study) is not designed to digest ensemble weather forecasts of forcing data. For future versions of ICEA to ingest ensemble forcing, the SCF will have to be fixed at a calibrated value (not to introduce biases to the ensemble forcing). However, the uncertainty of other SNOW17 parameters can be determined in the same way (as the current ICEA does). And that uncertainty can be merged into an ensemble with forcing uncertainty and initial condition uncertainty.

3/ metrics used for the evaluation: the annual bias, the correlation, the RMSE, NRR and UR95.When presenting the results about the dispersion, I would suggest making sure that not only the dispersion is being discussed as what matters is if the dispersion represents the observed variability. The authors mention in the discussion that other metrics could be used to assess the reliability. I would suggest using some of them here, like rank histogram for example. They would help assessing if the dispersion is fast enough for short lead times, and if the information in the ensemble is right, because too large of an ensemble has no value, too narrow either. It needs to bracket the observation in a representative way (uniform histogram). I believe that this is another important component of your approach – you need to evaluate if the uncertainty is better assessed when individually characterizing the uncertainties, or if blending them drive to the same result.

In this study, we evaluated the ensemble streamflow predictions via both statistic metrics (i.e., NRR and UR95, as presented in Table 6 and Figure 9) and visual inspection (i.e., hydrograph of the wettest year as presented in Figure 7). Both NRR and UR95 have been widely applied in evaluating the dispersion of ensembles (see references listed in Section 3.5). NRR is a measure of ensemble dispersion relative to the deviation of the ensemble mean. UR95 is a measure of the aggregate variability of the 95th percentile prediction range relative to the observations. In addition, we evaluated the ensemble mean prediction in terms of other metrics including correlation, bias, RMSE, and NSE (i.e., Table 5, Figures 5 and 9). We currently have a paper in progress that evaluates performance of the ICEA against traditional operational methods in the NFARB using a range of probabilistic metrics (including rank histogram and others).

This study aims to evaluate a first version of the ICEA in the context of providing streamflow predictions. In this preliminary study, we are particularly concerned about the width of ensemble dispersion and its variability relative to the magnitude of observations. We believe that this information, combined with the metrics quantifying the performance of ensemble mean predictions as well as a visual inspection of the ensemble predictions in the extremely wet year, produces an initial picture of the performance of the ICEA. We thus deem that the ensemble metrics employed serve the purpose of this study and more detailed assessment being undertaken can be provided in more detail in our next publication on this work.

4/ The results are presented for the entire WY. It is difficult to isolate the performance due to the SWE assimilation on top of the approach and the one due to the approach when the SWE assimilation is not in used. What is the performance of the system for the snow period only - overall. What is the performance of the snow-free period (hear glacial melt instead of perennial snow melt if applicable)? It is common to look at the 95th percentile for looking at figure 6, it is not obvious if low flow/average flow has improved. Please comment as to here add again about the overall performance of the approach and its best application / limitations.

The study basin typically has snow cover (no glacier) from Nov. to June (Lines 18-20, Page 7715) when most of the annual precipitation occurs. The snow-free period over this basin is the summer and early fall (e.g., July – Oct., refer to Figure 8a as an example). During the snow-free period, the flow is generally low (e.g., refer to Figure 7 as an example), which has marginal contribution to the total annual flow volume.

Operational hydrologic forecasting generally focuses on two variables: the high flows (extreme events) and the total volume of flow (water supply). That is why we presented, in the original manuscript, the predictions on high flows (Figure 6) and the statistics (including the bias) between predicted and observed flow on both annual (Table 4) and inter-annual scale (Figure 5). In addition, we also presented the results at different lead times from day 1 up to day 7 (Figure 9, which is not on WY scale). We deemed these results well illustrate the performance of ICEA in producing predictions on high flows and total flow volume.

5/ the ISURF approach allows defining the optimized parameter sets, with their uncertainty structure. This is still more or less equivalent to a calibration prior to the data assimilation approach, which can affect the water balance. How does it theoretically affect the parameter uncertainty structure? And operationally, it might be okay (is it?) to not meet the annual water balance as long as we are getting the next few days peak flows right. It would contribute to the description of the approach (performances and limits of applicability discussion) to comment on it.

