We greatly appreciate referee K. Engeland for your thoughtful and positive comments on this manuscript. Below are our detailed responses to the points raised in bold.

K. Engeland (Referee)

General comments

The paper is interesting and deserves publication after a moderate revision. The scientific content and the modelling experiment carried out is excellent. I think, however, that the presentation and discussion can be improved in several ways.

Response: Thank you for the positive comment.

Structure of paper

I think there are at least two ways to improve the structure of the paper 1. It could be helpful if you in the introduction provide some explicit aims, objectives, hypotheses or research questions you want two answer, and provide the answers to those in the conclusions. 2. You discuss your results to a large degree in the results-chapter as well as in this discussion and conclusion chapter. It might be better to make a Results and discussion-chapter and make a much shorter conclusion chapter. Now follows comments to each chapter of the paper.

Response to comment 1 on the structure of the paper: We agree with this point. It would be good to add more explicit motivation and specific questions examined in the paper to the introduction. Corresponding answers would be added to the conclusions as you note.

Response to comment 2 on the structure of the paper: We agree with this point as well. It is possible to include a vast majority of the discussion in a new Results and Discussion section. We'd likely then rename the final section to "Summary" and include only key discussion points for reference with concluding statements and answers to the study questions.

1 Introduction

In the introduction a couple of references could be added: (Griessinger et al., 2016) and (Bergeron et al., 2016) are very fresh paper in this journal and could be included. Some background material from Scandinavian assimilation experiments could be added, see (Udnæs et al., 2007) and (Engeset et al., 2003) for Norway and (Arheimer et al., 2011) for Sweden. For the background information, I think it could be interesting to add a few sentences how data assimilation is used operationally in western US. I guess there are several reports (grey literature) that covers this topic, and that in many cases subjective methods are used. On page 6 lines 148-49 your write a bit about manual practice, this could be moved to the introduction as a background information.

Response: Thank you for the references, they will be great to add to the introduction. As mentioned in lines 148-149, snow data assimilation is implemented manually in operation currently. We plan to move this to the introduction with a few more sentences describing the state of operational DA in the Western US. This helps more clearly defines the motivation of the paper.

2 Models and calibration

I would like to have some more details on the snow model: (1) Do you divide the catchment into elevation zones? This is standard for operational forecasting models in Scandinavia and is important for the performance in catchments with seasonal snow cover. (2) Do you have any sub-catchment distribution of snow (uniform, gamma, lognormal) or do you consider the snow depth to be equal all over the catchment? In Table 1 you list the mean elevation (please specify) but it would also be interesting to show the min and max elevation.

Response: We did not divide the catchment into elevation zones currently. We agree that elevation bands are standard practice in many regions, including the Western US. For this study, the reference models (no DA simulations) are also lumped, thus we feel the DA work and improvements are still relevant.

We are working toward elevation band simulations with DA across many basins currently, but it is not included in this manuscript.

The snow model assumes uniform depth across the basin, but does have an empirical snow covered area curve (see Snow-17 references in this paper).

We will add the min and max elevation in Table 1.

3.2 Generating ensembles of estimated observed watershed SWE

I think the use of the term "observation" is confusing since it might refer both to the point observations and the estimated catchment SWE from the regression equations. Especially "estimated observed watershed SWE" is confusing. Maybe it arise from the Ensemble Kalman filtering setting where the term "observation" is standard terminology. In the text it is not always evident when "observation" refers to the point measurement and when "observation" refers to the observation based catchment SWE. E.g. in line 111 "observation" refers to point observation, whereas for lines 112 and 113 I am confused if you refer to "observation based catchment SWE”"or the point measurements. Some suggestion are to write "observation based SWE", "observation based catchment SWE" or "observed catchment SWE". At least you should use a consistent terminology in order to distinguish between the point measurements and the estimated catchment SWE from the point measurements.

Response: We agree with this point and will clarify and change the terminology upon revision.

3.2.1 Percentile-regression

- In lines 121-122 you write: “within a sample of all SWE observations at the same site within a time-window of +/- n days centered on the date of the observation.” For me it is not evident if you then use all SWE observations from the year y, from the years preceding y or from all years in your dataset. Maybe the term “date” means “month and day” in this context and not “year, month, day”. Please specify.

Response: We mean all the years in our dataset, and will revise the sentences in our manuscript accordingly.

- Why do you do the regression on the percentile? Does the percentile give you different information than the observed SWE? Please explain why with some sentences.

Response: We did this mainly for reducing interpolation uncertainty caused by spatial heterogeneity of SWE gauge sites following Slater and Clark (2006). We will include clarification in the revised draft.

- Did you have any challenges since p has a lower and upper bound? On line 125, did you need to truncate the simulated p-values to be between 0 and 100?

Response: Yes, we needed to truncate the simulated p-values to between 0 and 100. We experimented with various assumptions related to this truncation and found that straight truncation (e.g. a regression percentile of 110 is set to 100) worked the best in these cases.

- The LOO cross validation approach is similar to the Jackknife approach. What is the difference since you do not call it Jackknife.

Response: We call it LOO cross validation approach just for specifying that it is a special Jackknife approach that only one sample is cut out each time.

- Lines 127-129 could be explained better. Do you calculate the percentile p for each ensemble member in order to get 100 pairs of p and SWE from the model? Both the observation based and the model based ensembles are random, it is not evident how the transformation works. Do you order the samples of p-values?

Response: Yes, we calculated the percentile p for each ensemble member in order to get 100 pairs of p and SWE from the model.

We will include the following discussion to improve clarity. We did not order the samples of p-values. We just calculate the corresponding percentiles of the full ensemble model SWE (i.e., a sample of (2n+1)*Y*100 members, where (2n+1) is the length of time window each year, Y is the total number of years of our dataset, 100 is the number of ensemble model SWE members) according to the estimated sample p-values $\hat{p}_y^o(j)$.

- Line 129. Why is capital J used?

Response: Initially we used capital J just for distinguishing it from the index of percentile ($\hat{p}_y^o(j)$), but the capital J does seem unnecessary. We will revise to use just lowercase *j* in our manuscript.

3.2.2 Z-score regression

- Why do you use the term z-score. It might be a bit confusing since the term"score" is often used for model evaluation

Response: It is just a conventional name referring to the transformation of Eq. (1). When using a Gaussian distribution, distance from the mean is often discussed in terms of standard deviations, and when normalized by the standard deviation of a particular distribution, a deviation is termed "Z-score" for Z-standard deviations from the mean in statistics.

- Lines 140-141: "long-term non-zero mean and standard deviation of the full ensemble model SWE within the time-window of +/- n days". Does this mean that you calculate the mean and standard deviation over a sample of 2*n*100 model simulations?

Response: We calculated the mean and standard deviation over a sample no greater than (2n+1)*Y*100 (only non-zero members are used), where (2n+1) is the length of time window of +/- n days in each year, Y is the total number of years of our datasets, and 100 is the number of ensemble model SWE members.

3.3 EnKF approach and experimental design

For the data assimilation, it could be useful to (i) write eq. 5 also without the h operator that is actually not used. (ii) describe in two sentences how the analysis works.

Response: We prefer to keep the transformation vector h as it is the formal terminology and if the transformation is not done in a pre-processing step as we have done it does need to be performed.

We calculate an analysis via eq. 5 and use that analysis to update the Snow-17 SWE states. We then run the model system with the updated states until the end of the WY. This clarification would be included as revised text at the end of section 3.3.

3.6 Verification metrics

It could be useful to write for which variables the verification metrics is calculated.

Response: The verification metrics are for seasonal streamflow volume. Text will be modified to state this.

4 Results

It is not necessary to put tables 2 and 3 in the paper, move them to supplementary material. Figures 3 and 4 are sufficient.

Response: We will move them to the supplementary material upon revision.

From the text and the figures it is confusing for which variable the evaluation statistics is calculated: On line 216 it is written: "The evaluation statistics for ensemble SWE observations". Whereas in the Figure captions it is written that the evaluation is for ensemble mean streamflow. It is not evident for which period the verification metrics is calculated.

Response: Our evaluation statistics are all calculated for streamflow. We can clarify the text to clearly state that the evaluation metrics are for seasonal streamflow volume, applied to the two SWE interpolation approaches using varying window lengths for the SWE transformation.

In Figure 3 and 4 it is not evident which forcing you use. Is it "perfect forecast" or one of the two ESP forecasts? What is the difference between the evaluations in Figures 3 and 4 versus Figures 5 and 6. Both pairs of figures show evaluation statistics for streamflow forecasts, but I am not able, based on the text, to tell the difference between the two set of plots.

Response: In figures 3 and 4, perfect forcing is used. This will be clarified in the text and figure captions.

The difference between the evaluations in Figures 3 and 4 versus Figures 5 and 6 is that their focuses are different. Figure 3 and 4 show the evaluation results of the sensitivity analysis of model and observation error variance (i.e., P, 0.5P and 2P).

Figure 5 and 6 show the evaluation results of seasonal ESP (two types of ESP forecasts) compared with perfect forcing. Further clarification will be added to the text.

There are two results and comments that seem to be contradicting: Line 229: Comment to Figures 2 and 3: "although the DA does not help correct forecast biases." Line 243-245: Comments to Figures 7,8 and 9: "Increasing the ensemble model SWE through DA will lead to increased model runoff, and vice versa. For basins with a strong seasonal cycle of streamflow (e.g. Greys and Merced River), SWE DA generally improves daily runoff forecasts in addition to seasonal volume forecast improvements" How is it possible that DA does not help correct forecast biases whereas it improves seasonal volume forecasts?

Response: We believe this comes through modification of both negative and positive runoff errors. Bias is a sum of signed errors, thus the noDA and DA runs can have similar total error even if the no DA run has large year to year errors. The DA run can improve a statistic like RMSE which is a squared error metric without changing bias. For example Figure 12, lower left panels highlights that DA reduces large positive error for a few years and conversely, increases negative runoff error over many years. This improves correlation, RMSE, etc, but leave bias nearly unchanged.

5 Discussion and conclusion:

- In general, it is helpful if you refer to specific tables and figures in the discussion to make it evident which results you discuss.

Response: We will revise the paper to include more specifics figure and table references in the summary section.

- Lines 264-273 could be moved to section 3.2 since it is a good description of the method used and not a discussion of the results presented in this paper.

Response: Agreed, this will be moved.

- I would like more discussion of Figures 10-12, and I would like to know how often the DA improves the simulated seasonal runoff and how often it becomes worse. Figures 7-9 could also include on year when DA makes the simulated seasonal runoff worse. For the subplots to the right in Figures 10-12, it could be interesting to know more about the cases when the points are located in the lower left or upper right quadrants, i.e. to little/much runoff is simulated and you decrease/increase the simulated runoff.

Response: Reviewer 2 had a similar comment to this. Most of this response is repeated there as well. Generally, when the SWE increment is incorrect, it is less than 10% of that year's SWE and runoff with the exception of the Merced River where five of the years have SWE increment errors larger than 10% of that year's runoff. In the Greys River, all incorrect increments are less than 10% of the observed runoff for that year and also in years where the noDA runoff error is less than 10% of observed. A small increment implies that the estimated observed and model SWE are very similar, and thus in years with small model error, the model SWE climatology closely matches observed climatology after transformation for this basin. Since SWE-Runoff are not perfectly correlated and there is likely information loss in the EnKF and modeling systems, it would be expected that in years where there is weak signal in the observations, the increment may end up being incorrect. Overall, there are 11 of 28 (39%), 4 of 24 (17%), and 8 of 26 (31%) years for the Greys, Tolt and Merced rivers where the DA increment is in the incorrect direction. However, the years with large SWE increments are always of the same sign as the runoff error except for the Merced River.

The Merced River is the only basin to use state of California SWE observations, and these may be of lower quality as evidenced by the large amount of manual quality control we had to perform on the data and the quality control discussion of these data in Lundquist et al. (2015). This suggests that observed SWE data need to be of higher quality (or information content) than the calibrated model SWE to have a positive impact in the DA system. Conversely, there are years where the noDA runoff error is large, but the SWE increment is small in all three basins. This is not unexpected as spring SWE is not perfectly correlated with subsequent runoff. This may also hint at a level of data loss in the EnKF and modeling system, future work should compare streamflow hindcasts using this type of system with traditional statistical methods using SWE as a primary input.

We will look a bit more into these years and try to identify if there was anything that makes them "special."

Reference:

Lundquist, J. D., M. Hughes, B. Henn, E. D. Gutmann, B. Livneh, J. Dozier, and P. Neiman, 2015: High-elevation

precipitation patterns: using snow measurements to assess daily gridded datasets across the Sierra Nevada, California. *J. Hydrometeorology*, 16, 1773-1792. doi: 10.1175/JHM-D-15-0019.1.

Details:

(i) Why is ESP an abbrevation for "Ensemble Streamflow Forecast"? (ii) What is X in 1981-201X on line 174?

Response: We will fix the text, ESP should be an abbreviation for “Ensemble Streamflow Prediction”.

Suggested references:

Arheimer, B., Lindström, G., Olsson, J., 2011. A systematic review of sensitivities in the Swedish flood-forecasting system. Atmos. Res. 100, 275–284. doi:10.1016/j.atmosres.2010.09.013

Bergeron, J.M., Trudel, M., Leconte, R., 2016. Combined assimilation of streamflow and snow water equivalent for mid-term ensemble streamflow forecasts in snowdominated regions. Hydrol. Earth Syst. Sci. Discuss. 1–34. doi:10.5194/hess-2016-166

Engeset, R.V., Udnæs, H.C., Guneriussen, T., Koren, H., Malnes, E., Solberg, R., Alfnes, E., 2003. Improving runoff simulations using satellite-observed time-series of snow covered area. Nord. Hydrol. 34, 281–294.

Griessinger, N., Seibert, J., Magnusson, J., Jonas, T., 2016. Assessing the benefit of snow data assimilation for runoff modelling in alpine catchments. Hydrol. Earth Syst. Sci. Discuss. 1–18. doi:10.5194/hess-2016-37

Udnæs, H.-C., Alfnes, E., Andreassen, L.M., 2007. Improving runoff modelling using satellite-derived snow covered area? Nord. Hydrol. 38, 21. doi:10.2166/nh.2007.032

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-185, 2016.

We greatly appreciate this anonymous referee for your thoughtful and positive comments on this manuscript. We have revised the manuscript accordingly. Below are detailed responses to the points raised.

Anonymous Referee #2

I found the topic relevant to HESS and a contribution to DA understanding for water resources in snow-dominated watersheds. While I found the paper well written, it was often difficult to follow because of the number of DA-model scenarios and corresponding acronyms (though I struggled to come up with good alternatives). I also thought the results section lacked specifics and overly asked the reader to interpret the figures/tables. Finally, I found the major contribution of the paper to be its potential utility for improving streamflow prediction in watersheds with relatively low model skill. I would like to see the authors leverage their previous work to highlight the utility of the approach presented. It should be noted that I reviewed 'version 2' of the manuscript.

