

---

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

4 Chengcheng Huang<sup>1,2</sup>, Andrew J. Newman<sup>2</sup>, Martyn P. Clark<sup>2</sup>, Andrew W. Wood<sup>2</sup>  
5 and Xiaogu Zheng<sup>1</sup>

6 <sup>1</sup> College of Global Change and Earth System Science, Beijing Normal University, Beijing, China

7 <sup>2</sup> National Center for Atmospheric Research, Boulder CO, 80301, USA

8 *Correspondence to:* Andrew J. Newman ([anewman@ucar.edu](mailto:anewman@ucar.edu))

9

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

21 *Keywords:*

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

---

## 23 1 Introduction

24 In the snow-dominated watersheds of the Western US, spring snowmelt is a major source of runoff (Barnett et al., 2005; Clark  
25 and Hay, 2004; Singh and Kumar, 1997; Slater and Clark, 2006). In such basins, the initial conditions of the basin, primarily  
26 in the form of snow water equivalent (SWE), drive predictability out to seasonal time scales (Wood et al., 2005; Wood and  
27 Lettenmaier, 2008; Mahanama et al. 2012; Staudinger and Seibert 2014; Wood et al. 2015). Thus better estimates of basin  
28 mean initial SWE should lead to better seasonal streamflow predictions (Arheimer et al., 2011; Clark and Hay, 2004; Slater  
29 and Clark, 2006; Wood et al. 2015). For various reasons (e.g., the uncertainty in model parameters, forcing data, model  
30 structures), simulated SWE in hydrological models can be very different from reality (Pan et al., 2003). Fortunately, a variety  
31 of snow observations (including point gauge and spatial satellite data) contain valuable information (Andreadis and  
32 Lettenmaier, 2006; Barrett, 2003; Engeset et al., 2003; Mitchell et al., 2004; Su et al., 2010; Sun et al., 2004).

33 Many studies have explored the role of snow data assimilation in different modeling frameworks (Moradkhani, 2008; Takala  
34 et al., 2011; McGuire et al, 2006; Wood and Lettenmaier, 2006). Of particular focus here are papers that have examined the  
35 impact of SWE data assimilation (DA) on runoff modelling and prediction (e.g. Bergeron et al., 2016; Griessinger et al., 2016;  
36 Wood and Lettenmaier, 2006; Franz et al., 2014; Jörg-Hess et al., 2015; Moradkhani, 2008; Slater and Clark, 2006). Among  
37 the major challenges facing SWE-based DA are that the time-space resolution of remote sensing SWE data are too coarse or  
38 period-limited for many watershed-scale hydrological applications in mountainous regions (Dietz et al., 2012; Jörg-Hess et al.,  
39 2015), and point gauge snow data have sparse and uneven spatial coverage (Slater and Clark 2006). For point measurements,  
40 spatial interpolation of SWE measurements is typically used to estimate observed SWE state in a watershed of interest  
41 (e.g., Franz et al., 2014; Jörg-Hess et al., 2015; Slater and Clark, 2006; Wood and Lettenmaier, 2006).

42 Here we use the Ensemble Kalman Filter (EnKF) method for DA using an implementation that allowing for seasonally varying  
43 estimates of observation and model error variances (Evensen, 1994, 2003; Evensen et al., 2007). The EnKF framework has  
44 been successfully implemented in research basins in several previous studies (Clark et al., 2008; Franz et al., 2014; Moradkhani  
45 et al., 2005; Slater and Clark, 2006; Vrugt et al., 2006). The EnKF provides an objective analytical framework to optimize the  
46 update of model states based on observed values and their corresponding uncertainties. While the EnKF approach has a formal  
47 theory, its overall objectivity in an application (contrasting with an arbitrary DA approach such as direct insertion) nonetheless  
48 depends on several methodological choices that are often empirical when applied to SWE DA.

49 Following Slater and Clark (2006), this study uses two slightly different approaches to estimate ensemble SWE observations  
50 with point gauge SWE data from surrounding gauge sites for study basins. When using calibrated hydrologic modeling systems,  
51 model SWE states may exhibit systematic biases from observed SWE estimates for a number of reasons – e.g., all hydrologic  
52 models must simplify real watershed physics and structure, and model parameter estimation (calibration) may result in SWE

---

53 behavior that in part compensates for forcing or model errors (e.g. Slater and Clark, 2006). Therefore, transformation of snow  
54 observations to model space is needed before they are used to update the model states to ensure that the model ingests SWE  
55 estimates that are as close to unbiased relative to the model climatology as possible. We explore two variations on an approach  
56 using cumulative density function (CDF) transformations of observations to model space (following Wood and Lettenmaier,  
57 2006, among others). Additionally, we undertake a sensitivity analysis to highlight the importance of robust observations and  
58 model uncertainty estimates. We focus on the impacts of updates made just once per snow accumulation season, noting that an  
59 important choice that is not examined as a result is the selection of DA dates and frequency. For a given generally optimal  
60 selection of the EnKF approach, the Ensemble Streamflow Prediction (ESP) approach is used to test the impact of SWE DA  
61 on subsequent streamflow forecasts.

62 For context, operational seasonal streamflow forecasts in the US currently do not use formalized DA. If the initial states of the  
63 model are suspected to contain error (He et al. 2012), DA is performed through subjective forecaster intervention. Manual  
64 adjustments (termed ‘MODs’, e.g. Anderson 2002) to model states (e.g. SWE) are applied repeatedly throughout the water  
65 year, and particularly before initializing seasonal forecasts. This manual nature of the correction hinders the ability to scale up  
66 DA procedures to many basins, to benchmark DA performance, and quantify improvements to the forecast system as skill  
67 depends on the forecaster’s experience (Seo et al. 2003).

68 The central motivating aim of this study is thus to assess the potential benefits of objective, automated SWE DA against a  
69 reference model configuration to identify forecast improvement opportunities. We apply the EnKF DA approach to nine river  
70 basins in the Western US that have a range of basin features and environmental conditions, over a period of multiple decades.  
71 This experimental scope differs from many previous studies that focus on one or two basins (e.g., Clark et al., 2008; Franz et  
72 al., 2014; He et al., 2012; Moradkhani et al., 2005), or assess DA performance over shorter periods. We also use ensemble  
73 simulations driven by a new probabilistic forcing dataset (Newman et al, 2015) as a basis for estimating model SWE uncertainty,  
74 in contrast to prior studies that relied on more arbitrary distributional assumptions. This range of basins permits us to explore  
75 the question of: “In what types of basins might automated SWE DA improve seasonal streamflow forecasts?”

