Responses to Referee #2
[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.]

This study presents a procedure for combining Bayesian parameter estimation with data assimilation to improve operational streamflow forecasts. The Integrated unCertainty and Ensemble-based data Assimilation (ICEA) method combines the previously presented ISURF (combination of GSA and DREAM to estimate parameters) and data assimilation with the EnKF. The authors had developed this method to be applicable to operational forecasting to increase the skill of predictions. This is applied to the SAC-SMA/SNOW-17 models to compare performance with operational forecasts. The method utilizes SNOTEL and CDEC observations of SWE to improve streamflow estimates. Overall the paper shows improvement over current operational forecasts but several issues need to be addressed (clarified and corrected) before the paper is ready for publication. Therefore, I recommend major revision of the manuscript given the comments below:
We appreciate the referee's comments which helped to improve the manuscript. We detailed the responses to these comments as follows.

Comments:
1.) Page 7719, Line 17 states that assimilation is only performed once a week. Why is the assimilation only performed once a week when the SNOTEL observations are typically daily?
Indeed SNOTEL observations are on daily or even hourly (for some sites) time step and any relevant timestep data could be assimilated into the developed framework . However, in operation (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."

2.) Page 7720, line 3 states that the second scenario used "ISURF-derived optimal model parameters" but ISURF estimates parameter distributions not the optimal parameter set. How is optimal defined here (e.g. mode, mean)?
Except for providing posterior distribution information, ISURF also provides likelihood information associated with each posterior parameter set. The optimal parameters here refer to the parameter set with the maximum likelihood. To make this point clearer, we changed "ISURF-derived optimal model parameters" to be "ISURF-derived parameter set with the maximum likelihood".

3.) Equation 5 appears to have an error. If the difference between the SNOTEL and modeled SWE is minimized, this will not produce an areal estimate of the SWE, as suggested in this equation. It seems from this method that the sum of the SNOTEL SWE multiplied by the weights is the areal SWE estimate but this needs to be clarified in the manuscript.

We agree and 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:

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

4.) Section 3.3 describes the method for estimating the areal SWE for the upper elevation band. This method combines SNOW17 model estimates (with prior RFC parameters?) and SNOTEL observations to estimate the spatially averaged SWE. The Ck values are then estimated based on the model estimates and the in-situ observations. This makes the SWE values highly dependent on the SNOW-17 model. The model dependent SWE observations are then used for calibration and assimilated into the model. In my opinion this is very problematic because the model is calibrated, in part, to the prior model runs and not solely on observations. In addition, it cannot be ensured that the estimate of SWE is representative of the spatial average and thus an accurate calibration and assimilation will not necessarily lead to more accurate stream-flow forecast. A further clarification of the technique with justification or a method for estimating areal SWE independently of the model is necessary.

We would like to highlight that the way we determine the areal SWE (for assimilation into the model via EnKF) mimics the operational way of assimilating SWE information into the model (via the direct insertion method). More specifically, the operational direct insertion method does not use the SNOTEL-derived SWE (gauge_SWE) to replace the entire SNOW17-simulated SWE (model_SWE) and then use the former as new model states to run the model forward. Instead, the new SWE is generally a weighted value from both gauge_SWE and model_SWE. Put it in a different way, forecasters trust neither model_SWE nor gauge_SWE completely but believe that both of them contain meaningful information. In this study, since we are not using the direct insertion method (which inserts weighted SWE into the model), we need to come up with a technique to get the modeled SWE into play along with the SNOTEL SWE (as in the operational environment). That is why we use the current technique, as we specified in Lines 9-13 on Page 7721. Though this technique might not be perfect, it mimics operational purposes. Additionally, we would like to clarify that SWE information is not used in calibration of SNOW17 model at all RFCs in U.S. In most cases, observed streamflow is solely applied to calibrate the SNOW17 model coupled with the SAC-SMA model.

As for the representativeness of the areal SWE derived, we would like to highlight that SWE has so much variability in space that there are actually no widely agreed upon representative observations. Specifically, in-situ observations (e.g. SNOTEL SWE) contains limited information on the spatial variability of SWE. As an example (also as discussed in Lines 19-21, Page 7720), there are over 1700 snow stations which provide SWE observations in the Western U.S. However, they are still not adequate to resolve the variability of SWE at the basin scale (Bales et al., 2006). While remotely sensing techniques are able to provide continuous spatial distribution of SWE, they are generally suffering from poor resolution and thus quality, as reviewed in section 3.3. In fact, none of those (in-situ and remotely sensed) products can be claimed as spatially "representative" and thus are not yet being widely applied in operations. In the study, we set out to mimic the operational method of deriving areal SWE, however, we do not deem the SWE produced is the "truth" or more accurate than other spatial SWE estimates. Instead, we assigned an error term to it when assimilating it into the model via the EnKF (as specified in Lines 1-2, Page 7722). In addition, we further discussed alternative options in defining this error term in Section 5 (Lines 21-26, Page 7732).