Response: Thank you for your overall summary comments of the paper. We agree with the general comment that additional specific analysis can be added to the results section. We will clarify the text throughout, with emphasis on the results section. Our replies to your specific comments give more detail to this general response.

Comment 1: Include more detail in the results. The reader is left to do most of the work in interpreting and quantifying many statements. Tell us how much and where things were improved and where they were not. Statements like this on line 211: "However, we also note that the ensemble observations of 7-day window can have a large variance, likely due to the more limited sample size for the regression, which can negatively impact DA performance (see Supplement Tables S1.1 and S1.2)." would strongly benefit from specific number. What is large variance? What is a negative impact to DA?

Response: The negative impact is truly a reduction in the positive impact of DA when comparing the 7-day window to the 3-month window. We will include specific numbers and clarify this text.

I became frustrated having to look at all the figures and table to understand what was meant by sentences like this. A number of examples are listed below, but I encourage the authors to re-read the manuscript to address this problem completely. Lines 221-225: Where by how much?

Response: We agree that the results section needs more analysis clearly stated in the text. We will tabulate key results for the metrics across example metrics for the entire basin set and add those results to the text.

Line 227-228: Which basins? By how much?

Response: Again, we will revise this discussion

Lines 243-246: Improves runoff forecasts by how much

Response: We will quantify the improvement to daily flow for the example basins given.

Comment 2: Can you remove some of the acronyms or more clearly explain every acronym in the figure captions.

Response: Yes, we agree these need to be more clearly defined for each figure, or removed entirely in the captions. We will do that upon revision.

Comment 3: There should be more discussion of why the DA could make predictions worse and where that occurred. Should we be worried about this for future DA efforts? How might we screen sites to ensure that DA does not make predictions worse?

Response: We can see from the right two subplots in Figures 10-12 that the years when DA makes the simulated runoff worse is when runoff error is generally very small. Generally, those SWE increments are less than 10% of that year's SWE and runoff with the exception of the Merced River where five of the years have SWE increment errors larger than 10% of that year's runoff. Overall, there are 11 of 28 (39%), 4 of 24 (17%), and 8 of 26 (31%) years for the Greys, Tolt and Merced rivers where the DA increment is in the incorrect direction.

In terms of observational sites, the Merced River is the only basin to use state of California SWE observations, and these may be of lower quality as evidenced by the large amount of manual quality control we had to perform on the data and the quality control discussion of these data in Lundquist et al. (2015). This suggests that observed SWE data need to be of higher quality (or information content) than the calibrated model SWE to have a positive impact in the DA system. Conversely, there are years where the noDA runoff error is large, but the SWE increment is small in all three basins. This is not unexpected as spring SWE is not perfectly correlated with subsequent runoff. This may also hint at a level of data loss in the EnKF and modeling system, future work should compare streamflow hindcasts using this type of system with traditional statistical methods using SWE as a primary input.

We believe screening of observational sites is a difficult task. The above discussion and our results in California does suggest screening is needed. High quality sites with no information content would also need to be screened as well (also see discussion in reply to comment 4). It is possible that guidelines for this could be developed and then potentially automated, but this is likely a major undertaking. In this study, site selection was first taken using closest distances to the basin, then manual screening of suspect sites and sites that had little relationship with runoff were removed. It is possible some formalization of this methodology could be developed.

That being said, the relationship between SWE and runoff will likely be basin dependent and the addition of an assimilation system and model forecast introduces information losses that are also likely basin dependent since the hydrologic modeling system is basin dependent, such that a screening methodology based solely on observations is likely to misidentify potential degradation or improvement when DA is applied.

Comment 4: It seems that one of the major contributions of the paper is pointing out that DA methods are likely only make improvements in snow dominated watersheds when model performance was <0.80 NSE. Given that Newman et al., 2015a has quantified the performance of SAC-SMA skill in >500 watersheds, I think a major contribution would be to discuss how many watersheds could benefit from DA and how they are spatially distributed. I think that this should be discussed in the context of where the DA methods did not perform well, i.e. comment 2.

Response: This idea you mention is an interesting topic. We will look back through the database and add some additional analysis examining spatial location of basins that may benefit from DA using the basic metrics of noDA NSE and contribution of SWE to runoff. That being said, a comprehensive description and analysis about how many watersheds could benefit from DA and how they are spatially distributed is a large topic and could be a separate paper.

Preliminary screening of candidate basins would not only require the basic metrics of being snow dominated, generally lower noDA skill, but also somehow assessing the quality of information from the nearby observation sites. Furthermore, we'd expect that implementation of the enKF DA would result in potential differences as there may be data loss in the observation transformation operator, etc.

Minor comments: 1. It seems odd to combine the discussion and conclusions section.

Response: We will revise the last two sections to be Results and Discussion and Summary. More discussion will be included in section 4, while the summary section will restate key discussion points and then findings of the study.

References:

Lundquist, J. D., M. Hughes, B. Henn, E. D. Gutmann, B. Livneh, J. Dozier, and P. Neiman, 2015: High-elevation precipitation patterns: using snow measurements to assess daily gridded datasets across the Sierra Nevada, California. *J. Hydrometeorology*, 16, 1773-1792. doi: 10.1175/JHM-D-15-0019.1.

Evaluation of snow data assimilation using the Ensemble Kalman Filter for seasonal streamflow prediction in the Western United States

Chengcheng Huang[1, 2], Andrew J. Newman[2], Martyn P. Clark[2], Andrew W. Wood[2] and Xiaogu Zheng[1]

[1] College of Global Change and Earth System Science, Beijing Normal University, Beijing, China

[2] National Center for Atmospheric Research, Boulder CO, 80301, USA

Correspondence to: Andrew J. Newman (anewman@ucar.edu)

Abstract. In this study we examine the potential of snow water equivalent data assimilation (DA) using the ensemble Kalman Filter (EnKF) to improve seasonal streamflow predictions. There are several goals of this study. First, we aim to examine some empirical aspects of the EnKF, namely the observational uncertainty estimates and the observation transformation operator. Second, we use a newly created ensemble forcing dataset to develop ~~our~~ ensemble model states ~~(e.g.~~that provide an estimate of model state uncertainty~~)~~. ~~Finally, we also~~Third, we examine the impact of varying the observation and model state uncertainty on forecast skill. We use basins from the Pacific Northwest, Rocky Mountains, and California in the western United States with the coupled Snow17 and Sacramento Soil Moisture Accounting (SAC-SMA) models. ~~Results show that~~We find that most EnKF implementation variations result in improved streamflow prediction, but the methodological choices in the examined components impact predictive performance in a non-uniform way across the basins. Finally, basins with relatively higher calibrated model performance (> 0.80 NSE) without DA generally have lesser improvement with DA, while basins with poorer historical model performance show greater improvements.

Keywords:

Hydrological data assimilation; SWE; EnKF; Snow-17; SAC

1 Introduction

In the snow-dominated watersheds of the Western US, spring snowmelt is a major source of runoff (Barnett et al., 2005; Clark and Hay, 2004; Singh and Kumar, 1997; Slater and Clark, 2006). In such basins, the initial conditions of the basin, primarily in the form of snow water equivalent (SWE), drive predictability out to seasonal time scales (Wood et al., 2005; Wood and Lettenmaier, 2008; Harrison and Bales, 2015; Wood et al. 2015). Thus better estimates of basin mean initial SWE should lead to better seasonal streamflow predictions (Arheimer et al., 2011; Clark and Hay, 2004; Slater and Clark, 2006; Wood et al. 2015). For various reasons (e.g., the uncertainty in model parameters, forcing data, model structures), simulated SWE in hydrological models can be very different from reality (Pan et al., 2003). Fortunately, a variety of snow observations (including point gauge and spatial satellite data) contain valuable information (Andreadis and Lettenmaier, 2006; Barrett, 2003; Engeset et al., 2003; Mitchell et al., 2004; Su et al., 2010; Sun et al., 2004).

Many studies have explored the role of snow data assimilation in different modeling frameworks (Kerr et al., 2001; Moradkhani, 2008; Takala et al., 2011; McGuire et al, 2006; Wood and Lettenmaier, 2006). Of particular focus here are papers that have examined the impact of SWE data assimilation (DA) on runoff modelling and prediction (e.g. Bergeron et al., 2016; Griessinger et al., 2016; Wood and Lettenmaier, 2006; Franz et al., 2014; Jörg-Hess et al., 2015; Moradkhani, 2008; Slater and Clark, 2006). Among the major challenges facing SWE-based DA are that the time-space resolution of remote sensing SWE data are too coarse or period-limited for many watershed-scale hydrological applications in mountainous regions (Dietz et al., 2012; Jörg-Hess et al., 2015), and point gauge snow data have sparse and uneven spatial coverage. For point measurements, spatial interpolation based on distance are typically used to estimate observed SWE state in a watershed of interest (e.g.,Franz et al., 2014; Jörg-Hess et al., 2015; Slater and Clark, 2006; Wood and Lettenmaier, 2006).

Here we use the Ensemble Kalman Filter (EnKF) method for DA using an implementation that allowing for seasonally varying estimates of observation and model error variances (Evensen, 1994, 2003; Evensen et al., 2007). The EnKF framework has been successfully implemented in research basins in several previous studies (Clark et al., 2008; Franz et al., 2014; Moradkhani et al., 2005; Slater and Clark, 2006; Vrugt et al., 2006). The EnKF provides an objective analytical framework to optimize the update of model states based on observed values and their corresponding uncertainties. While the EnKF approach has a formal theory, its overall objectivity in an application (contrasting with an arbitrary DA approach such as direct insertion) nonetheless depends on several methodological choices that are often empirical when applied to SWE DA. ~~Here we examine DA performance sensitivities related to three elements: 1) the estimation of watershed mean SWE from surrounding point measurements; 2) the transformation operator that relates watershed mean SWE to model mean SWE; and 3) sensitive analyses of the relative size of observed and model error variance.~~

Following Slater and Clark (2006), this study uses two slightly different approaches to estimate ensemble SWE observations with point gauge SWE data from surrounding gauge sites for study basins. When using calibrated hydrologic modeling systems, model SWE states may exhibit systematic biases from observed SWE estimates for a number of reasons – e.g., all hydrologic models must simplify real watershed physics and structure, and model parameter estimation (calibration) may result in SWE behavior that in part compensates for forcing or model errors (e.g. Slater and Clark, 2006). Therefore, transformation of snow observations to model space is needed before they are used to update the model states to ensure that the model ingests SWE estimates that are as close to unbiased relative to the model climatology as possible. We explore two variations on an approach using cumulative density function (CDF) transformations of observations to model space (following Wood and Lettenmaier, 2006, among others). Additionally, we undertake a sensitivity analysis to highlight the importance of robust observations and model uncertainty estimates. We focus on the impacts of updates made just once per snow accumulation season, noting that an important choice that is not examined as a result is the selection of DA dates and frequency. For a given generally optimal selection of the ~~EnKF system~~EnKF approach, the Ensemble Streamflow Prediction (ESP) approach is used to test the impact of SWE DA on subsequent streamflow forecasts.

For context, ~~essentially all~~ operational seasonal streamflow forecasts in the US ~~use no~~currently do not use formalized DA. ~~Typically~~ If the initial states of the model are ~~assumed correct~~suspected to contain error (He et al. 201~~2~~2~~1~~)~~. If any,~~ DA is performed it is through subjective forecaster intervention~~; ~~. M~~m~~anual adjustments (termed 'MODs', e.g. Anderson 2002) to model states (e.g. SWE) are applied repeatedly throughout the water year, and particularly before initializing seasonal forecasts. This manual nature of the correction hinders the ability to scale up DA procedures to many basins, to benchmark ~~model~~DA performance, and ~~make~~quantify improvements to the forecast system as skill depends on forecaster experience (Seo et al. 2003).

~~Thus, t~~The ~~main~~central motivating aim of this study is thus to ~~examine~~ assess the potential benefits of objective, automated SWE DA ~~using an EnKF system~~ against a reference model configuration to identify forecast improvement opportunities. We ~~do this via application of~~apply ~~We apply~~ the EnKF DA approach ~~system~~ to nine river basins in the Western US that have a range of basin features and environmental conditions, over a period of multiple decades. This experimental scope ~~is~~ ~~distinct~~differs from many previous studies that focus on one or two basins ~~in more detail~~ (e.g., Clark et al., 2008; Franz et al., 2014; He et al., 201~~1~~2; Moradkhani et al., 2005), or assess DA performance over shorter periods. We also use ensemble simulations driven by a new probabilistic forcing dataset (Newman et al, 2015) as a basis for estimating model SWE uncertainty, in contrast to prior studies that ~~which used~~relied on more arbitrary distributional assumptions. This ~~experimental design~~ range of basins permits us to explore the question of: "In what types of basins might automated SWE DA improve seasonal streamflow forecasts?"

Additionally, as discussed throughout the introduction, the ~~EnKF system~~EnKF approach has several empirical components that require tuning. ~~T~~We therefore~~,~~ e~~Here we e~~xamin~~ation of~~e EnKF ~~DA~~ performance sensitivities related to three elements: 1) the estimation of watershed mean SWE from surrounding point measurements; 2) the transformation operator that relates watershed mean SWE to model mean SWE; and 3) sensitivity~~e~~ analyses of the relative size of observed and model error variance~~ is also undertaken~~.

The following sections discuss the study basins and data sets, and the model and EnKF DA approach, before the presenting study results and discussion, and ~~discussion and conclusions~~a summary.

2 Study basins and data

In this study, nine basins across the Western US are selected for SWE DA evaluation. They are in the Pacific Northwest, California (Sierra Nevada Mountains), and central Rocky Mountains. We focus on these three areas as they span a range of snow accumulation and melt conditions of the Western US and are in areas with active seasonal streamflow prediction and water resource management. ~~Note w~~We do not examine rain driven low-lying basins ~~as~~ because they do not have significant SWE contributions to runoff. The locations of the basins and nearby SWE gauge sites are shown in Figure 1, illustrating that all of the study watersheds have SWE measurements distributed in and/or around the basins. The main features of these basins are shown in Table 1. The basin areas range from 16 to 1163 km^2 and the mean elevations of the basins range from 998 to 3459 m with a large spread in basin mean slopes (as estimated from a fine-resolution digital elevation model) and forest percentage. Two sources of SWE observations are used in this study: (1) the widely used Snow Telemetry (~~Snotel~~SNOTEL) network for Natural Resources Conservation Service (NRCS)~~ (www.wcc.nrcs.usda.gov/snow/)~~, which covers most of the western US; and (2) the California Department of Water resources (DWR, ~~) (cdec.water.ca.gov/snow) (~~denoted as CADWR sites hereafter), which maintains a snow pillow network for California. The SWE data from CADWR sites have frequent missing data and some unrealistic extreme values, thus extensive manual quality control was required before using the CADWR data in the study.