76 Additionally, as discussed throughout the introduction, the EnKF approach has several empirical components that require  
77 tuning. We therefore examine performance sensitivities related to three elements: 1) the estimation of watershed mean SWE  
78 from surrounding point measurements; 2) the transformation operator that relates watershed mean SWE to model mean SWE;  
79 and 3) sensitivity analyses of the relative size of observed and model error variance.

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

82

---

## 83 2 Study basins and data

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

97

## 98 3 Methodology

### 99 3.1 Models and calibration

100 The Snow-17 temperature index snow model is coupled to the Sacramento Soil Moisture Accounting (SAC-SMA) conceptual  
101 hydrologic model (Anderson, 2002; Anderson, 1973; Burnash and Singh, 1995; Burnash et al., 1973; Franz et al., 2014;  
102 Newman et al., 2015a) to simulate streamflow in this study. This model combination has been in operational use by US National  
103 Weather Service (NWS) River Forecast Centers (RFCs) since the 1970s (Anderson, 1972; 1973). The Snow-17 model is a  
104 conceptual snow pack model that employs an air temperature index to partition precipitation into rain and snow and  
105 parameterize energy exchange and snowpack evolution processes. The only required forcing inputs are near-surface air  
106 temperature and precipitation. The output rain-plus-snowmelt (RAIM) time series from Snow-17 is part of the forcing input  
107 of the SAC-SMA model. SAC-SMA is a conceptual hydrologic model that uses five moisture zones to describe the movement  
108 of water through watersheds. The required forcing input is the potential evaporation and the surface water input from Snow-  
109 17.

110 Daily streamflow data from United States Geological Survey (USGS) National Water Information System server  
111 (<http://waterdata.usgs.gov/usa/nwis/sw>) are used to calibrate 20 parameters of Snow-17 and SAC-SMA model. The calibration  
112 is obtained using the shuffled complex evolution global search algorithm (SCE; Duan et al, 1992) via minimizing daily

---

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

### 122 **3.2 Generating ensembles of estimated observed watershed SWE**

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

#### 133 **3.2.1 Percentile interpolation**

134 First, the non-exceedance percentile  $p_y^o(k)$  of each SWE observation (observation based values noted with superscript o) at  
135 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  
136 years at the same site within a time-window of +/-  $n$  days centered on the date of the observation in each year.

137 Then we use the percentiles to do linear regression on geographic features latitude, longitude and elevation to estimate the  
138 SWE percentile for the target basin:  $\hat{p}_y^o$ , where the hat indicates the basin mean estimate. By LOO cross validation, the  
139 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  
140 ensemble percentiles  $\{\hat{p}_y^o(j)\}$ , where  $j = 1, \dots, 100$  represents ensemble member.

141 Finally, we take the corresponding  $\hat{p}_y^o(j)$  percentile from the full ensemble model SWE within the time-window of +/-  $n$

142 days centered on the DA date each year in all years, denoted as  $\hat{S}_y^f(j)$ . The final ensemble SWE observations on DA date at  
 143 year  $y$  for the target basin are  $\{\hat{S}_y^f(j)\}$ , where  $j = 1, \dots, 100$ .

### 144 3.2.2 Z-score interpolation

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

$$146 \text{Zscore}_y(k) = \frac{S_y^o(k) - \overline{S^o(k)}}{\sigma(S^o(k))}, \quad (1)$$

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

$$153 \hat{S}_y^o(j) = \hat{Z}\text{score}_y^o \times \sigma(S^f(k)) + \overline{S^f(k)}, \quad (2)$$

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

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

### 164 3.3 EnKF approach and experimental design

165 For evaluating the relative performance of DA and for re-initializing the soil moisture of DA runs at the beginning of each  
 166 water year (WY), an open loop or 'control' retrospective simulation (denoted No DA) is performed using the calibrated model  
 167 parameters with ensemble forcing data. This control run is one continuous simulation per ensemble member for the entire  
 168 hindcasting and evaluation period (1981-201X) for each basin. Because this study focuses on assessing variations in  
 169 methodological aspects of the DA approach rather than differences in performance throughout a forecasting season, we apply

170 DA updates only once per year, using the date on which the SWE correlation with future runoff is highest for the study basin,  
171 but no later than 1 April, a common date for initiation of spring seasonal runoff forecasts.

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

$$174 \mathbf{x}_{i+1}^t = M(\mathbf{x}_i^t) + \boldsymbol{\eta}_i, \quad (3)$$

$$175 \mathbf{y}_i^o = \mathbf{h}(\mathbf{x}_i^t) + \boldsymbol{\varepsilon}_i, \quad (4)$$

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

182 The SWE DA approach is implemented via the following procedure:

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

186 2) Calculate the ensemble analysis states:

$$187 \mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{s}_i \mathbf{h}_i^T (\mathbf{h}_i \mathbf{s}_i \mathbf{h}_i^T + \mathbf{o}_i)^{-1} \mathbf{d}_i, \quad (5)$$

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

$$190 \mathbf{d}_i = \mathbf{y}_i^o - \mathbf{h}_i(\mathbf{x}_i^f), \quad (6)$$

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

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

### 197 3.4 Model and observation error variance

198 In this study, only the uncertainty of the forcing data is taken into account in our model uncertainty, and uncertainty that arises  
199 from model structural and parameter errors could cause the true model error to be larger. Thus we assess the impacts of inflating  
200 model error variance to evaluate the relative size of observed and forecast error variance. We simply set the model SWE error

---

201 variance to 1/2 and 2 times of the original size to see how the DA performances change. If increasing the model error variance  
202 results in DA performance improvements, it would indicate that the model error variance is underestimated, and vice versa.  
203 This sensitivity analysis underscores the importance of a careful effort to properly estimate both model and observational  
204 uncertainty when using the EnKF – a challenge that is well known in the DA community.