5.) In paragraph 2 of section 3.4, the method of using the SCF and PXTEMP parameters to account for forcing error is described. This technique adds noise to these two parameters to account for errors in forcing data. From my understanding in He et al. 2011a, this is not actually noise added to the parameters but an estimated posterior distribution from DREAM. If this is the case here, the forcing uncertainty will likely be underestimated. Since the DREAM algorithm works in a batch framework, the uncertainty in these parameters (and thus the forcing error) will be estimated over the entire length of the batch instead of daily, as is more commonly performed. Estimating the SCF across the entire batch length will find the uncertainty relating to the average error of the precipitation measurements as opposed to the daily measurement errors. It is likely that the daily precipitation measurement error has a much greater variance than its long term average uncertainty, which will likely lead to degraded prior distributions.
In my opinion, the added ease of application through assuming the SCF and PXTEMP parameters handle forcing error is not worth risking the potential problems associated with this approach. Further, I would suggest examining this method as compared to the traditional methods of adding forcing error to ensure that this method is not significantly altering the results.

We would like to clarify that the technique applied in this study is not adding "noise to these two parameters to account for errors in forcing data", but applying the same technique as presented in He et al. 2011a which produces posterior uncertainty distributions of these two parameters. We realized that the description "we perturb parameters SCF and PXTEMP" (Line 26, Page 7722) might be confusing. To make it clearer, we changed the last three sentences of Section 3.4 from

"Hence, instead of perturbing precipitation and air temperature timeseries, we perturb parameters SCF and PXTEMP and assume that the uncertainty identified for these two parameters implicitly represent the uncertainty in precipitation and air temperature. This method is relatively easier to understand in concept and requires no explicit quantification of the distribution type of precipitation and air temperature. This method has been recently applied in defining SNOW17 model forcing uncertainty (He et al., 2011a)."

to:

"Hence, instead of perturbing precipitation and air temperature timeseries, we determine the uncertainty of parameters SCF and PXTEMP and assume that the uncertainty identified implicitly represent the uncertainty in precipitation and air temperature. This is achieved by applying the same method used in defining the uncertainty of these two SNOW17 parameters in one of our previous studies (He et al., 2011a). This method is relatively easier to understand in concept and requires no explicit quantification of the distribution type of precipitation and air temperature."

In the meantime, we agree with the referee that estimating SCF uncertainty range based on long record likely makes the estimated range conditioned more on average error of observed precipitation other than errors in individual observations. Applying the traditional methods to add noise to (directly perturb) the observed precipitation might provide more realistic estimates on precipitation uncertainty and thus provide further improved streamflow predictions. However, there is no widely agreed upon method which provides most "realistic" estimate on precipitation uncertainty in research community yet. In the operational environment, forecasters prefer tools easy to understand in concept while providing predictions in satisfactory quality (e.g., perturbation of forcing data would be very confusing to forecasters based on the first author's extensive experience communicating with them). As such, in this study, our intent is to come up with a method easy for forecaster to understand and apply, other than comparing various methods and trying to pick up the "best" one (without universal agreement from the research community though) for forecasters. More importantly, the ICEA as is (with simplification on handling precipitation error) is demonstrated to provide improved predictions, which is what the forecasters are interested in. Nevertheless, from a researcher's perspective, we agree with the referee that there are definitely potential values of the traditional perturbation methods. These values are worth being explored and will be evaluated in the future work.

In light of above discussions invoked by the referee's constructive comment on precipitation uncertainty, we added the sentence "Moreover, the uncertainty of parameter SCF derived from ISURF represents only partial uncertainty of precipitation. This is due to the fact that ISURF is applied over the entire training period to produce the uncertainty range and distribution information of SCF. Thus, the uncertainty of SCF is more representative of the average precipitation error in the training period rather than errors associated with individual precipitation events. To more comprehensively account for uncertainty in precipitation, it may be necessary to implicitly consider the error for each precipitation event. This could be achieved by assigned a random multiplier (with a certain distribution) to observed precipitation at each time step following previous studies (e.g. Margulis et al., 2002; Leisenring and Moradkhani, 2011)." after the references in Line 26, Page 7733.

Lastly, we would like to highlight that we are presenting a first version of ICEA, evaluating its viability in streamflow prediction against the current prediction, and identifying the potential enhancements (as discussed in the last paragraph of Section 5) for future versions of ICEA. As illustrated by the results, this first version of ICEA has already shown advantages over the current prediction, even with several simplifications specifically designed to mimic the operational environment (e.g., sparse assimilation frequency, parsimonious methods utilized to determine areal SWE data and forcing uncertainty as also mentioned in the previous paragraph). This inspires us to believe that an enhanced ICEA would provide further improved streamflow predictions, which is being explored in our ongoing work (as also highlighted in the last paragraph of Section 5).

6.) Page 7724 line 7 states that the UR95 has a perfect score of 0%. This statement is not entirely correct because an uncertainty ratio of 0% indicates no uncertainty is estimated. In any practical scenario, there would be some uncertainty, due to forcing, model, parameter and observation error, and therefore an UR95 of 0 % will be an overconfident prediction.