3 Methodology

3.1 Models and calibration

The Snow-17 temperature index snow model is coupled to the Sacramento Soil Moisture Accounting (SAC-SMA) conceptual hydrologic model (Anderson, 2002; Anderson, 1973; Burnash and Singh, 1995; Burnash et al., 1973; Franz et al., 2014; Newman et al., 2015a) to simulate streamflow in this study. This model combination has been in operational use by US National Weather Service (NWS) River Forecast Centers (RFCs) since the 1970s (Anderson, 1972; 1973). The Snow-17 model is a conceptual snow pack model that employs an air temperature index to partition precipitation into rain and snow and parameterize energy exchange and snowpack evolution processes. The only required forcing inputs are near-surface air temperature and precipitation. The output rain-plus-snowmelt (RAIM) time series from Snow-17 is part of the forcing input of the SAC-SMA model. SAC-SMA is a conceptual hydrologic model that uses five moisture zones to describe the movement of water through watersheds. The required forcing input is the potential evaporation and the surface water input from Snow-17.

Daily streamflow data from United States Geological Survey (USGS) National Water Information System server (http://waterdata.usgs.gov/usa/nwis/sw) are used to calibrate 20 parameters of Snow-17 and SAC-SMA model. The calibration is obtained using the shuffled complex evolution global search algorithm (SCE; Duan et al, 1992) via minimizing daily simulation Root Mean Square Error (RMSE). ~~UGSS~~ USGS streamflow data are also used to verify the model predictions.

Model uncertainty arises from model parameter and structural uncertainty (e.g. Clark et al., 2008) and forcing input uncertainty (e.g., Carpenter and Georgakakos, 2004). Focusing on the latter, we drive the hydrology models with 100 equally likely members of meteorological data ensemble generated as described in Newman et al. (2015b), producing an 100 member ensemble of model moisture states, including SWE, and streamflow. The daily-varying spread of the ensemble model states serve as the estimate of model uncertainty. Because this method estimates SWE uncertainty without also considering sources other than forcing input uncertainty, and therefore may underestimate model uncertainty in initial SWE (e.g. Franz et al. 2014), we also include a sensitivity analysis to explore the sensitivity of DA results to variations in the estimated observation and model uncertainty magnitudes.

3.2 Generating ensembles of estimated observed watershed SWE

Since the SWE gauge observations are point measurements that do not represent the watershed mean conditions and have observation error, observation uncertainty needs to be robustly estimated to ensure reasonable DA performance. In this study, we follow Slater and Clark (2006) to generate ensemble estimated catchment SWE from gauge observations using a multiple linear regression in which the predictors are the attributes of SWE gauge sites (longitude, latitude and elevation). The observation uncertainty is estimated by leave-one-out (LOO) cross validation: i.e., each station is left out of the regression training and then its SWE is predicted and verified against its actual measurement. For reducing interpolation uncertainty caused by spatial heterogeneity of SWE gauge sites, t~~T~~he SWE values are transformed into percentiles or Z-scores (eg, standard normal deviates) before the regression is performed, and the corresponding inverse transformations are used to convert them back to SWE values. These two approaches are denoted as percentile and Z-score interpolation respectively and detailed descriptions for them are as follows.

3.2.1 Percentile interpolation

First, the non-exceedance percentile $p_y^o(k)$ of each SWE observation (observation based values noted with superscript o) at gauge site k on DA date in year y is calculated based on its rank, or percentile, within a sample of all SWE observations in all years at the same site within a time-window of +/- n days centered on the date of the observation in each year.

Then we use the percentiles to do linear regression on geographic features latitude, longitude and elevation to estimate the SWE percentile for the target basin: $\hat{p}_y^o$, where the hat indicates the basin mean estimate. By LOO cross validation, the interpolation error of the linear regression is estimated as $\hat{e}_y^o$. We sample from normal distribution $N(\hat{p}_y^o, \hat{e}_y^o)$ to get the ensemble percentiles $\{\hat{p}_y^o(j)\}$, where j = 1,…, 100 represents ensemble member.

Finally, we take the corresponding $\hat{p}_y^o(j)$ percentile from the full ensemble model SWE within the time-window of +/- n days centered on the DA date each year in all years ~~in year *y*~~, denoted as $\hat{S}_y^{\mathrm{f}}(j)$. The final ensemble SWE observations on DA date at year y for the target basin are $\{\hat{S}_y^{\mathrm{f}}(j)\}$, where ~~*J*~~ j = 1,…, 100.

3.2.2 *Z*-score interpolation

First, we use the observed SWE at gauge site k on DA date in year y to calculate the Z-score:

$$Zscore_y(k) = \frac{S_y^o(k) - \overline{S^o(k)}}{\sigma(S^o(k))}, \tag{1}$$

where $\overline{S^o(k)}$ and $\sigma(S^o(k))$ are the ~~are the~~ long-term mean and standard deviation of a sample of all non-zero SWE observations at the same site within a time-window of +/- n days centered on the date of the observation respectively. Here we use the Z-score in the linear regression and again use LOO cross validation to estimate the mean and interpolation error of the Z-score for a target basin. Then we sample from normal distribution to get ensemble Z-scores for target basin, denoted as $\{\hat{Z}\text{-score}_y^o(j)\}$, where j = 1,…, 100 represents ensemble member. Finally we use the following equation to transform Z-score to back to SWE values:

$$\hat{S}_y^o(j) = \hat{Z}score_y^o \times \sigma\left(S^f(k)\right) + \overline{S^f(k)}, \tag{2}$$

where $\overline{S^f(k)}$ and $\sigma\left(S^f(k)\right)$ are the long-term non-zero mean and standard deviation of the full ensemble model SWE within the time-window of +/- n days centered on the DA date each year in all years~~in year *y*~~ respectively. The final ensemble SWE observations on DA date at year y for the target basin are $\{\hat{S}_y^o(j)\}$, where j = 1,…, 100.

Both percentile and Z-score transformations normalize the original SWE values ~~in a way~~ to decrease their spatial variability (Slater and Clark 2006; Wood and Lettenmaier, 2006). The latter~~former~~ ensures the ensemble observations have the same mean as the ensemble model SWE and the variance of ensemble observations is proportional to ensemble model SWE variance. The former~~latter~~ emphasizes the shape of the observation time series. SWE observations in and near a watershed but at different elevations may have greatly varying values, but their percentile and Z-score statistics will show reduced variation because they arise from similar relative weather conditions with respect to conditions in other years. Using normalized statistics significantly reduces the interpolation uncertainty and systematic biases relative to the watershed's SWE climatology.

3.3 EnKF approach and experimental design

For evaluating the relative performance of DA and for re-initializing the soil moisture of DA runs at the beginning of each water year (WY), an open loop or 'control' retrospective simulation (denoted No DA) is performed using the calibrated model parameters with ensemble forcing data. This control run is one continuous simulation per ensemble member for the entire hindcasting and evaluation period (1981-201X) for each basin.

~~In operational practice, manual adjustments to model SWE are applied repeatedly throughout the water year, and particularly before initializing seasonal forecasts.~~ Because this study ~~is~~ focuses~~ing~~ on assessing variations in methodological aspects of the DA approach rather than differences in performance throughout a forecasting season, we ~~simplify this configuration by~~ apply~~ing~~ DA updates only once per year, using the date on which the SWE correlation with future runoff is highest for the study basin, but no later than 1 April, a common date for initiation of spring seasonal runoff forecasts.

The EnKF method used in this study is a time-discrete forecast and linear observation system described by two relationships (generally following the notation of Ide et al. (1997) and Wu et al. (2012)) :

$$\boldsymbol{x}_{i+1}^{t} = M(\boldsymbol{x}_{i}^{t}) + \boldsymbol{\eta}_{i}, \tag{3}$$

$$\boldsymbol{y}_{i}^{o} = \boldsymbol{h}(\boldsymbol{x}_{i}^{t}) + \boldsymbol{\varepsilon}_{i}, \tag{4}$$

where i is the time step, M is the coupled Snow17 and SAC-SMA model~~ system~~, $\boldsymbol{x}$ is the state variable and $\boldsymbol{y}$ is the observation variable (in this study both $\boldsymbol{x}$ and $\boldsymbol{y}$ are the one-dimensional vector containing basin mean SWE for the target watershed across all ensemble members), the superscripts t and o stand for truth and observed respectively, $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are the model and observation errors respectively, and $\mathbf{h}$ is the observation operator that maps the model states to the observation variable. In this study, $\mathbf{h}$ is simply the identity vector as we regard the SWE estimates that have been transformed to model space as observation $\boldsymbol{y}$, as a pre-processing step.

The SWE DA approach is implemented via the following procedure:

1) Run the watershed model once for each ensemble forcing member from the beginning of a WY until the DA date with initial states $\boldsymbol{x}_0$ taken from the retrospective control runs, producing the ensemble forecast states $\boldsymbol{x}_i^f$. The superscript f denotes forecast.

2) Calculate the ensemble analysis states:

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^f + s_i \boldsymbol{h}_i^T (\boldsymbol{h}_i s_i \boldsymbol{h}_i^T + o_i)^{-1} \boldsymbol{d}_i, \tag{5}$$

where superscript a means analysis, $\mathbf{o}$ and $\mathbf{s}$ are the observed and model simulation error variances (estimated by the variance of ensemble observations and model states respectively) respectively, and the innovation vector (residual) is calculated as:

$$\boldsymbol{d}_i = \boldsymbol{y}_i^o - \boldsymbol{h}_i(\boldsymbol{x}_i^f), \tag{6}$$

3) Update the Snow-17 SWE states with the analysis states to use for initialization of forecasts through the end of the WY.

Steps 1-3 are repeated for all WY available in the hindcast period (1981-201X). Soil states are re-initialized using the states from the retrospective (No DA) run at the start of every WY (October 1), when there is no SWE. To summarize, we calculate an analysis via E~~e~~q. 5 and use that analysis to update the Snow-17 SWE states. We then run the model~~ system~~ with the updated states until the end of the WY.

3.4 Model and observation error variance

In this study, only the uncertainty of the forcing data is taken into account in our model uncertainty, and uncertainty that arises from model structural and parameter errors could cause the true model error to be larger. Thus we assess the impacts of inflating model error variance to evaluate the relative size of observed and forecast error variance. We simply set the model SWE error variance to 1/2 and 2 times of the original size to see how the DA performances change. If increasing the model error variance results in DA performance improvements, it would indicate that the model error variance is underestimated, and vice versa. This sensitivity analysis underscores the importance of a careful effort to properly estimate both model and observational uncertainty when using the EnKF – a challenge that is well known in the DA community.

3.5 Seasonal Ensemble Streamflow Prediction

Although the impacts of the SWE DA on forecast accuracy can be assessed through verification of post-adjustment simulations using 'perfect' future forcing, we demonstrate the performance of SWE DA by initializing seasonal ESP forecasts for a streamflow forecast product that is widely used in water management, the snowmelt-period runoff volume from April through July. ESP uses historical climate data to represent the future climate conditions each year from the start point of forecast period to predict streamflow. Two typical ESP applications are tested in this study. Because we have an ensemble of historical forcing instead of the traditional application in which only a single historical forcing time series is available, there are different ways to construct an ESP. We adopt two: (1) We construct the ESP forcing ensemble by randomly selecting one year of the historical ensemble forcing data for each historical member of the ESP; and (2) We use all historical years of ensemble mean forcing data for each ESP historical year member, yielding a 30*100 member ensemble for an ESP based on meteorology from 1981-2010 (variations are noted ens forcing and ens mean forcing respectively in subsequent figures discussing ESP results).

3.6 Verification metrics

In this study, five frequently used statistics are calculated for April through July seasonal streamflow volume expressed as runoff (mm) for evaluating the two DA approaches. The b~~B~~ias, c~~C~~orrelation coefficient (R), r~~R~~elative root mean squared error (R-RMSE), Nash-Sutcliffe efficiency (NSE) are based on the ensemble averages. ~~And~~ The c~~C~~ontinuous r~~R~~anked p~~P~~robability s~~S~~core (CRPS) is a measurement of error for probabilistic prediction (Murphy and Winkler, 1987). It is defined as the integrated squared difference between cumulative distribution function (CDF) of forecasts and observations:

$$\mathrm{CRPS}=\int_{-\infty}^{+\infty}\left[F^{\mathrm{f}}(x)-F^{\mathrm{o}}(x)\right]^{2}dx, \tag{7}$$

where F^{f} and F^{o} are CDF for forecasts and observations of streamflow respectively. Smaller CRPS means more accurate forecasts.

4 Results and D~~d~~iscussion

4.1 Overall performance in the case basins

Using the two approaches described in Section 3.2 with three different window lengths (7 days, 3 months, 1 year), a sample comparison from one year (2004) of the results for estimated watershed SWE from the two methods versus the model SWE ensemble on DA date (DA dates for the case basins are listed in Table 1) for the case basins are shown in Figure 2. The distributions of SWE from the model ensemble and from the percentile and *Z*-score interpolation methods differ in ways that are not consistent across all watersheds. The variance of the estimated observed SWE for both methods is generally largest for the 1-year, an effect that is more pronounced for the *Z*-score interpolation. However, we also note that the ensemble observations of 7-day window can have a larger variance -than the 3-month window, and as large as the 1-year window in some cases. See the ~~(e.g.,~~ percentile interpolation for ~~Smith~~the Payette River for 7-d window in Figure 2 where the 7-day window interquartile range is about 250 mm, the 1-year window range is 300 mm while the 3-month window is only about 120 mm. This ~~),~~is likely due to the more limited sample size for the regression, which can ~~negatively impact~~reduce the positive impact of DA performance. For example, the ~~(e.g.,~~ SF Payette River and the Greys River have positive DA impact for both the 7-day and 3-month windows ~~in Table 3 of Supplement S1 and Crystal River in Table 4 of Supplement S1 where DA improvements is limited.see Supplement Tables S1.1 and S1.2)~~but for the 7-day window the positive impact is reduced by roughly half in both basins for most metrics (Tables S.1 and S.3 of Supplement S1). Increased estimated observation variance decreases the weight of the observations in an ~~EnKF system~~EnKF approach and thus decreases the impact of the observations. In this study, a 3-month window of SWE observations generally gives the best performance~~Therefore, a 3-month window is recommended for both approaches.~~. However, in some basins a different window length may bring larger improvements. ~~Generally, l~~Longer windows mean that ~~more SWE values are used for transformation, and~~ the transformation ~~tends to be~~is more statistically representative of the long-term model-observation climatology. ~~. However, s~~Shorter time windows ~~mean~~ imply that the model SWE values used for transformation are more relevant to a specific seasonal time period, avoiding aliasing for seasonality, but have much smaller sample sizes and may not properly represent the relationship between model and observation climatologies. The window length must be a balance between these two considerations. Therefore, a 3-month window is recommended for both approaches.