### 205 **3.5 Seasonal Ensemble Streamflow Prediction**

206 Although the impacts of the SWE DA on forecast accuracy can be assessed through verification of post-adjustment simulations  
207 using ‘perfect’ future forcing, we demonstrate the performance of SWE DA by initializing seasonal ESP forecasts for a  
208 streamflow forecast product that is widely used in water management, the snowmelt-period runoff volume from April through  
209 July. ESP uses historical climate data to represent the future climate conditions each year from the start point of forecast period  
210 to predict streamflow. Two typical ESP applications are tested in this study. Because we have an ensemble of historical forcing  
211 instead of the traditional application in which only a single historical forcing time series is available, there are different ways  
212 to construct an ESP. We adopt two: (1) We construct the ESP forcing ensemble by randomly selecting one year of the historical  
213 ensemble forcing data for each historical member of the ESP; and (2) We use all historical years of ensemble mean forcing  
214 data for each ESP historical year member, yielding a 30\*100 member ensemble for an ESP based on meteorology from 1981-  
215 2010 (variations are noted ens forcing and ens mean forcing respectively in subsequent figures discussing ESP results).

### 216 **3.6 Verification metrics**

217 In this study, five frequently used statistics are calculated for April through July seasonal streamflow volume expressed as  
218 runoff (mm) for evaluating the two DA approaches. The bias, correlation coefficient (R), relative root mean squared error (R-  
219 RMSE), Nash-Sutcliffe efficiency (NSE) are based on the ensemble averages. The continuous ranked probability score (CRPS)  
220 is a measurement of error for probabilistic prediction (Murphy and Winkler, 1987). It is defined as the integrated squared  
221 difference between cumulative distribution function (CDF) of forecasts and observations:

$$222 \text{CRPS} = \int_{-\infty}^{+\infty} [F^f(x) - F^o(x)]^2 dx, \quad (7)$$

223 where  $F^f$  and  $F^o$  are CDFs for forecasts and observations of streamflow respectively. Small CRPS values mean more  
224 accurate forecasts, with 0 value indicating a perfect forecast accuracy.

## 225 **4 Results and Discussion**

### 226 **4.1 Overall performance in the case basins**

227 Using the two approaches described in Section 3.2 with three different window lengths (7 days, 3 months, 1 year), a sample  
228 comparison from one year (2004) of the results for estimated watershed SWE from the two methods versus the model SWE  
229 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  
230 distributions of SWE from the model ensemble and from the percentile and Z-score interpolation methods differ in ways that



---

231 are not consistent across all watersheds. The variance of the estimated observed SWE for both methods is generally largest for  
232 the 1-year, an effect that is more pronounced for the  $Z$ -score interpolation. However, we also note that the ensemble  
233 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  
234 cases. See the percentile interpolation for the Payette River for 7-d window in Figure 2 where the 7-day window interquartile  
235 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  
236 due to the more limited sample size for the regression, which can reduce the positive impact of DA performance. For example,  
237 the SF Payette River and the Greys River have positive DA impact for both the 7-day and 3-month windows but for the 7-day  
238 window the positive impact is reduced by roughly half in both basins for most metrics (Tables S.1 and S.3 of Supplement S1).  
239 Increased estimated observation variance decreases the weight of the observations in an EnKF approach and thus decreases  
240 the impact of the observations. In this study, a 3-month window of SWE observations generally gives the best performance.  
241 However, in some basins a different window length may bring larger improvements. Longer windows mean that the  
242 transformation is more statistically representative of the long-term model-observation climatology. Shorter time windows  
243 imply that the model SWE values used for transformation are more relevant to a specific seasonal time period, avoiding aliasing  
244 for seasonality, but have much smaller sample sizes and may not properly represent the relationship between model and  
245 observation climatologies. The window length must be a balance between these two considerations. Therefore, a 3-month  
246 window is recommended for both approaches.

247 The evaluation statistics for simulated streamflow using perfect forcing after DA with ensemble SWE observations estimated  
248 by the percentile and  $Z$ -score interpolation approaches for the 3-month window are shown in Figures 3 and 4. They are also  
249 compiled in Tables S.1-6 in supplement S1. In those tables, the 2<sup>nd</sup> column shows the forecast error variance used to calculate  
250 analysis states, where “No DA” means the open loop control run (see Section 3.3), and the P,  $1/2 \cdot P$  and  $2 \cdot P$  refer to the DA  
251 runs with the model error variance estimated by 1,  $1/2$  and 2 times the original size of the ensemble model variance. Both  
252 percentile and  $Z$ -score interpolation approaches exhibit enhanced DA performance among the case basins, indicating that both  
253 approaches are effective in adding observation based information to the model simulations. Overall, using the original model  
254 variance estimate (case P) the mean improvement for the percentile interpolation method ( $Z$ -score method) is a reduction in  
255 relative RMSE (R-RMSE) of about 11% (12%) and an increase in NSE of 0.03 (0.05). The percentile interpolation and  $Z$ -  
256 score interpolation methods vary in performance across the basins with both performing better in some basins and not others  
257 (e.g, percentile interpolation performs slightly better than  $Z$ -score interpolation in Grey River using NSE as the evaluation  
258 metric (0.94 vs 0.93) and slightly worse that in SF Tolt River (0.82 vs 0.88)). Using NSE, percentile interpolation performs  
259 better in the Greys River, while  $Z$ -score interpolation performs better in the Vallecito, South Fork of the Tolt, Merced, and  
260 Smith Rivers. To the hundredth NSE value (0.01) both methods are equivalent in the South Fork of the Payette River, and