We agree with the referee that a UR95 value of 0% indicates no uncertainty is estimated. In practice, it is unlikely to produce such a value. Accordingly, we changed the statement "UR95 is a measure of the aggregate variability of the 95th percentile prediction range relative to the observations (Moradkhani et al., 2006). It ranges from 0 to 100 %, with a perfect score equal to 0 %." (Lines 5-7, Page 7724) to be "UR95 is a measure of the aggregate variability of the 95th percentile prediction range relative to the observations, generally ranging from 0 to 100 % (Moradkhani et al., 2006; Leisenring and Moradkhani, 2011)."

7.) Page 7729 line 24 states "from day 230 to day 252, the EnKF ensemble is much wider than ICEA ensemble" but it is not mentioned that during this time the EnKF ensemble encompasses the observation while the ICEA ensemble does not. This means that the EnKF actually performed better than ICEA during this period. This is followed by the statement "from day 265 to day 289, the ICEA ensemble reasonably captures the recession pattern while the EnKF ensemble follows the variation of RFC predictions which deviate from the observed streamflow". Though the ICEA is closer to the observation during this period, it appears that the observation is outside the ensemble of the ICEA for the majority of this period. For this reason, I disagree that the "ICEA ensemble reasonably captures the recession pattern".

We agree with the referee that the EnKF outperforms the ICEA from day 230 to 252, while the latter performs better the former from day 265 to 289. We also agree that our current description on this can be improved to avoid any potential confusion. Specifically, we changed "First, from day 230 to day 252, the EnKF ensemble is much wider than ICEA ensemble; second, from day 265 to day 289, the ICEA ensemble reasonably captures the recession pattern, while the EnKF ensemble follows the variation of RFC predictions which deviate from the observed streamflow with a negative bias." (Lines 24-27, Page 7729) to be "First, from day 230 to day 252, in comparison to the ICEA ensemble, the EnKF ensemble is much wider and well encompasses the observations; second, from day 265 to day 289, the ICEA ensemble reasonably captures the variation pattern of streamflow observations in this period, while the EnKF ensemble follows the variation of RFC predictions which deviate from the observed streamflow with a negative bias."

8.) Page 7730 lines 14-17 explains that after day 265 the EnKF SWE melts more rapidly than the ICEA SWE leading to poorer performance from the EnKF than the ICEA in terms of streamflow. However, the EnKF appears to match the observed SWE more closely than the ICEA throughout the melt season. Given that the ICEA performs worse in matching the observed SWE but better in matching the streamflow, it is likely that the SWE used for assimilation is not representative of the true basin SWE (this relates back to problems suggested in comment 3). This would explain the poor performance of the EnKF in terms of streamflow despite a relatively accurate assimilation of SWE. Once again I find it necessary to show that the method for spatially averaged SWE generation is representative of the true basin SWE.
We would like to highlight again that it is hardly possible to evaluate whether a SWE product is "representative" or not (and thus best suitable for operational applications) since there is no widely agreed upon "true" basin SWE available. In this study, we aimed to mimic the operational way (of NWS RFCs) in generating areal SWE and evaluate the applicability of ICEA in streamflow predictions by assimilating this SWE information. And results showed that ICEA did provide improved predictions. Different methods can be employed to derive different areal SWE products (how to ensure they are "representative"?) which might lead to further improved predictions of ICEA, but this is out of the scope of the current study and could be a topic for further research. In fact, we discuss the uncertainty in areal SWE produced in the study and proposed an alternative method to better account for this uncertainty in the conclusion section (Lines 21-26, Page 7732).

As for the performance of both EnKF and ICEA after day 265, both techniques merge observations and model simulations to provide updated model states. If the technique identifies that observations are more erroneous than model simulations, it puts more weight on the former, and vice versa (reflected via the Kalman Gain). The descriptions in Lines 14-17, Page 7730 (as the referee highlighted in this comment) indicate that ICEA is more skillful in merging the observed information and modeled information than the EnKF, given the same observation information available (in operations) even this information is flawed. To reiterate, our intention was to mimic operational methods in producing observation information rather than producing the most representative information (which can not be verified in reality and is never being applied in operations).

9.) Page 7731 lines 7-10 states "the whole EnKF predicted streamflow ensemble is wider than the ICEA ensemble at several lead times (day 2, day 6, and day 7), while the ensemble is narrower at other lead times (figure 9f)". Figure 9f shows the NRR which is a measure of the accuracy of the ensemble spread, not a direct measure of the width of the ensemble. The UR95 is a more accurate measure for comparing which has a wider ensemble spread. This statement would also be more accurate if phrased "the whole EnKF predicted streamflow ensemble is less overconfident than the ICEA ensemble at several lead times (day 2, day 6, and day 7), while the ensemble is more overconfident at other lead times (figure 9f)".
We thank the referee for the thoughtful comment on this point. We accordingly updated the sentence according to the referee's suggestion.