~~Both percentile and Z-score transformations normalize the original SWE values in a way to decrease the spatial variability (Slater and Clark 2006; Wood and Lettenmaier, 2006). The former ensures the ensemble observations have the same mean as the ensemble model SWE and the variance of ensemble observations is proportional to ensemble model SWE variance. The latter emphasizes the shape of the observation time series. SWE observations in and near a watershed but at different elevations may have greatly varying values, but their percentile and Z-score statistics will show reduced variation because they arise from similar relative weather conditions with respect to conditions in other years. Using normalized statistics significantly reduces the interpolation uncertainty and systematic biases relative to the watershed's SWE climatology..~~

The evaluation statistics for simulated streamflow using~~with~~ perfect forcing after DA with ensemble SWE observations estimated by the percentile and Z-score interpolation approaches for the 3-month ~~7-day and 1-year~~ windows are shown in Figures 3 and 4. They are also compiled in Tables S.1-6~~2 and Table 3 respectively. Only results for the 3-month window are shown in these two tables, the tables for 7-day and 1-year windows are~~ in supplement S1. In those~~ese~~ tables~~Tables 2 and 3~~, the 2nd column shows the forecast error variance used to calculate analysis states, where "No DA" means the open loop control run (see Section 3.3), and the P, 1/2·P and 2·P refer to the DA runs with the model error variance estimated by 1, 1/2 and 2 times the original size of the ensemble model variance. Both percentile and *Z*-score interpolation approaches exhibit enhanced DA performance~~s~~ among the case basins, indicating that both ~~these~~ approaches are effective in adding observation based information to the model simulations. Overall, using the original model variance estimate (case P) the mean improvement for the percentile interpolation method (Z-score method) is a reduction in relative RMSE (R-RMSE) of about 11% (12%) and an increase in NSE of 0.03 (0.05). The percentile interpolation and Z-score interpolation methods vary in performance across the basins with both performing better in some basins and not others (e.g. ~~comparing the results in Table 1 and Table2 in supplement S2,~~ percentile interpolation performs slightly better than Z-score interpolation in Grey River using NSE as the evaluation metric (0.94 vs 0.93) and slightly worse ~~than~~ that in SF Tolt River (0.82 vs 0.88)). Using NSE, percentile interpolation performs better in the Greys River, while Z-score interpolation performs better in the Vallecito, South Fork of the Tolt, Merced, and Smith Rivers. To the hundredth NSE value (0.01) both methods are equivalent in the South Fork of the Payette River, and General and Blackwood Creeks.

The results of forecast error variance inflation shows that for both percentile and Z-score interpolation, "2·P" has better performance than "P" in most of the case basins – i.e., increasing the model error variance leads the assimilation to trust observations more and improves the DA performance (circles in both figures generally have improved evaluation metrics than squares or triangles). Using NSE, the percentile (Z-score) interpolation "2·P" case is on average another 0.01 (0.01) better than the "P" case across the nine basins. Th~~ese~~is ~~indicates that the model error variance tends to be underestimated or our observation uncertainties tend to be overestimated.~~ sensitivity analysis~~es~~ of model uncertainty impacts on DA performance suggest that either the forcing-alone based estimation of model errors underestimate the total model error variance, or the observed SWE error estimation approaches (interpolation plus the SWE regression) tend to overestimate observation uncertainty, or both. It is likely we are underestimating model uncertainty because we have not taken model structural and parameter uncertainty into consideration~~.~~

~~The evaluation statistics of Table 12 and 2 3 in supplement S1 are also presented as scatter plots in Figures 3 and 4 respectively~~. B~~The metrics R-RMSE, R and NSE indicate that b~~oth approaches bring incremental enhancements to the ensemble mean streamflow hindcast in most basins when evaluated across the R-RMSE, R and NSE metrics, ~~although~~ however ~~the~~ DA does not help correct forecast biases in these simulations~~s~~. Post-processing procedures (e.g. bias correction) could be used to further enhance the ~~system~~ forecast performance, but is not a focus of this study. These figures also show that ~~nNo DA~~ forecasts without DA ("No DA" in figures, "NoDA" in text) ~~with~~ that have relatively better performance, mostly due to better simulations of forecast initial conditions, benefit less from DA. -Three of the basins have a No-DA seasonal runoff NSE of less than 0.8, with an average improvement of 0.05 for the percentile regression and 0.12 for the Z-score regression versus 0.03 and 0.05 across all nine basins. Four basins have seasonal runoff NSE values of at least 0.89 and the two DA methods result in minimal improvement, 0.02 for both methods. With a sample size of nine, ~~no~~little statistical significance can be attached to these results, but they do suggest DA is more beneficial in poorly calibrated basins. -Future work will examine the potential for DA based on No-DA (open loop) model performances and the characteristics of nearby observed SWE data.~~.~~

Figure ~~5~~ 5 summarizes the ESP evaluation statistics. For simplicity, only the percentile interpolation approach with a 3-month window is shown without forecast error inflation. It shows that for both ESP forcing methodologies used (Section 3.5) in all the case study watersheds, SWE DA enhances seasonal runoff prediction skill, including the probabilistic prediction metric CRPS. Again, higher skill ~~n~~No~~No~~ DA watersheds saw smaller DA improvements. The DA evaluation metric improvement increment versus the corresponding ~~No noDA~~NoDA evaluation metric score for the case basins are shown in Figure 6. ~~It can be seen that t~~The DA improvements in all evaluation metrics have a generally weak negative correlation with N~~n~~o~~No~~ DA performance, which again highlights that better simulated basins benefit less from SWE DA.

4.1.1 Broader DA Potential

~~SWE information from the two data sources (CADWR and SnotelSNOTEL) are available across the western US. Here~~ In general, ~~we find general trends where~~ the incremental DA improvements are ~~generally~~ relatively smaller ~~as~~where the NoDA model performance ~~increases~~is relatively better. However, specific basin performance is dependent on many factors including: 1) representativeness of nearby observations to basin conditions; 2) quality of observations; 3) specific basin characteristics of the calibrated hydrologic model. Because we ~~are operating~~use calibrated, watershed scale hydrologic models, transferability of performance characteristics of the DA ~~system~~approach without implementation in each basin is limited. That being said, Figure 7 displays the difference between the rank correlation of SWE and runoff for the calibrated model (NoDA) and highest correlated observation site (from the nearest 10 sites). It highlights the same general ~~trends~~spatial patterns seen in the 9 basins simulated here. The potential for larger DA improvement appears to be in the Pacific Northwest (upper left of figure). Basins in the Dakotas (upper right basins) are far from ~~Snotel~~SNOTEL sites and have little areal SWE; basins along the far southern US have little SWE and runoff as well. Throughout the central Rockies (central basins), model-observation correlation differences are small, ~~-~~potentially indicating reduced DA improvement potential, in agreement with the results seen above. ~~Again, we note that actual DA performance will vary from basin to basin and actual system implementation is needed.~~

4.2 Case study analyses

To provide a more in-depth examination of the SWE DA impacts to the watershed model states and fluxes, time series of runoff and SWE are shown in Figures 8~~7~~, 9~~8~~ and 10~~9~~ for three example basins, one for each region (the same figures for the other six basins are included in the supplemental material), and for one hindcast year. The feedback from the change of SWE on DA date to seasonal runoff is readily apparent. Increasing the ensemble model SWE through DA will lead to increased model runoff, and vice versa. For basins with a strong seasonal cycle of streamflow (e.g. Greys and Merced River), SWE DA ~~generally~~ may improve~~s~~ daily runoff forecasts in~~in addition to~~ years when seasonal volume forecast improvements are seen, although this is not true in every watershed (e.g. Tolt River). ~~-~~For example, the daily NSE for the Greys River in 1997 after DA was improved from 0.53 to 0.80 in the perfect forcing example, and this is via bias reduction as the daily flow time series is unchanged. In Figure 9~~8~~, the NSE of the daily flow prediction of the Tolt River is essentially unchanged (0.54 for DA, 0.53 for NoDA) even though the seasonal volume prediction is improved (1990 mm observed, 1968 mm DA, 1534 mm NoDA). In this case improvements to bias did not improve NSE as the bias improvements did not improve the squared daily flow differences (e.g. RMSE: 7.76 vs 7.88 for DA vs NoDA).~~, although this is not true in every watershed (e.g. Tolt River).~~

Figures 11~~0~~, 12~~1~~ and 13~~2~~ show several scatter plots of forecast period runoff for the ESP ensemble forcing and perfect forcing forecasts, versus observed runoff, in the three case basins for all of the hindcast years. The left two columns show the comparison for N~~n~~o~~No~~ DA and DA simulated seasonal runoff vs observed runoff for perfect (top row) and ESP ensemble forcing (bottom row) respectively. The 1:1 lines are shown as grey dashed lines and regression lines for the results are shown as green solid line. The results after DA have higher correlation and are generally closer to the 1:1 line, which indicates that for both forcing types SWE DA improves seasonal runoff simulation and prediction skill. The rightmost columns in these three figures show the scatter plots of SWE increment (i.e., SWE analyses states minus model SWE without DA) vs runoff error (i.e., the simulated seasonal runoff without DA minus the observed seasonal runoff). If the runoff errors are positive (the seasonal runoff is overestimated), we would expect the SWE increment to be negative in order to decrease the model seasonal runoff (counteract model error) and vice versa. Thus the ideal results are that the points fall onto different sides of y=0 and x=0 lines (shown as grey dashed lines in this panel), i.e., the points all fall into the 2^{nd} (upper left) and 4^{th} (lower right) quadrants. This is generally the case for our case basins for both perfect and ESP forcing, which again shows that the SWE DA approach is successful in reducing model and forecast error.

For the three basins highlighted here, there are years where the DA SWE increment is not in the 2^{nd} or 4^{th} quadrants. In these years, the increment decreases subsequent forecast skill. Overall, there are 11 of 28 (39%), 4 of 24 (17%), and 12 of 26 (46%) years for the Greys, Tolt and Merced rivers where this is the case using perfect forcing. ~~-~~These years generally correspond to small SWE increments relative to that year's SWE and runoff in all basins except for five years in the Merced River where the SWE increment is larger than 10% of that year's streamflow production and incorrect. ~~–~~In the Greys River, all incorrect increments are less than 10% of the observed runoff for that year and also in years where the NoDA runoff error is less than

10% of observed. A small increment implies that the estimated observed and model SWE are very similar, and thus in years with small model error, the model SWE climatology closely matches observed climatology after transformation for this basin. Figure 14~~3~~ highlights an example WY in the Merced River where the SWE increment and runoff error are both negative, indicating ~~so~~ that DA increased the model forecast error.~~.~~

The Merced River is the only basin to use state of California SWE observations, and these may be of lower quality as evidenced by the large amount of manual quality control we had to perform on the data and the discussion of these data in Lundquist et al. (2015). This suggests that observed SWE data need to be of higher quality (or information content) than the calibrated model SWE to have the positive impact in the DA ~~system~~approach. The calibrated Merced model has -19% April-July runoff bias with 23 (88%) of years having a negative runoff error. EnKF SWE increments are negative in 15 (58%) and positive in 11 (42%) of the years. This indicates ~~suggests~~ that the ~~model~~observed SWE transformation to model space is largely unbiased, but the calibrated model bias impacts SWE DA performance. Calibration of the model specifically for seasonal flow to ensure minimal bias, or hydrologic parameter estimation within the ~~EnKF system~~EnKF approach (e.g. He et al. 2012) would likely improve hydrologic model performance and thus seasonal SWE DA ~~performance~~forecasts in the Merced. Finally, examination of El Nino/La Nina signals (not shown) revealed no clear pattern with degradation of DA forecast skill ~~(not shown)~~.

Finally, there are years where the NoDA runoff error is large, but the SWE increment is small in all three basins. This is not unexpected as spring SWE is not perfectly correlated with subsequent runoff. This may also hint at a level of data loss in the EnKF ~~and modeling system~~approach, and future work should compare streamflow hindcasts using this type of ~~system~~DA approach with traditional statistical methods using SWE as a primary input. It also suggests that improved model calibration, or in combination with model parameter estimation in the ~~EnKF system~~EnKF approach (e.g. He et al. 2012) may improve DA performance across all basins, not just the Merced.

5 ~~Discussion and Conclusions~~Summary and Conclusions

This study tests variants of EnKF SWE DA approaches in 9 case basins in Western US. These basins have seasonal runoff representative of basins used for water resource management across the Western US and have at least 6 close SWE gauge sites with 20+ years of observation history. ~~While SWE observations generally containing valuable information that has potential to enhance seasonal runoff forecasts,. However, relating point SWE measurements that have uneven spatial distribution and varying environmental conditions to watershed mean conditions which is a challenge that is often met by empirical solutions.~~ Two approaches of constructing SWE ensemble observations are examined in this study in an effort to reduce the spatial variability and decrease the interpolation uncertainty while also transforming the observations to model space (e.g., the range of the model climatology). ~~In this study,~~ A~~a~~ 3-month window of SWE observations generally gives the best performance for these two approaches in this study (Figs. 2-4, Tables S.1-6 in S1). However, in some basins a different window length may bring larger improvements. A suitable window length needs to include sufficient samples for transformation as well as including the most relevant samples (i.e., a specific seasonal time period).

~~-~~Sensitivity analyses of model uncertainty impacts on DA performance suggest that either the forcing-alone based estimation of model errors underestimate the total model error variance, or the observed SWE error estimation approaches (interpolation plus the SWE regression) tend to overestimate observation uncertainty, or both (Figs. 3-4, Tables S.1-6 in S1) . ~~It is likely we are underestimating model uncertainty because we have not taken model structural and parameter uncertainty into consideration.~~ Future work should examine this in more detail, as this work clearly indicates that uncertainty scaling approaches (for the model and/or the observations) are likely to be a valuable step for ~~achieving successful DA performance~~further DA ~~r DA .~~

~~The~~improvements.

Encouragingly, t~~T~~he ESP-based assessment of automated SWE DA in the case study watersheds shows clearly the potential for SWE DA to enhance seasonal runoff forecasts ~~in an automated fashion~~, which is notable as the objective incorporation of observed SWE~~is~~ has been a long-standing challenge in operational forecasting. We show at least minor improvement in seasonal runoff forecasts in all nine basins (Figs 5-6). A notable finding is also that the benefits of SWE are linked to the quality of the model simulations of the basin, which can help to target the application of DA to locations where it will have the most benefit (Figs 5-6). For the basins with poor no DA simulations (e.g., the SF Tolt River Fig. 1~~1~~2), the SWE DA can potentially have greater model performance impacts. ~~Broadly speaking t~~The Pacific Northwest and California was found to have the ~~most~~greatest potential for DA improvements to seasonal forecasting in this study (Fig. 7). This stems from ~~reduced~~weaker No DA model performance; the NoDA model run will have more years with larger runoff errors. However, there are still individual years where DA may not improve the forecast. This likely stems from ~~the calibrated~~ hydrologic model~~ not being unbiase~~ bias that ~~d so that~~leads to SWE state corrections ~~often~~ enhancing rather than reducing~~e~~ runoff errors (e.g. Merced River, Figs. 13-14). ~~Additionally, SWE DA can benefit daily streamflow forecasts in some cases (Figs. 7-9).~~

We chos~~e ae~~e a DA update frequency of once per year, the date of climatological maximum correlation of modeled and observed runoff. In operational practice, updates would be applied more frequently, pointing to an area for future research. We note also that this study was conducted using conceptual lumped watershed models, similar to those used in operational practice in the US. As a result, this study ~~did not~~does not shed light on how to address additional challenges that may be associated with using SWE DA in~~for the~~ spatially distributed models, or with spatially continuous datasets (e.g., satellite and remote sensing SWE estimates) that are increasingly being developed or applied in streamflow forecasting contexts. ~~Although S~~SWE DA has been implemented in distributed models in ~~limited~~ prior experimental contexts across large domains (e.g., Wood and Lettenmaier, 2006), but a systematic examination of EnKF DA in spatially distributed hydrological models, coupled with a thoughtful accounting for model parameter and structural errors~~,~~ remains a potentially fruitful area of research and development.