---

261 General and Blackwood Creeks.

262 The results of forecast error variance inflation shows that for both percentile and Z-score interpolation, “2·P” has better  
263 performance than “P” in most of the case basins – i.e., increasing the model error variance leads the assimilation to trust  
264 observations more and improves the DA performance (circles in both figures generally have improved evaluation metrics than  
265 squares or triangles). Using NSE, the percentile (Z-score) interpolation “2·P” case is on average another 0.01 (0.01) better than  
266 the “P” case across the nine basins. This sensitivity analysis of model uncertainty impacts on DA performance suggest that  
267 either the forcing-alone based estimation of model errors underestimate the total model error variance, or the observed SWE  
268 error estimation approaches (interpolation plus the SWE regression) tend to overestimate observation uncertainty, or both. It  
269 is likely we are underestimating model uncertainty because we have not taken model structural and parameter uncertainty into  
270 consideration. Both approaches bring incremental enhancements to the ensemble mean streamflow hindcast in most basins  
271 when evaluated across the R-RMSE, R and NSE metrics, however DA does not help correct forecast biases in these simulations.  
272 Post-processing procedures (e.g. bias correction) could be used to further enhance the forecast performance, but is not a focus  
273 of this study. These figures also show that forecasts without DA (“No DA” in figures, “NoDA” in text) that have relatively  
274 better performance, mostly due to better simulations of forecast initial conditions, benefit less from DA. Three of the basins  
275 have a NoDA seasonal runoff NSE of less than 0.8, with an average improvement of 0.05 for the percentile regression and  
276 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  
277 least 0.89 and the two DA methods result in minimal improvement, 0.02 for both methods. With a sample size of nine, little  
278 statistical significance can be attached to these results, but they do suggest DA is more beneficial in poorly calibrated basins.  
279 Future work will examine the potential for DA based on NoDA (open loop) model performances and the characteristics of  
280 nearby observed SWE data.

281 Figure 5 summarizes the ESP evaluation statistics. For simplicity, only the percentile interpolation approach with a 3-month  
282 window is shown without forecast error inflation. It shows that for both ESP forcing methodologies used (Section 3.5) in all  
283 the case study watersheds, SWE DA enhances seasonal runoff prediction skill, including the probabilistic prediction metric  
284 CRPS. Again, higher skill NoDA watersheds saw smaller DA improvements. The DA evaluation metric improvement  
285 increment versus the corresponding NoDA evaluation metric score for the case basins are shown in Figure 6. The DA  
286 improvements in all evaluation metrics have a generally weak negative correlation with NoDA performance, which again  
287 highlights that better simulated basins benefit less from SWE DA.

#### 288 **4.1.1 Broader DA Potential**

289 In general, the incremental DA improvements are relatively smaller where the NoDA model performance is relatively better.  
290 However, specific basin performance is dependent on many factors including: 1) representativeness of nearby observations to

---

291 basin conditions; 2) quality of observations; 3) specific basin characteristics of the calibrated hydrologic model. Because we  
292 use calibrated, watershed scale hydrologic models, transferability of performance characteristics of the DA approach without  
293 implementation in each basin is limited. That being said, Figure 7 displays the difference between the rank correlation of SWE  
294 and runoff for the calibrated model (NoDA) and highest correlated observation site (from the nearest 10 sites). It highlights  
295 the same general spatial patterns seen in the 9 basins simulated here. The potential for larger DA improvement appears to be  
296 in the Pacific Northwest (upper left of figure). Basins in the Dakotas (upper right basins) are far from SNOTEL sites and have  
297 little areal SWE; basins along the far southern US have little SWE and runoff as well. Throughout the central Rockies (central  
298 basins), model-observation correlation differences are small, potentially indicating reduced DA improvement potential, in  
299 agreement with the results seen above.

#### 300 **4.2 Case study analyses**

301 To provide a more in-depth examination of the SWE DA impacts to the watershed model states and fluxes, time series of  
302 runoff and SWE are shown in Figures 8, 9 and 10 for three example basins, one for each region (the same figures for the other  
303 six basins are included in the supplemental material), and for one hindcast year. The feedback from the change of SWE on DA  
304 date to seasonal runoff is readily apparent. Increasing the ensemble model SWE through DA will lead to increased model  
305 runoff, and vice versa. For basins with a strong seasonal cycle of streamflow (e.g. Greys and Merced River), SWE DA may  
306 improve daily runoff forecasts in years when seasonal volume forecast improvements are seen, although this is not true in  
307 every watershed (e.g. Tolt River). For example, the daily NSE for the Greys River in 1997 after DA was improved from 0.53  
308 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,  
309 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  
310 seasonal volume prediction is improved (1990 mm observed, 1968 mm DA, 1534 mm NoDA). In this case improvements to  
311 bias did not improve NSE as the bias improvements did not improve the squared daily flow differences (e.g. RMSE: 7.76 vs  
312 7.88 for DA vs NoDA).

313 Figures 11, 12 and 13 show several scatter plots of forecast period runoff for the ESP ensemble forcing and perfect forcing  
314 forecasts, versus observed runoff, in the three case basins for all of the hindcast years. The left two columns show the  
315 comparison for NoDA and DA simulated seasonal runoff vs observed runoff for perfect (top row) and ESP ensemble forcing  
316 (bottom row) respectively. The 1:1 lines are shown as grey dashed lines and regression lines for the results are shown as green  
317 solid line. The results after DA have higher correlation and are generally closer to the 1:1 line, which indicates that for both  
318 forcing types SWE DA improves seasonal runoff simulation and prediction skill. The rightmost columns in these three figures  
319 show the scatter plots of SWE increment (i.e., SWE analyses states minus model SWE without DA) vs runoff error (i.e., the  
320 simulated seasonal runoff without DA minus the observed seasonal runoff). If the runoff errors are positive (the seasonal runoff

---

321 is overestimated), we would expect the SWE increment to be negative in order to decrease the model seasonal runoff  
322 (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  
323 (shown as grey dashed lines in this panel), i.e., the points all fall into the 2<sup>nd</sup> (upper left) and 4<sup>th</sup> (lower right) quadrants. This  
324 is generally the case for our case basins for both perfect and ESP forcing, which again shows that the SWE DA approach is  
325 successful in reducing model and forecast error.

326 For the three basins highlighted here, there are years where the DA SWE increment is not in the 2<sup>nd</sup> or 4<sup>th</sup> quadrants. In these  
327 years, the increment decreases subsequent forecast skill. Overall, there are 11 of 28 (39%), 4 of 24 (17%), and 12 of 26 (46%)  
328 years for the Greys, Tolt and Merced rivers where this is the case using perfect forcing. These years generally correspond to  
329 small SWE increments relative to that year's SWE and runoff in all basins except for five years in the Merced River where the  
330 SWE increment is larger than 10% of that year's streamflow production and incorrect. In the Greys River, all incorrect  
331 increments are less than 10% of the observed runoff for that year and also in years where the NoDA runoff error is less than  
332 10% of observed. A small increment implies that the estimated observed and model SWE are very similar, and thus in years  
333 with small model error, the model SWE climatology closely matches observed climatology after transformation for this basin.  
334 Figure 14 highlights an example WY in the Merced River where the SWE increment and runoff error are both negative,  
335 indicating that DA increased the model forecast error.