In operational hydrologic forecasting, reliable streamflow predictions should mimic the actual (observed) streamflow in the context of high flows (particularly the peak flow) and the total flow volume (e.g. the water balance as the referee indicated). The way ISURF determines posterior parameter uncertainty information in this study is conditioned on minimizing the squared residuals between predicted and observed streamflow. A minimum of squared residuals generally indicates a minimum overall bias in total flow and a best match of high flows. Therefore, the ISURF-derived parameter sets lead to predictions resembles the observations in terms of both total volume and high flows. To make this point clear, in Line 17, Page 7717(before “In DREAM…”), we added the following sentence “This is accomplished by minimizing the sum of squared residuals between model-predicted and observed variables.”

6/ Why did you choose one week for the frequency of the analysis? Others have use 3 days for example (Clark and Slater 2006), agreeing that the prior distribution does not necessarily change that fast. I would think though that during snowmelt period, when snow depletes faster, the frequency of the assimilation should not be any longer than the time of concentration of the basin in order to avoid any incoherency in the flow and in the ensemble flow forecast characteristics, i.e. about 3-4 days over the American? For example, it would help looking at the ensemble/dispersion over a continuous period of time; i.e. day 7 might display a large dispersion and then assimilation is performed and the ensemble dispersion narrows again? Please explain or show a time series of the ensemble dispersion over a period of time.

The selection of one week for assimilation frequency was undertaken to mimic the operational environment. Specifically, in operations (e.g., in western River Forecast Centers), SWE is typically assimilated into the coupled SNOW17/SAC-SMA model on a weekly or even bi-weekly basis via a direct insertion method, coupled with subjective judgments from experienced forecasters (Donald Laurine (Development and Operations Hydrologist of NWRFC), personal communication, 2010). This fact inspired the usage of one week as assimilation frequency in this study. To make this point clearer, we changed the sentence “The assimilation frequency is one week.” (Line 17, Page 7719) to be “The assimilation frequency is set to be one week to mimic the operational environment.”

As for the change of ensemble dispersion over a period of time, in real-time ensemble forecasting, the dispersion generally increases in magnitude with increasing lead time. This is usually due to the fact that the accuracy of forecasted forcing (precipitation and temperature) decreases with increasing lead time. This study, however, does not use forecasted forcing. Instead, historical MAP and MAT data (and simulations) are used. The accuracy of forcing is thus consistent and not dependent on the lead time. As such the ensemble dispersion is not lead time-dependent, as can be told from Figures 7b and 7c. Of course, we realized that in the future, we would apply the enhanced ICEA in a real-time environment. We therefore presented relevant discussions in Lines 14-16, Page 7731 and Lines 10-11, Page 7732 in the original manuscript.

Technical revisions:
Why 6 years of training and 6 years of validation when 23 years are available?
This study is the first study of ICEA. The selection of 6-year training period and equivalent length of prediction is to demonstrate the applicability of this method in streamflow prediction. These years selected are representative of the entire period, as presented in Figure 2 and discussed in Lines 17-19, Page 7716.

P7716, line9: what is the model providing the meteorological forcing to SNOW17/SACMA?
This study is not a real-time case study. All the forcing applied to SNOW17/SAC-SMA are not forecasted data produced from numerical weather models. Instead, historical observed mean areal precipitation (MAP) and temperature (MAT) data are used. As such, these data are archived observations (not provided by any models). To make it clearer, we changed the statement in Line 9, Page 7716 from “…including the MAP, MAT…” to be “…including the historical MAP, MAT…”

p7716 line 11: is the observed daily flow regulated?
The observed flow is not regulated at the point of our simulations (USGS gage # 11427000).

How do the uncertainty in the forcing compares to actual short or medium range ensemble forecasts?
Currently no short to medium range ensemble forcing forecasts are applied in operations at NWS River Forecast Centers (RFCs). Instead, RFCs generally use single-valued precipitation (QPE) and temperature (QTE) forecasts produced by NCEP in streamflow forecasting. Yet NCEP also provides GFS/GEFS ensemble precipitation and temperature ensemble forecasts, ingest of these products into real-time hydrologic forecasting is in the experimental stage at OHD and not available at RFCs yet. As such, there are actually no benchmark practical ensemble forecasts to compare with.