Data Availability

All data used in this study are publicly available. The watershed shapefiles and basin information are described in Newman et al. (2015a) at: doi:10.5065/D6MW2F4D. The forcing ensemble is described in Newman et al. (2015b) and are available at: doi:10.1065/D6TH8JR2. The streamflow data are available through the USGS via: http://waterdata.usgs.gov/usa/nwis/sw and in doi:10.5065/D6MW2F4D. The ~~Snotel~~SNOTEL observations are available at: www.wcc.nrcs.usda.gov/snow/ while the California SWE observations are available at: cdec.water.ca.gov/snow.

Acknowledgements

This work was supported by China Scholarship Council (No. 201406040164), and the NCAR/Research Applications Laboratory; US Department of the Interior Bureau of Reclamation, and US Army Corps of Engineers Climate Preparedness and Resilience Program.

References

Anderson, E., 2002. Calibration of conceptual hydrologic models for use in river forecasting. Office of Hydrologic Development, US National Weather Service, Silver Spring, MD.

Anderson, E.A., 1972. "NWSRFS Forecast Procedures", NOAA Technical Memorandum, NWS HYDRO-14, Office of Hydrologic Development, Hydrology Laboratory, NWS/NOAA, Silver Spring, MD, 1972

Anderson, E.A., 1973. National Weather Service River Forecast System: Snow accumulation and ablation model, 17. US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.

Andreadis, K.M., Lettenmaier, D.P., 2006. Assimilating remotely sensed snow observations into a macroscale hydrology model. Advances in Water Resources, 29(6): 872-886.

Arheimer, B., Lindström, G., Olsson, J., 2011. A systematic review of sensitivities in the Swedish flood-forecasting system. Atmospheric Research, 100: 275–284. doi:10.1016/j.atmosres.2010.09.013.

Barnett, T.P., Adam, J.C., Lettenmaier, D.P., 2005. Potential impacts of a warming climate on water availability in snow-dominated regions. Nature, 438(7066): 303-309.

Barrett, A.P., 2003. National operational hydrologic remote sensing center snow data assimilation system (SNODAS) products at NSIDC. National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences.

Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Pearson correlation coefficient, Noise reduction in speech processing. Springer, pp. 1-4.

Bergeron, J.M., Trudel, M., Leconte, R., 2016. Combined assimilation of streamflow and snow water equivalent for mid-term ensemble streamflow forecasts in snow-dominated regions. Hydrology and Earth System Science Discussions: 1–34. doi:10.5194/hess-2016-166.

Burnash, R., Singh, V., 1995. The NWS river forecast system-catchment modeling. Computer models of watershed hydrology: 311-366.

Burnash, R.J., Ferral, R.L., McGuire, R.A., 1973. A generalized streamflow simulation system, conceptual modeling for digital computers.

Carpenter, T.M., Georgakakos, K.P., 2004. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. Journal of Hydrology, 298(1): 202-221.

Clark, M.P., Hay, L.E., 2004. Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. Journal of Hydrometeorology, 5(1): 15-32.

Clark, M.P. et al., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Advances in Water Resources, 31(10): 1309-1324.

Clark, M.P., Slater, A.G., 2006. Probabilistic quantitative precipitation estimation in complex terrain. Journal of Hydrometeorology, 7(1): 3-22.

Dietz, A.J., Kuenzer, C., Gessner, U., Dech, S., 2012. Remote sensing of snow–a review of available methods. International Journal of Remote Sensing, 33(13): 4094-4134.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res, 28(4): 1015-1031.

Engeset, R.V., Udnæs, H.C., Guneriussen, T., Koren, H., Malnes, E., Solberg, R., Alfnes, E., 2003. Improving runoff simulations using satellite-observed time-series of snow covered area. Nordic Hydrology. 34, 281–294.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics, 53(4): 343-367.

Evensen, G. et al., 2007. Using the EnKF for assisted history matching of a North Sea reservoir model, SPE Reservoir Simulation Symposium. Society of Petroleum Engineers.

Franz, K.J., Hogue, T.S., Barik, M., He, M., 2014. Assessment of SWE data assimilation for ensemble streamflow predictions. Journal of Hydrology, 519: 2737-2746.

Griessinger, N., Seibert, J., Magnusson, J., Jonas, T., 2016. Assessing the benefit of snow data assimilation for runoff modelling in alpine catchments. Hydrology and Earth System Science Discussions: 1–18. doi:10.5194/hess-2016-37.

Harrison, B., Bales, R., 2015. Skill Assessment of Water Supply Outlooks in the Colorado River Basin. Hydrology, 2(3): 112-

131.

He, M., Hogue, T., Margulis, S., Franz, K., 2012~~1~~. An integrated uncertainty and ensemble-based data assimilation approach for improved operational streamflow predictions. Hydrology and Earth System Sciences Discussions, 8(4): 7709-7755.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation of data assimilation: operational, sequential and variational. *J. Meterol. Soc. Of Japan.*, **75**, pp. 181-189.

Jörg-Hess, S., Griessinger, N., Zappa, M., 2015. Probabilistic Forecasts of Snow Water Equivalent and Runoff in Mountainous Areas*. Journal of Hydrometeorology, 16(5): 2169-2186.

Kerr, Y.H. et al., 2001. Soil moisture retrieval from space: The Soil Moisture and Ocean Salinity (SMOS) mission. Geoscience and Remote Sensing, IEEE Transactions on, 39(8): 1729-1735.

Koren, V., Smith, M., Wang, D., Zhang, Z., 2000. Use of soil property data in the derivation of conceptual rainfall-runoff model parameters, 15th Conference on Hydrology, Long Beach, American Meteorological Society, Paper.

Lundquist J D, Hughes M, Henn B, et al.,2015 High-Elevation Precipitation Patterns: Using Snow Measurements to Assess Daily Gridded Datasets across the Sierra Nevada, California. Journal of Hydrometeorology, 16:177-1792.

Mitchell, K.E. et al., 2004. The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. Journal of Geophysical Research: Atmospheres (1984–2012), 109(D7).

Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors, 8(5): 2986-3004.

Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Advances in Water Resources, 28(2): 135-147.

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Monthly Weather Review, 115(7): 1330-1338.

Nash, J., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology, 10(3): 282-290.

Newman, A. J. et al., 2015a. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences, 19(1): 209-223.

Newman, A. J., M. P. Clark, J. Craig, B. Nijssen, A. Wood, E. Gutmann, N. Mizukami, L. Brekke, and J. R. Arnold, 2015b. Gridded ensemble precipitation and temperature estimates for the contiguous United States. J. Hydrometeorology, **16**, 2481-2500.

Pan, M. et al., 2003. Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent. Journal of Geophysical Research: Atmospheres (1984–2012), 108(D22).

Schlosser, C.A. et al., 2000. Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS Phase 2 (d). Monthly Weather Review, 128(2): 301-321.

Seo, D. J., Koren, V., and Cajina, N.: Real-Time Variational Assimilation of Hydrologic and Hydrometeorological Data into Operational Hydrologic Forecasting, J. Hydrometeorol., 4, 627–641, 2003.

Singh, P., Kumar, N., 1997. Impact assessment of climate change on the hydrological response of a snow and glacier melt runoff dominated Himalayan river. Journal of Hydrology, 193(1-4): 316-350.

Slater, A.G., Clark, M.P., 2006. Snow data assimilation via an ensemble Kalman filter. Journal of Hydrometeorology, 7(3): 478-493.

Su, H., Yang, Z.L., Dickinson, R.E., Wilson, C.R., Niu, G.Y., 2010. Multisensor snow data assimilation at the continental scale: The value of Gravity Recovery and Climate Experiment terrestrial water storage information. Journal of Geophysical Research: Atmospheres (1984–2012), 115(D10).

Sun, C., Walker, J.P., Houser, P.R., 2004. A methodology for snow data assimilation in a land surface model. Journal of Geophysical Research: Atmospheres (1984–2012), 109(D8).

Takala, M. et al., 2011. Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements. Remote Sensing of Environment, 115(12): 3517-3529.

Vrugt, J.A., Gupta, H.V., Nualláin, B., Bouten, W., 2006. Real-time data assimilation for operational ensemble streamflow forecasting. Journal of Hydrometeorology, 7(3): 548-565.

Wood, A.W. and D.P. Lettenmaier, 2006, A new approach for seasonal hydrologic forecasting in the western U.S., Bull. Amer. Met. Soc. 87(12), 1699-1712, doi:10.1175/BAMS-87-12-1699.

Wood, A., Kumar, A., Lettenmaier, D., 2005. A retrospective assessment of NCEP climate model-based ensemble hydrologic forecasting in the western United States. Journal of Geophysical Research, 110: D04105.

Wood, A.W., Lettenmaier, D.P., 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. Geophysical Research Letters, 35(14).

Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016. Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill. J. Hydrometeorology, 17: 651-668, doi:10.1175/JHM-D-14-0213.1.

Wu, G., X. Zheng, L. Wang, S. Zhang, X. Liang, and Y. Li, 2012: A new structure for error covariance matrices and their adaptive estimation in EnKF assimilation. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2000.

Table 1 Basin features of nine case basins.

Region	Basin ID	Elevation (m)	Minimum elevation (m)	Maximum elevation (m)	DA date	Basin area (km^{2})	Slope	Forest percent	Basin name
14	09081600	3092.15	2050	4250	April 1	436.88	150.58	0.6136	Crystal River
14	09352900	3459.15	2450	4250	April 1	187.74	156.09	0.5199	Vallecito Creek
17	13023000	2468.57	1750	3450	March 1	1163.72	98.51	0.6753	Greys River
17	12147600	998.25	550	1650	April 1	16.07	159.37	1	SF Tolt River
17	13235000	2077.16	1150	3250	April 1	1158.47	126.25	0.8604	SF Payette River
17	14158790	1210.48	750	1750	March 15	40.76	116.44	1	Smith River
16	10336645	2180.92	1850	2650	April 1	20.09	118.27	0.7136	General Creek
16	10336660	2188.08	1850	2650	April 1	32.46	83.46	0.7908	Blackwood Creek
18	11266500	2576.54	1150	3950	April 1	836.15	140.18	0.6741	Merced River

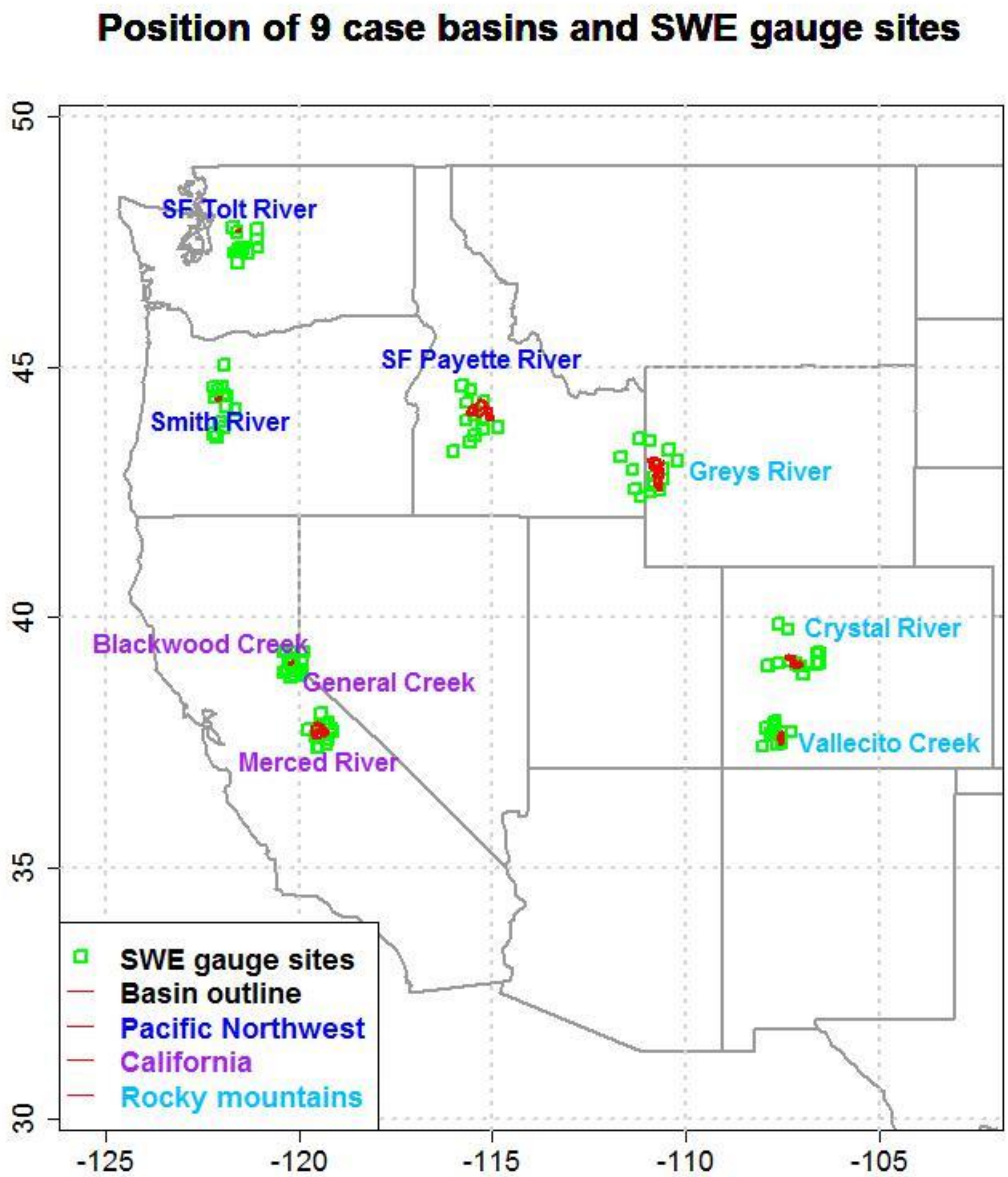

Figure 1. Location of nine case basins in the Western United States (US)~~S~~ and Snow Water Equivalent (SWE) gauge sites.