336 The Merced River is the only basin to use state of California SWE observations, and these may be of lower quality as evidenced  
337 by the large amount of manual quality control we had to perform on the data and the discussion of these data in Lundquist et  
338 al. (2015). This suggests that observed SWE data need to be of higher quality (or information content) than the calibrated  
339 model SWE to have the positive impact in the DA approach. The calibrated Merced model has -19% April-July runoff bias  
340 with 23 (88%) of years having a negative runoff error. EnKF SWE increments are negative in 15 (58%) and positive in 11  
341 (42%) of the years. This indicates that the observed SWE transformation to model space is largely unbiased, but the calibrated  
342 model bias impacts SWE DA performance. Calibration of the model specifically for seasonal flow to ensure minimal bias, or  
343 hydrologic parameter estimation within the EnKF approach (e.g. He et al. 2012) would likely improve hydrologic model  
344 performance and thus seasonal SWE DA forecasts in the Merced. Finally, examination of El Nino/La Nina signals (not shown)  
345 revealed no clear pattern with degradation of DA forecast skill.

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

---

351 just the Merced.

## 352 **5 Summary and Conclusions**

353 This study tests variants of EnKF SWE DA approaches in 9 case basins in Western US. These basins have seasonal runoff  
354 representative of basins used for water resource management across the Western US and have at least 6 close SWE gauge sites  
355 with 20+ years of observation history. Two approaches of constructing SWE ensemble observations, percentile and Z-score  
356 interpolation, are examined in this study in an effort to reduce the spatial variability and decrease the interpolation uncertainty  
357 while also transforming the observations to model space (e.g., the range of the model climatology). A 3-month window of  
358 SWE observations generally gives the best performance for these two approaches in this study (Figs. 2-4, Tables S.1-6 in S1).  
359 However, in some basins a different window length may bring larger improvements. A suitable window length needs to include  
360 sufficient samples for transformation as well as including the most relevant samples (i.e., a specific seasonal time period).  
361 Sensitivity analyses of model uncertainty impacts on DA performance suggest that either the forcing-alone based estimation  
362 of model errors underestimate the total model error variance, or the observed SWE error estimation approaches (interpolation  
363 plus the SWE regression) tend to overestimate observation uncertainty, or both (Figs. 3-4, Tables S.1-6 in S1). Future work  
364 should examine this in more detail, as this work clearly indicates that uncertainty scaling approaches (for the model and/or the  
365 observations) are likely to be a valuable step for further DA improvements.

366 Encouragingly, the ESP-based assessment of automated SWE DA in the case study watersheds shows clearly the potential for  
367 SWE DA to enhance seasonal runoff forecasts, which is notable as the objective incorporation of observed SWE has been a  
368 long-standing challenge in operational forecasting. We show at least minor improvement in seasonal runoff forecasts in all  
369 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  
370 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  
371 basins with poor no DA simulations (e.g., the SF Tolt River Fig. 12), the SWE DA can potentially have greater model  
372 performance impacts. The Pacific Northwest and California was found to have the greatest potential for DA improvements to  
373 seasonal forecasting in this study (Fig. 7). This stems from weaker NoDA model performance; the NoDA model run will have  
374 more years with larger runoff errors. However, there are still individual years where DA may not improve the forecast. This  
375 likely stems from hydrologic model bias that leads to SWE state corrections enhancing rather than reducing runoff errors (e.g.  
376 Merced River, Figs. 13-14).

377 We chose a DA update frequency of once per year, the date of climatological maximum correlation of modeled and observed  
378 runoff. In operational practice, updates would be applied more frequently, pointing to an area for future research. We note also  
379 that this study was conducted using conceptual lumped watershed models, similar to those used in operational practice in the  
380 US. As a result, this study does not shed light on how to address additional challenges that may be associated with using SWE

---

381 DA in spatially distributed models, or with spatially continuous datasets (e.g., satellite and remote sensing SWE estimates)  
382 that are increasingly being developed or applied in streamflow forecasting contexts. SWE DA has been implemented in  
383 distributed models in prior experimental contexts across large domains (e.g., Wood and Lettenmaier, 2006), but a systematic  
384 examination of EnKF DA in spatially distributed hydrological models, coupled with a thoughtful accounting for model  
385 parameter and structural errors remains a potentially fruitful area of research and development.

386

### 387 **Data Availability**

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

393

### 394 **Acknowledgements**

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

398

### 399 **References**

- 400 Anderson, E., 2002. Calibration of conceptual hydrologic models for use in river forecasting. Office of Hydrologic  
401 Development, US National Weather Service, Silver Spring, MD.
- 402 Anderson, E.A., 1972. "NWSRFS Forecast Procedures", NOAA Technical Memorandum, NWS HYDRO-14, Office of  
403 Hydrologic Development, Hydrology Laboratory, NWS/NOAA, Silver Spring, MD, 1972
- 404 Anderson, E.A., 1973. National Weather Service River Forecast System: Snow accumulation and ablation model, 17. US  
405 Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.
- 406 Andreadis, K.M., Lettenmaier, D.P., 2006. Assimilating remotely sensed snow observations into a macroscale hydrology model.  
407 *Advances in Water Resources*, 29(6): 872-886.
- 408 Arheimer, B., Lindström, G., Olsson, J., 2011. A systematic review of sensitivities in the Swedish flood-forecasting system.  
409 *Atmospheric Research*, 100: 275–284. doi:10.1016/j.atmosres.2010.09.013.
- 410 Barnett, T.P., Adam, J.C., Lettenmaier, D.P., 2005. Potential impacts of a warming climate on water availability in snow-

---

411 dominated regions. *Nature*, 438(7066): 303-309.

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

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

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

419 Burnash, R., Singh, V., 1995. The NWS river forecast system-catchment modeling. *Computer models of watershed hydrology*:  
420 311-366.

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

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

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

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

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

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

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

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

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

439 Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):  
440 343-367.