P7721, line17: How are the station weights computed? Or give a reference.

There is actually a typo in Equation (5). We have changed the following part "Specifically, the areal SWE is calculated via a non-negative least-squares algorithm as follows

$$\mathrm{SWE}_{\mathrm{areal}} = \min \left\| \sum_{t=1}^{T} \left[\sum_{k=1}^{3} (C_k \times \mathrm{SWE}_k^t) - \mathrm{SWE}_{\mathrm{model}}^t \right]^2 \right\|, \quad \forall \mathrm{SWE}_k^t \geq 0 \; \ldots(5) \quad "$$

to be

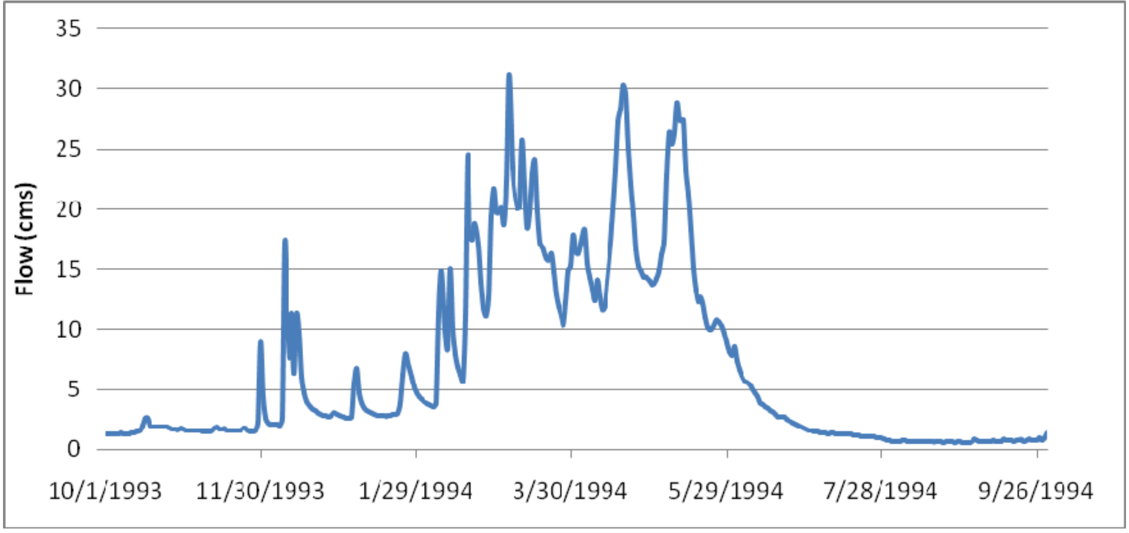
"Specifically, the areal SWE is calculated as a linear combination of the SWE observations from three SNOTEL sties. The weight associated with each SNOTEL site is determined via a non-negative least-squares algorithm as follows:

$$\min \left\| \sum_{t=1}^{T} \left[\sum_{k=1}^{3} (C_k \times \mathrm{SWE}_k^t) - \mathrm{SWE}_{\mathrm{model}}^t \right]^2 \right\|, \quad \forall \mathrm{SWE}_k^t \geq 0 \; \ldots(5) \quad "$$

The weight for each SNOTEL site is calculated via the updated Equation (5). The equation aims to mimic the operational way in determining the areal SWE, as we described in Lines 10-12, Page 7721.

P7732, line1: I would suggest substituting "current" with "automatic" as the current prediction depends heavily on the human forecasters who make a difference, as seen on the statistics used for this analysis and those seen on the RFC website.
We agree with the referee that the current prediction involves real-time adjustments from forecasters. We changed "current" to be "automatic".

P7743, table 1; It seems there a typo for year 1994 peak flow, should be higher
In fact, the peak flow of water year 1994 is 31.14 cms. The hydrograph of this year is depicted as follows.


Flow (cms)
10/1/1993
11/30/1993
1/29/1994
3/30/1994
5/29/1994
7/28/1994
9/26/1994