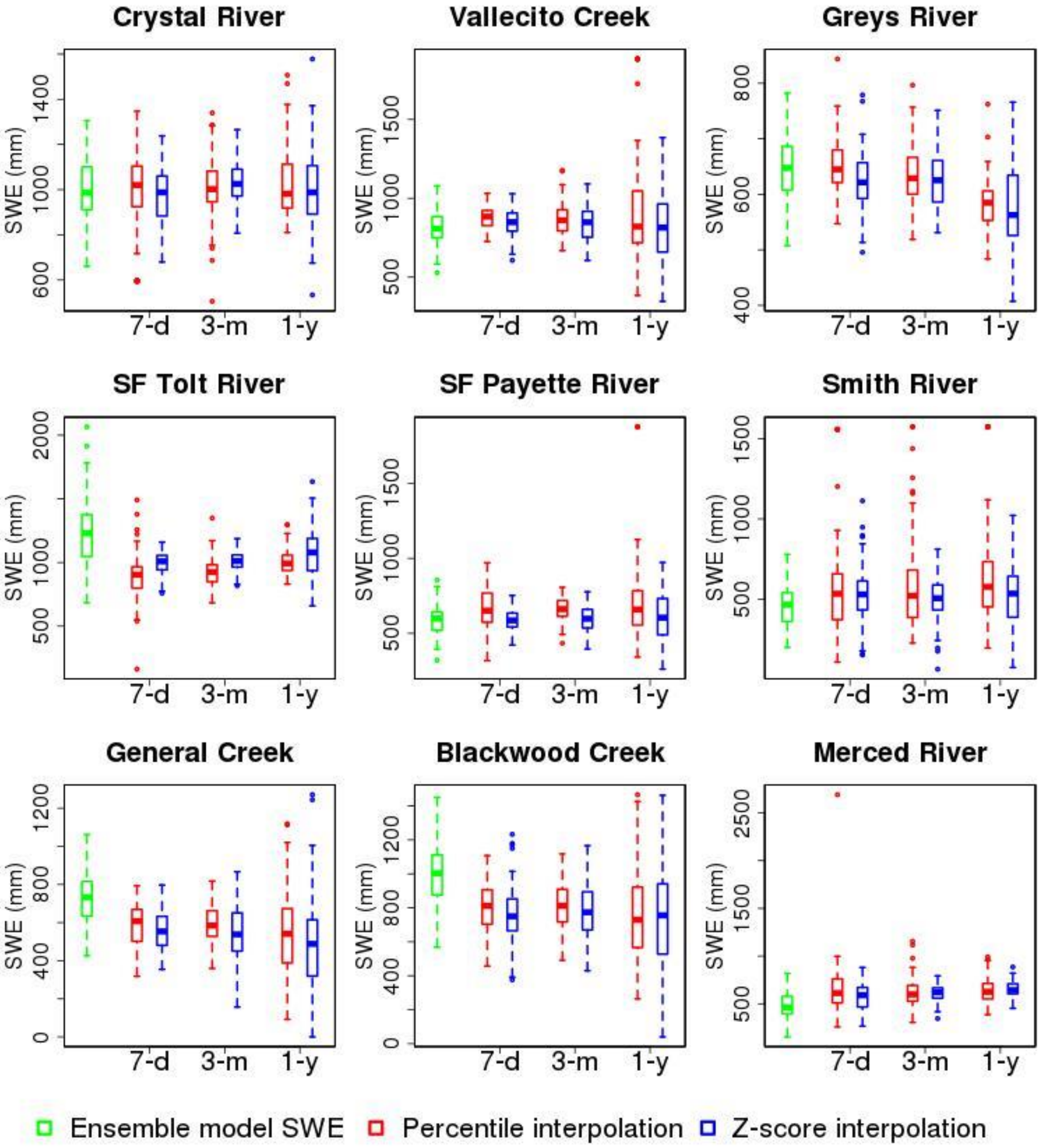

Figure 2. Boxplots of ensemble model SWE and estimated ensemble SWE observations for the nine case basins on the data assimilation~~DA~~ date in 2004, for three window lengths – 7 days, 3 months, and 1 year.

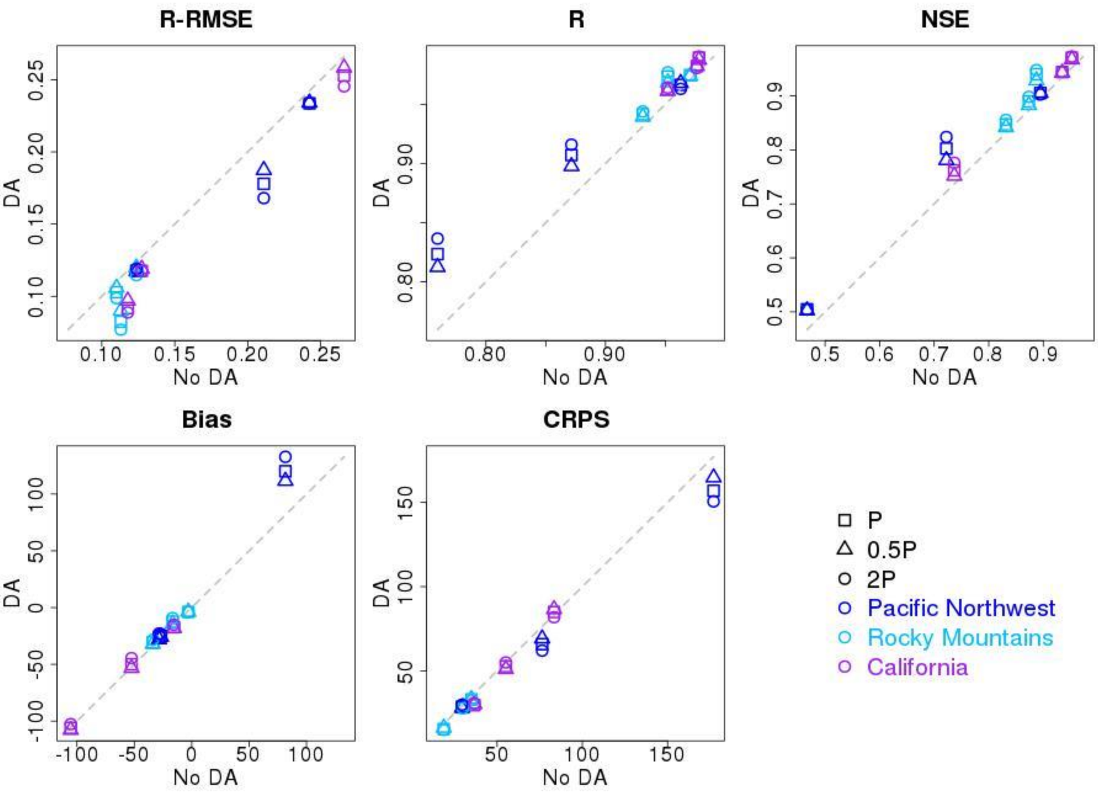

Figure 3. Evaluation metrics for April-July ensemble mean streamflow from the percentile-based interpolation method for the nine case basins using perfect forcing. The verification metrics from upper left to lower right are: R-RMSE is the relative (normalized) root mean squared error, R is the linear (Pearson) correlation coefficient, NSE is the Nash-Sutcliffe Efficiency, bias is the same as mean error, and CRPS is the continuous ranked probability skill scores.

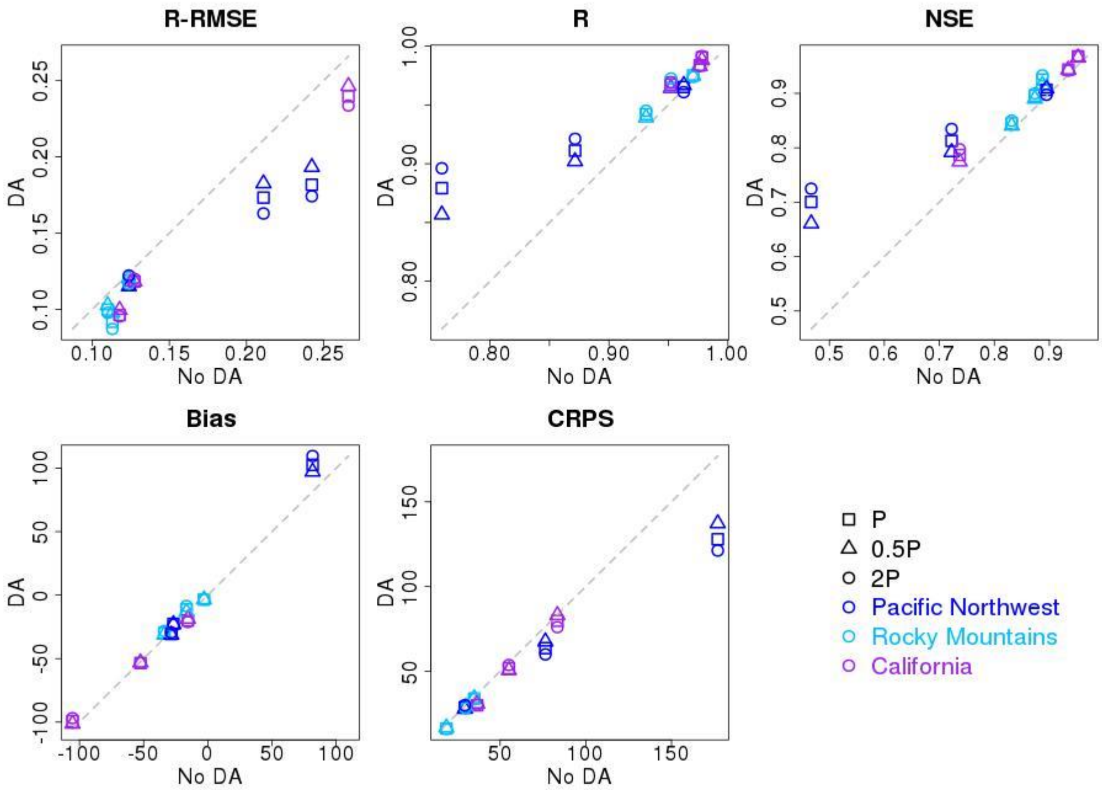

Figure 4. Evaluation metrics for April-July ensemble mean streamflow from the *Z*-score interpolation for the nine case basins using perfect forcing. Verification metrics are the same as Figure 3.

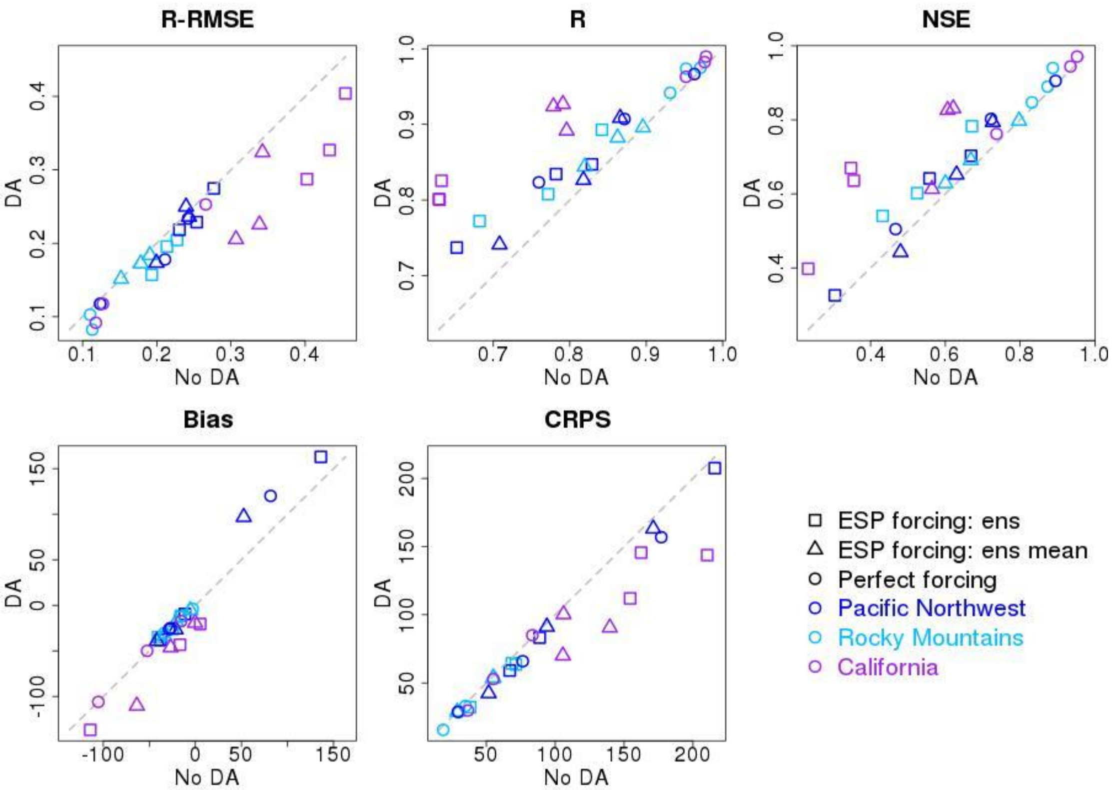

Figure 5. Evaluation statistics of percentile interpolation for the nine case basins with the two variations of Ensemble Streamflow Prediction (~~on~~ ESP) and with perfect forcing data (ens in the legend denotes ensemble). Verification metrics are the same as figure 3.