---

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

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

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

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

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

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

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

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

457 Mahanama, S., B. Livneh, R. Koster, D. Lettenmaier, and R. Reichle, 2012: Soil moisture, snow, and seasonal streamflow  
458 forecasts in the United States. *J. Hydrometeorol.* **13**, 189-203.

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

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

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

465 Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. *Monthly Weather Review*, 115(7): 1330-  
466 1338.

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

469 Newman, A. J. et al., 2015a. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous  
470 USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth*



---

471 System Sciences, 19(1): 209-223.

472 Newman, A. J., M. P. Clark, J. Craig, B. Nijssen, A. Wood, E. Gutmann, N. Mizukami, L. Brekke, and J. R. Arnold, 2015b.

473 Gridded ensemble precipitation and temperature estimates for the contiguous United States. *J. Hydrometeorology*, **16**, 2481-

474 2500.

475 Pan, M. et al., 2003. Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation

476 of model simulated snow water equivalent. *Journal of Geophysical Research: Atmospheres* (1984–2012), 108(D22).

477 Schlosser, C.A. et al., 2000. Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS Phase 2 (d). *Monthly*

478 *Weather Review*, 128(2): 301-321.

479 Seo, D. J., Koren, V., and Cajina, N.: Real-Time Variational Assimilation of Hydrologic and Hydrometeorological Data into Operational

480 Hydrologic Forecasting. *J. Hydrometeorol.*, 4, 627–641, 2003.

481 Singh, P., Kumar, N., 1997. Impact assessment of climate change on the hydrological response of a snow and glacier melt

482 runoff dominated Himalayan river. *Journal of Hydrology*, 193(1-4): 316-350.

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

484 478-493.

485 Staudinger, M., and J. Seibert, 2014: Predictability of low flow – An assessment with simulation experiments. *J. Hydrology*,

486 **519**, 1383-1393, doi:10.1016/j.jhydrol.2014.08.061.

487 Su, H., Yang, Z.L., Dickinson, R.E., Wilson, C.R., Niu, G.Y., 2010. Multisensor snow data assimilation at the continental scale:

488 The value of Gravity Recovery and Climate Experiment terrestrial water storage information. *Journal of Geophysical*

489 *Research: Atmospheres* (1984–2012), 115(D10).

490 Sun, C., Walker, J.P., Houser, P.R., 2004. A methodology for snow data assimilation in a land surface model. *Journal of*

491 *Geophysical Research: Atmospheres* (1984–2012), 109(D8).

492 Takala, M. et al., 2011. Estimating northern hemisphere snow water equivalent for climate research through assimilation of

493 space-borne radiometer data and ground-based measurements. *Remote Sensing of Environment*, 115(12): 3517-3529.

494 Vrugt, J.A., Gupta, H.V., Nualláin, B., Bouten, W., 2006. Real-time data assimilation for operational ensemble streamflow

495 forecasting. *Journal of Hydrometeorology*, 7(3): 548-565.

496 Wood, A.W. and D.P. Lettenmaier, 2006, A new approach for seasonal hydrologic forecasting in the western U.S., *Bull. Amer.*

497 *Met. Soc.* 87(12), 1699-1712, doi:10.1175/BAMS-87-12-1699.

498 Wood, A., Kumar, A., Lettenmaier, D., 2005. A retrospective assessment of NCEP climate model-based ensemble hydrologic

499 forecasting in the western United States. *Journal of Geophysical Research*, 110: D04105.

500 Wood, A.W., Lettenmaier, D.P., 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophysical*

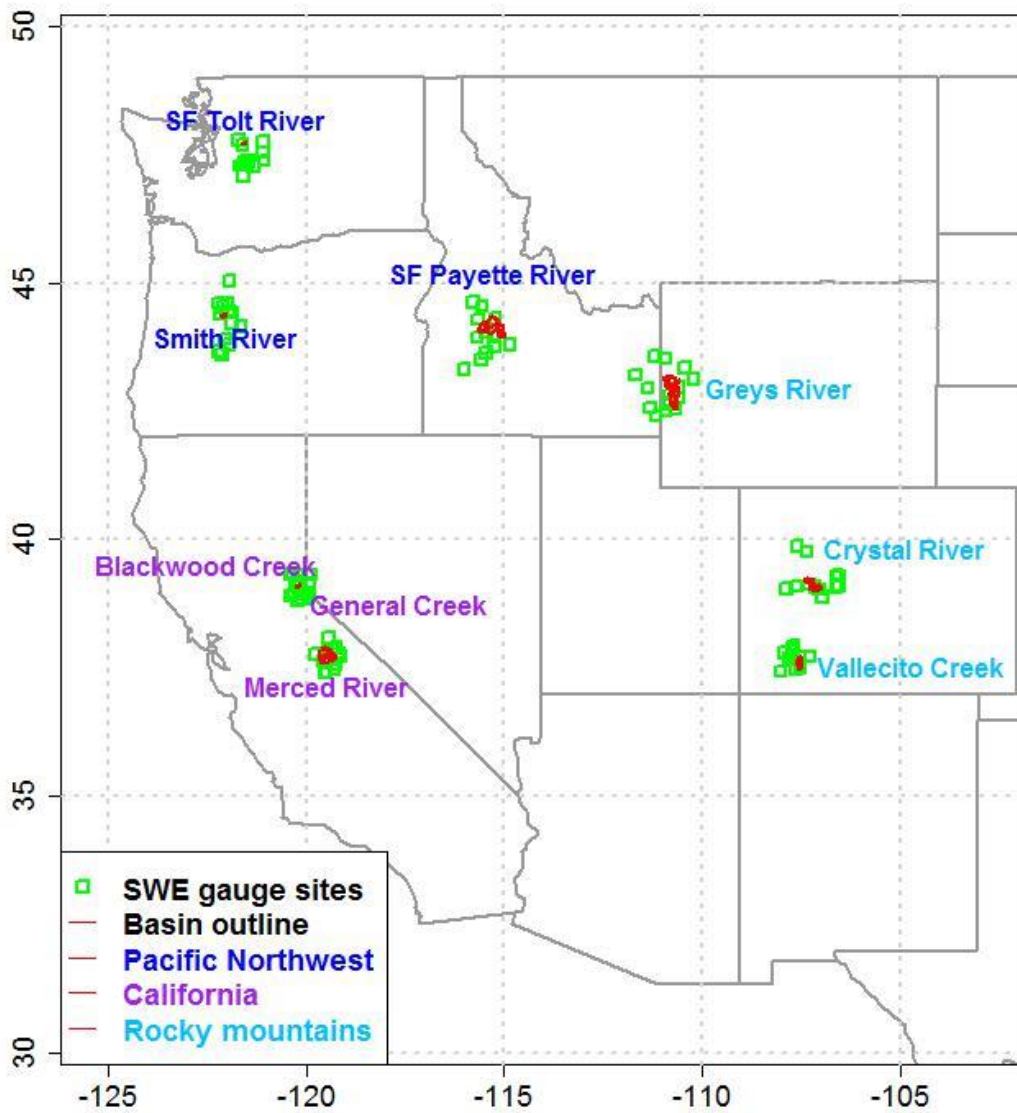
- 
- 501      Research Letters, 35(14).
- 502      Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016. Quantifying Streamflow Forecast Skill  
503      Elasticity to Initial Condition and Climate Prediction Skill. *J. Hydrometeorology*, 17: 651-668, doi:10.1175/JHM-D-14-  
504      0213.1.
- 505      Wu, G., X. Zheng, L. Wang, S. Zhang, X. Liang, and Y. Li, 2012: A new structure for error covariance matrices and their  
506      adaptive estimation in EnKF assimilation. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.2000.