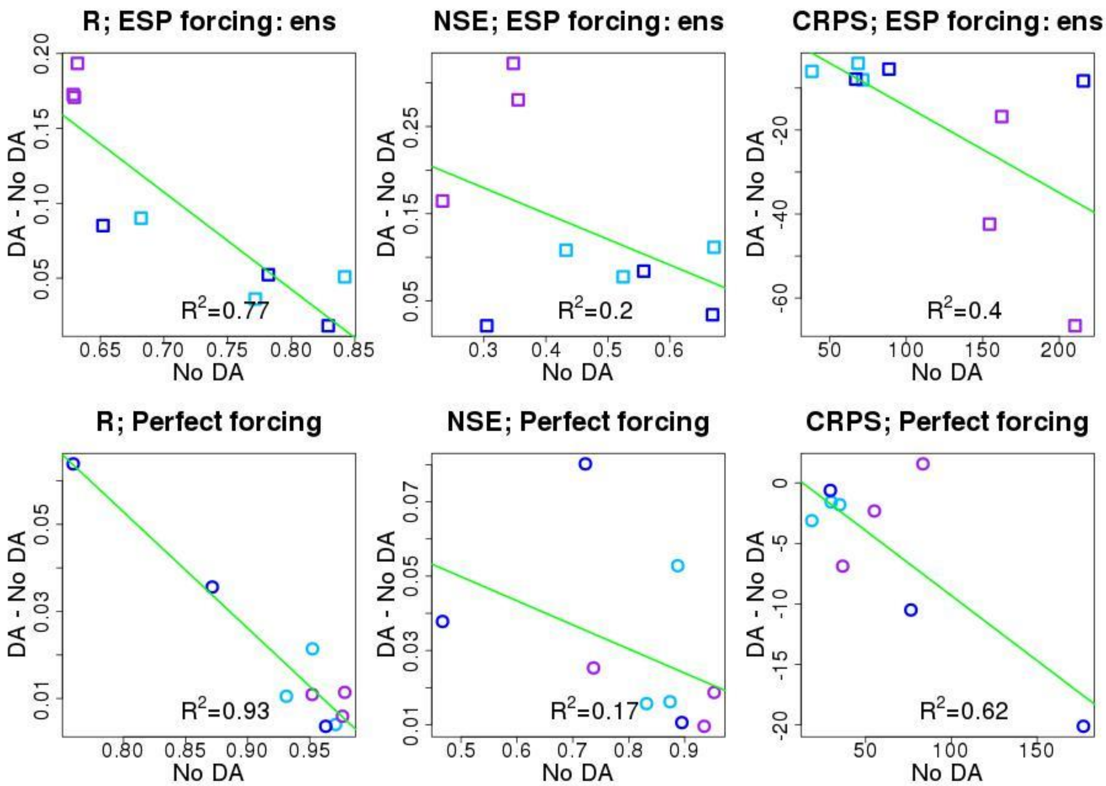

Figure 6. Incremental change in evaluation statistics for Ensemble Streamflow Prediction (ESP) and perfect forcing forecasts using percentile-based interpolation for the nine case basins. R is the linear (Pearson) correlation coefficient, NSE is the Nash-Sutcliffe Efficiency, and CRPS is the continuous ranked probability skill score.

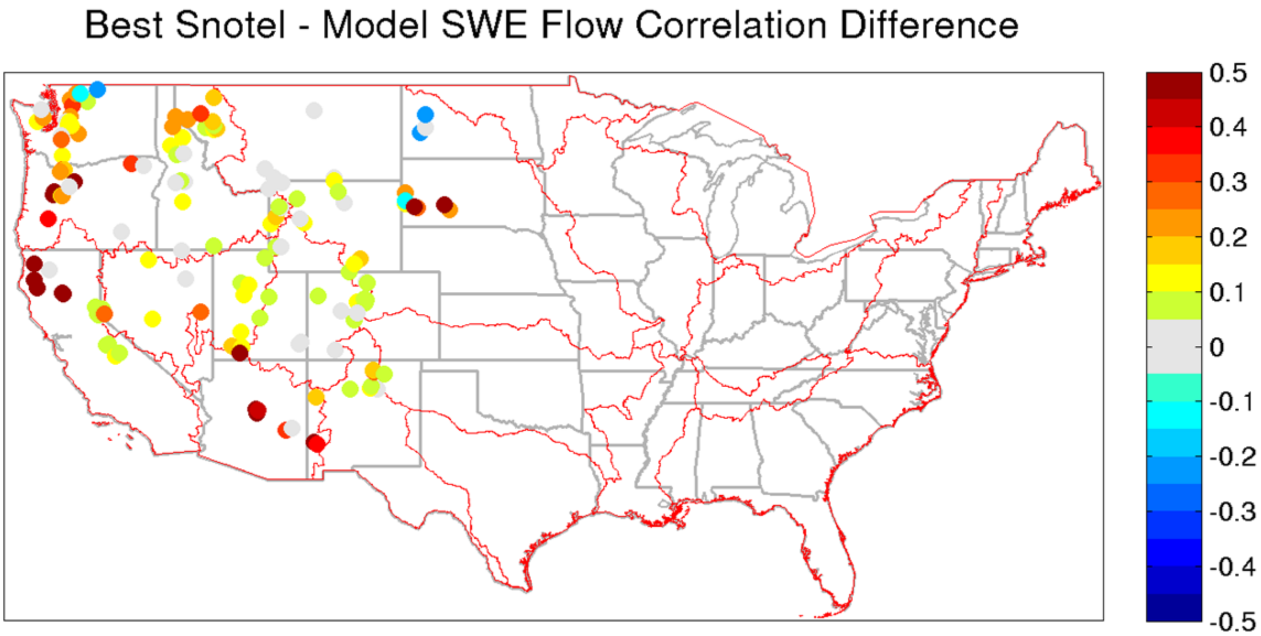

Figure 7. Difference of the rank correlation of SWE and runoff from the best ~~Snotel~~SNOTEL site (of nearest 10) and calibrated model without DA.

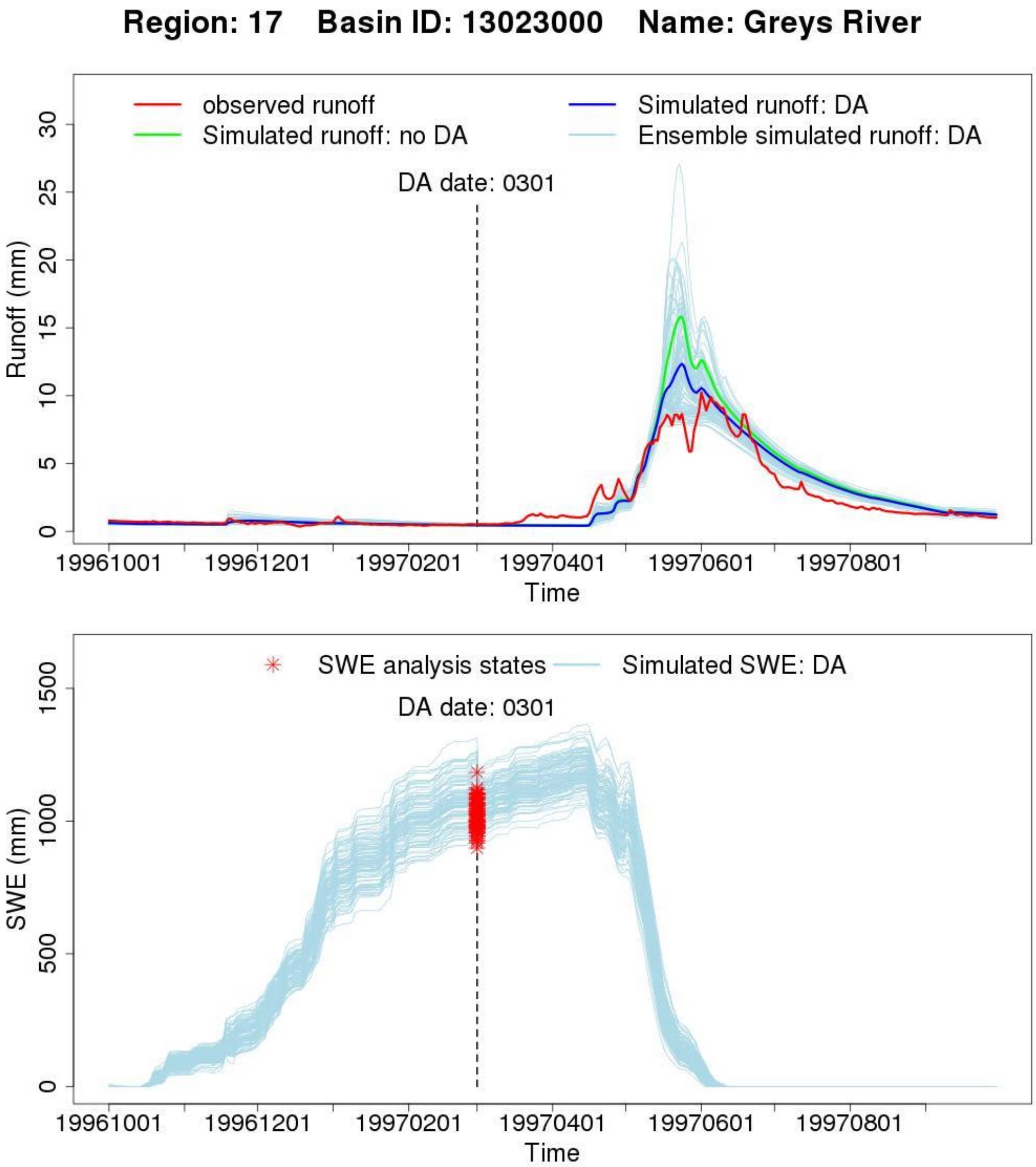

Figure ~~7~~8. Time series plots for runoff and SWE for Greys River for water year 1997. Light blue lines indicate individual ensemble member traces. Vertical black dashed line denotes the data assimilation (DA) date.

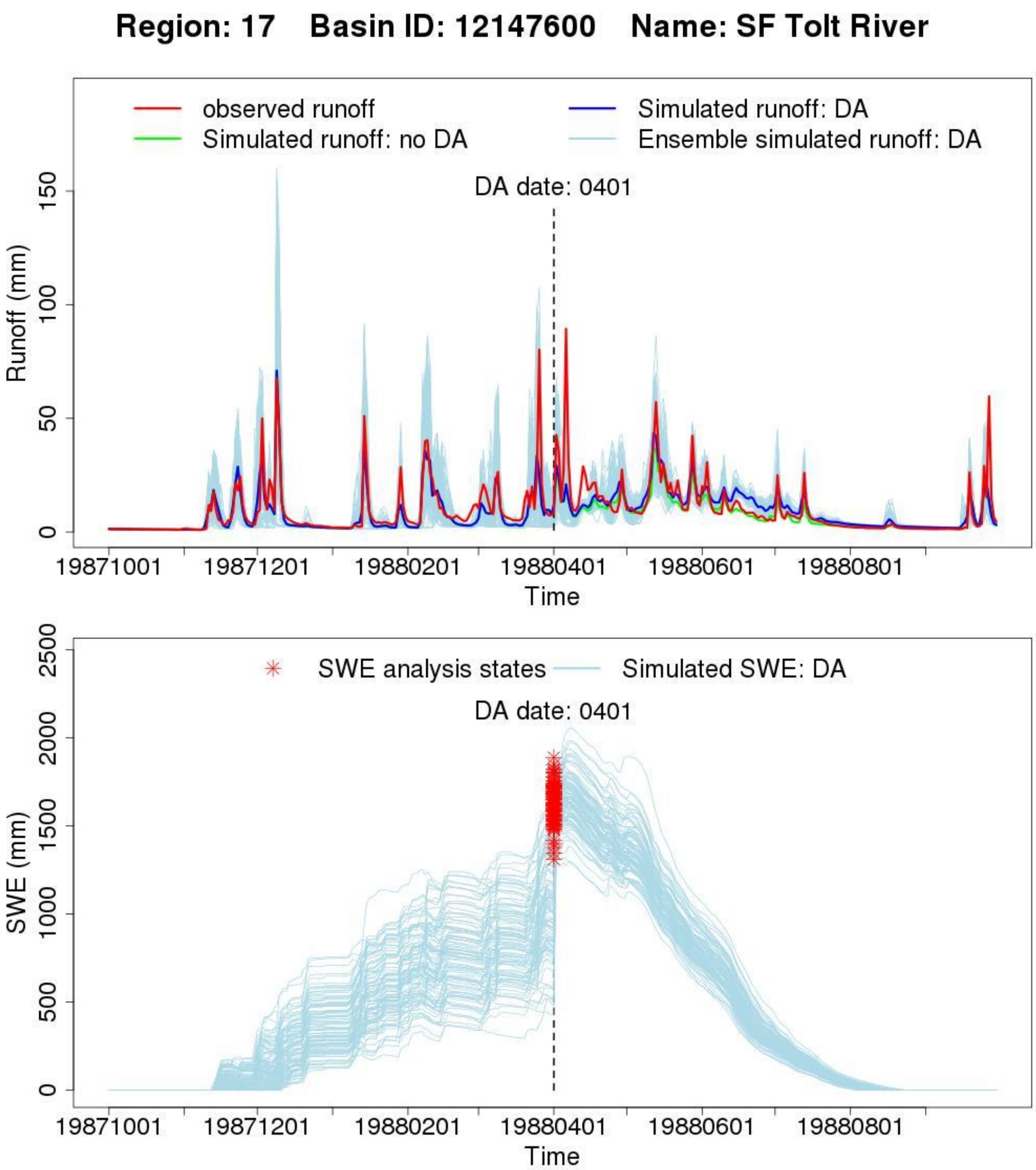

Figure 9~~8~~. Time series plots for runoff and SWE for the South Fork (SF) of the~~SF~~ Tolt River for water year 1988 following Figure 8~~7~~.

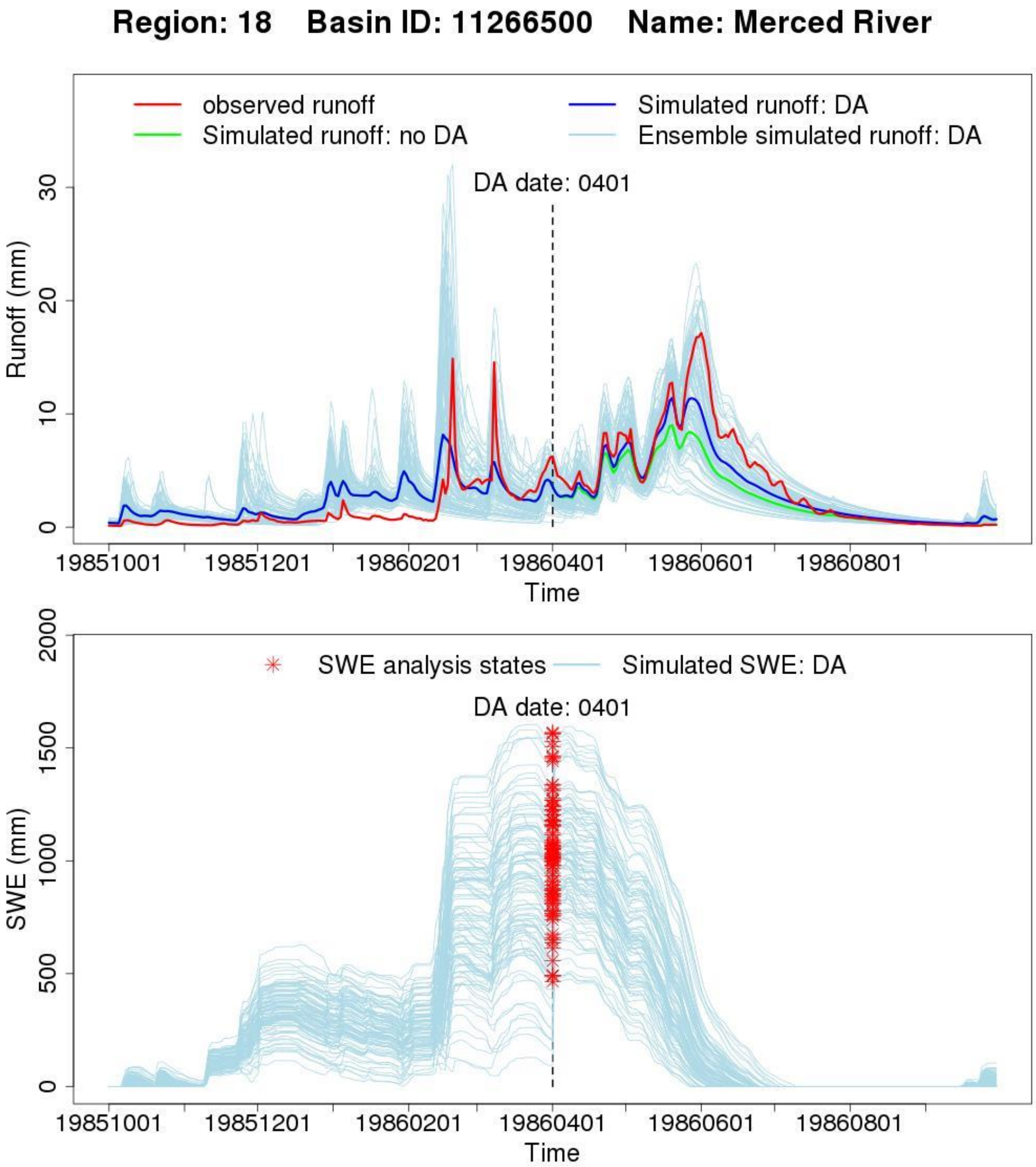

Figure ~~9~~ 10. Time series plots for runoff and SWE for the Merced River for water year 1986 following Figure ~~7~~8.

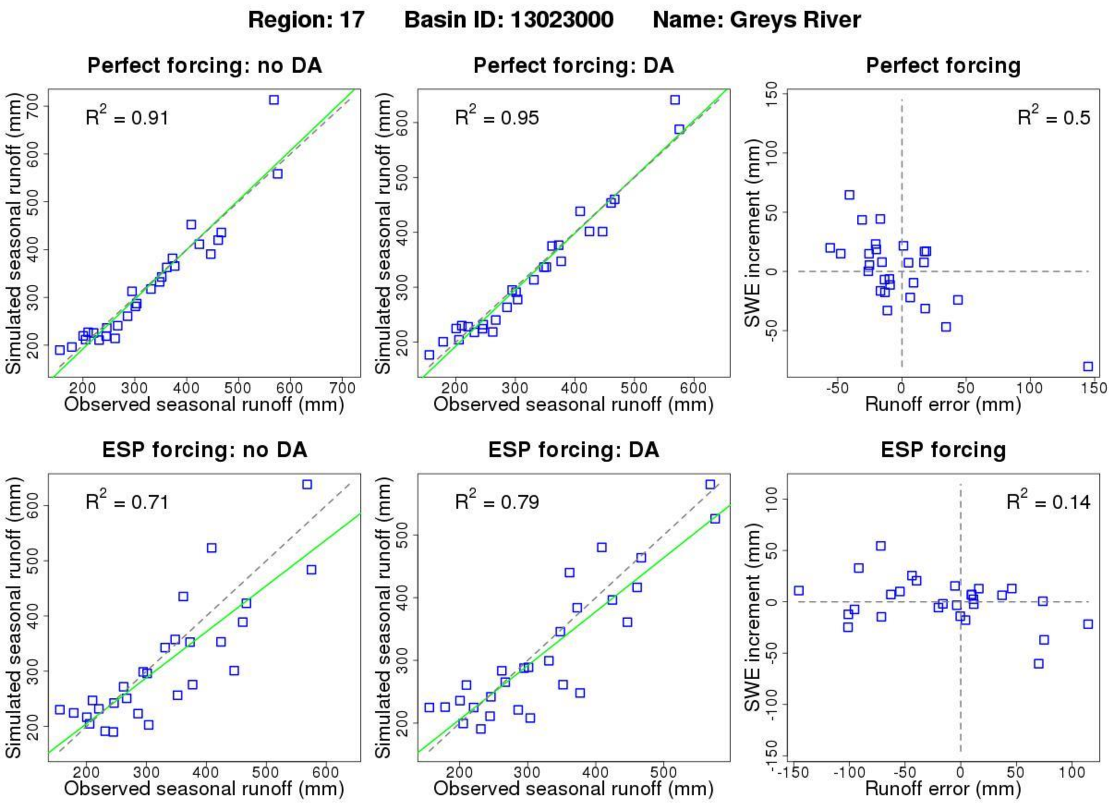

Figure ~~10~~11. Scatter plots for seasonal runoff and SWE on the data assimilation (DA) ~~DA~~ date ~~in the DA years~~ for the Greys River. Black dashed diagonal lines ~~is~~ are the 1:1 line, while the green lines indicates linear regression fits to data. Perfect forcing results are shown in the top row, while Ensemble Streamflow Prediction (ESP) results are in the bottom row.

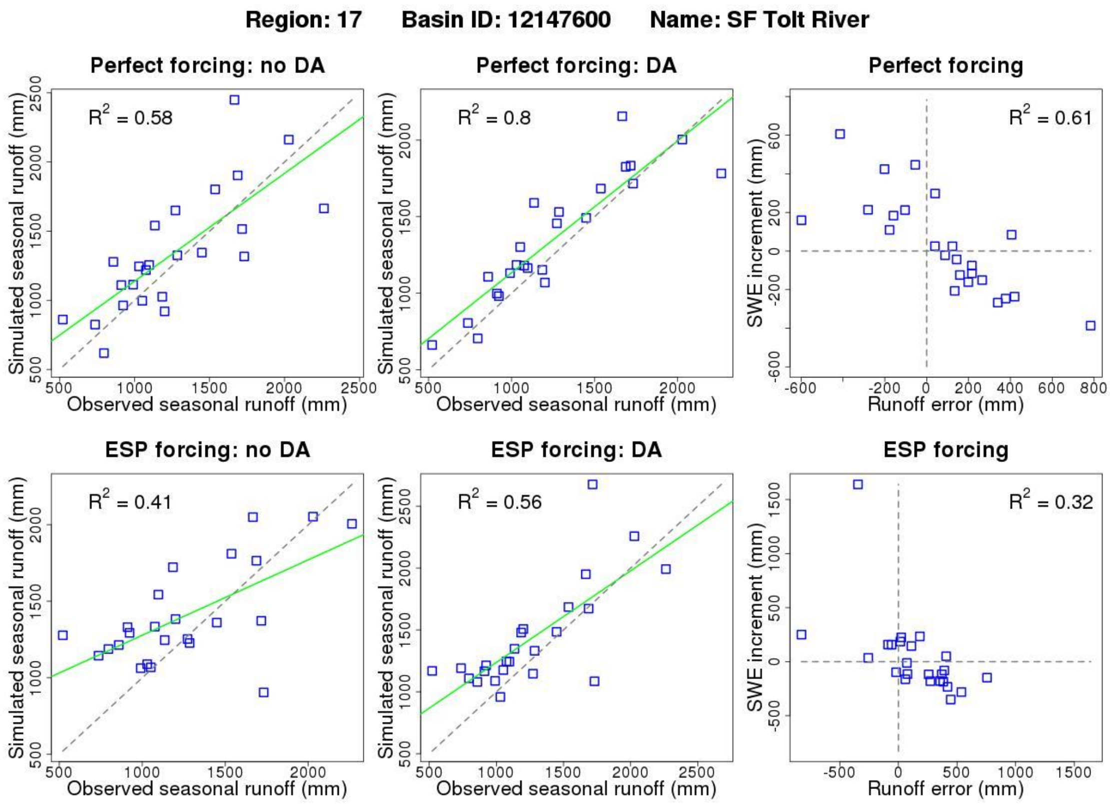

Figure 12~~1~~. Scatter plots for seasonal runoff and SWE on ~~DA~~ the data assimilation (DA) date ~~in the DA years~~ for ~~SF~~ the South Fork of the Tolt River following Figure 1~~0~~1.

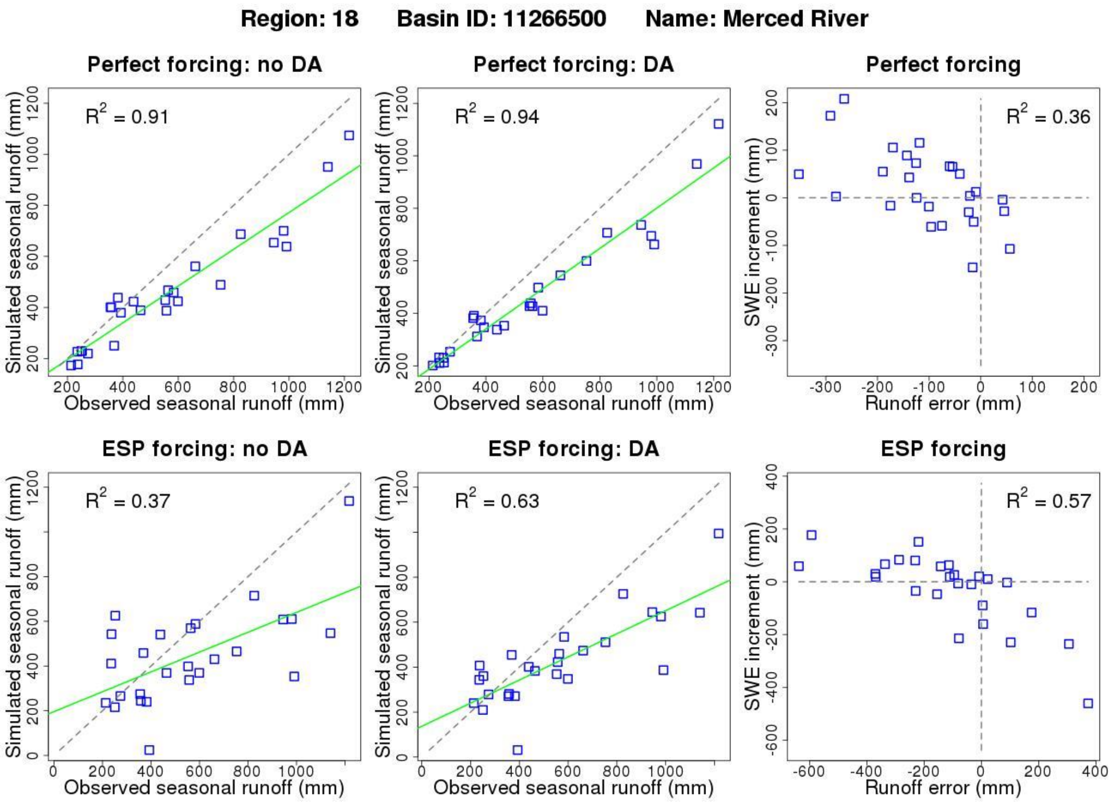

Figure 13~~2~~. Scatter plots for seasonal runoff and SWE on ~~DA~~ data assimilation date (DA) ~~in the DA years~~ for Merced River following Figure 11~~0~~.

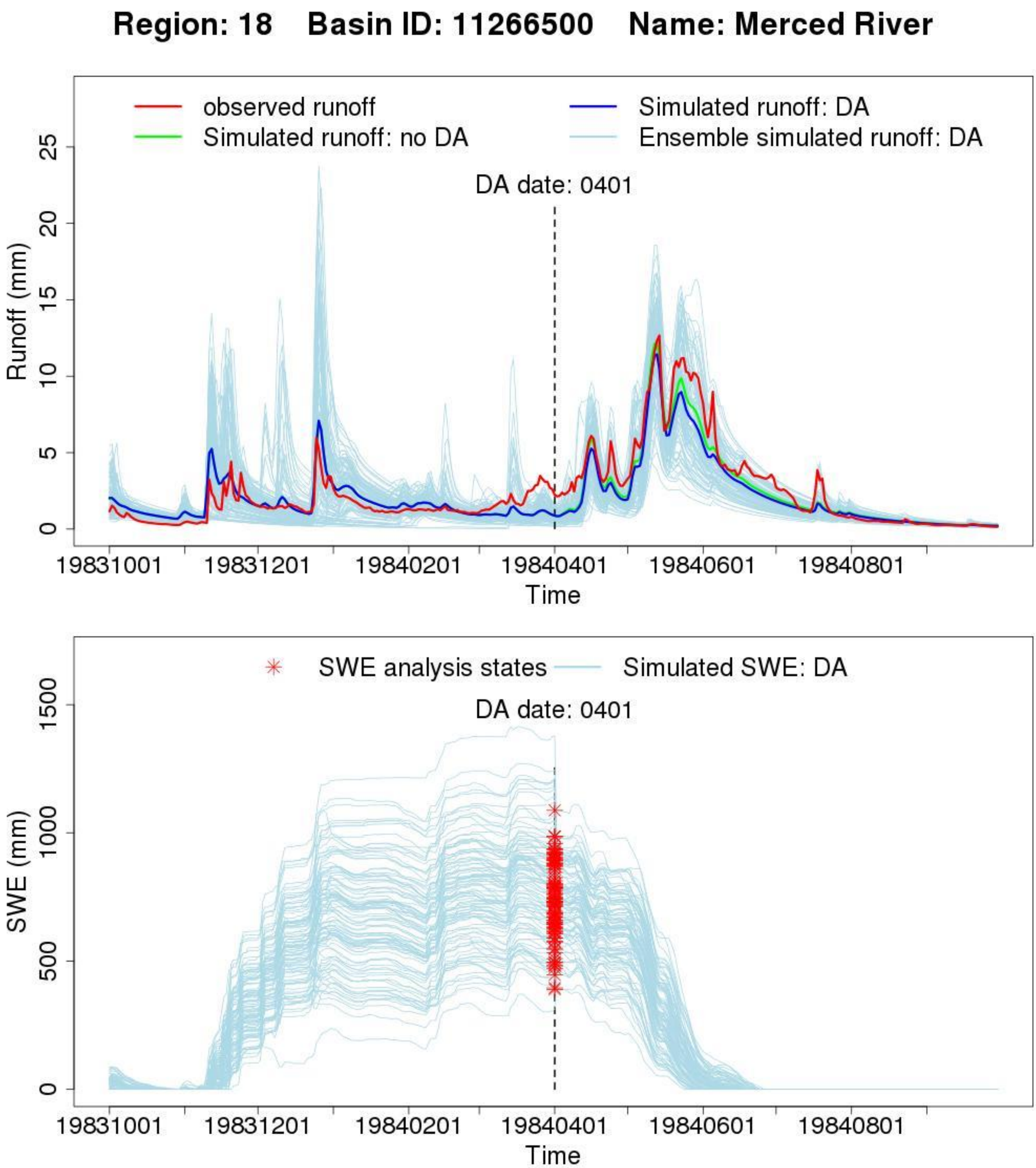

Figure 14. Time series plots for runoff and SWE for the Merced River for water year ~~WY~~ 1984 following Figure ~~7~~8.