507 **Table 1** Basin features of nine case basins.

| Region | Basin ID | Elevation<br>(m) | Minimum<br>elevation<br>(m) | Maximum<br>elevation<br>(m) | DA date  | Basin area<br>(km <sup>2</sup> ) | Slope<br>(m km <sup>-1</sup> ) | Forest<br>fraction | Basin name       |
|--------|----------|------------------|-----------------------------|-----------------------------|----------|----------------------------------|--------------------------------|--------------------|------------------|
| 14     | 09081600 | 3092.15          | 2050                        | 4250                        | April 1  | 436.88                           | 150.58                         | 0.61               | Crystal River    |
| 14     | 09352900 | 3459.15          | 2450                        | 4250                        | April 1  | 187.74                           | 156.09                         | 0.52               | Vallecito Creek  |
| 17     | 13023000 | 2468.57          | 1750                        | 3450                        | March 1  | 1163.72                          | 98.51                          | 0.68               | 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                         | 0.86               | 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                         | 0.71               | General Creek    |
| 16     | 10336660 | 2188.08          | 1850                        | 2650                        | April 1  | 32.46                            | 83.46                          | 0.79               | Blackwood Creek  |
| 18     | 11266500 | 2576.54          | 1150                        | 3950                        | April 1  | 836.15                           | 140.18                         | 0.67               | Merced River     |

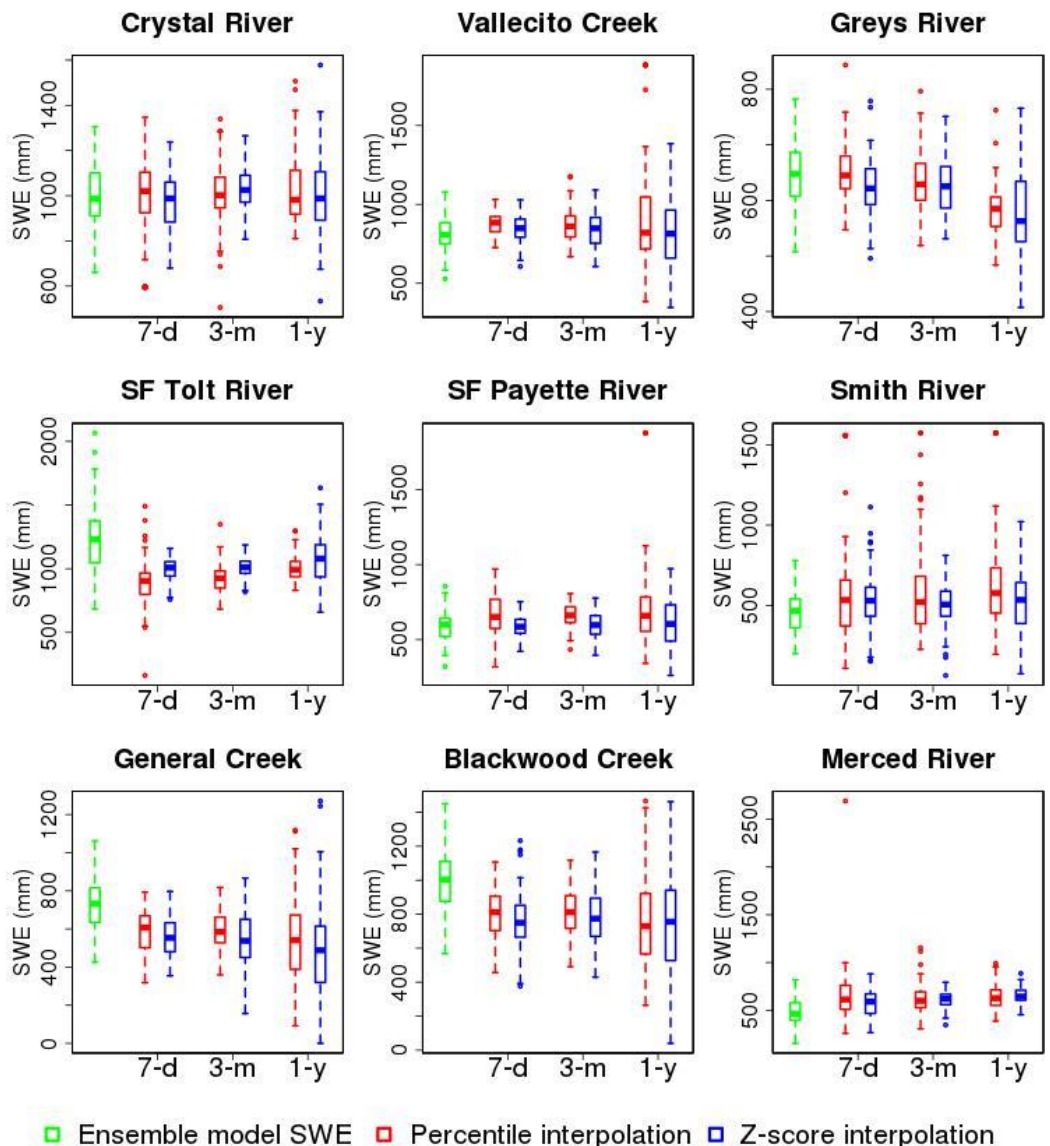
508

### Position of 9 case basins and SWE gauge sites



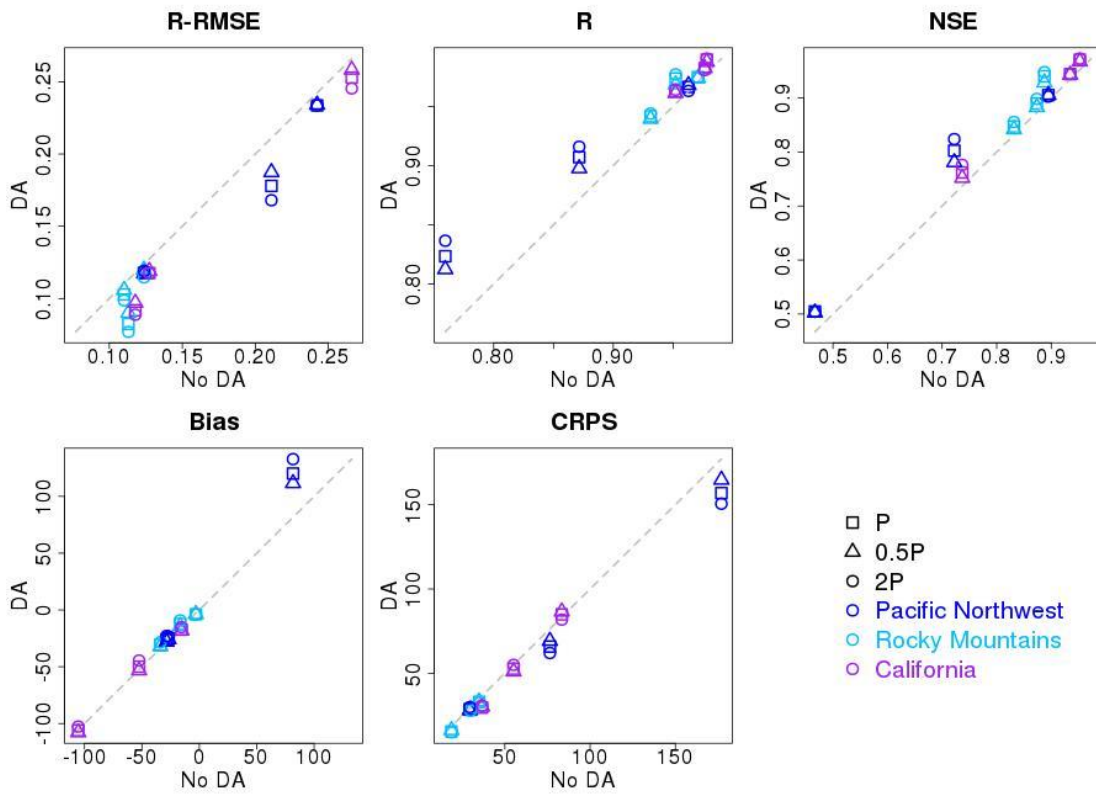
509

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



511  
512  
513  
514

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

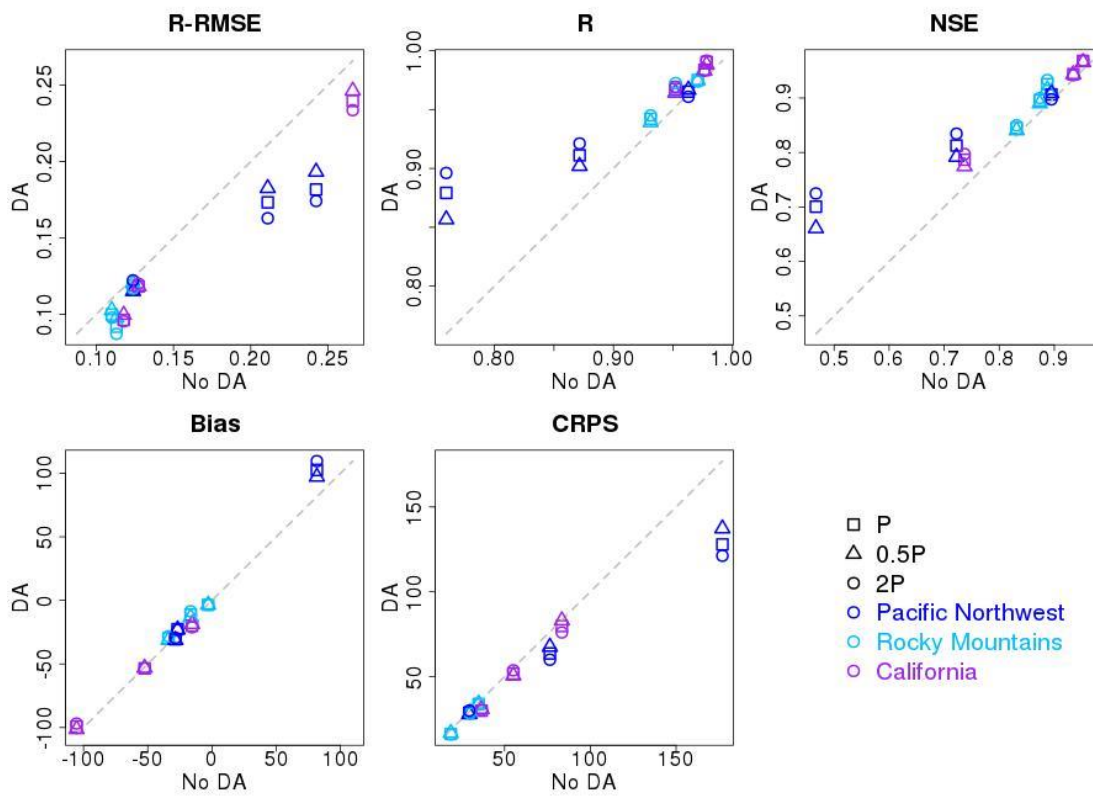


515

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

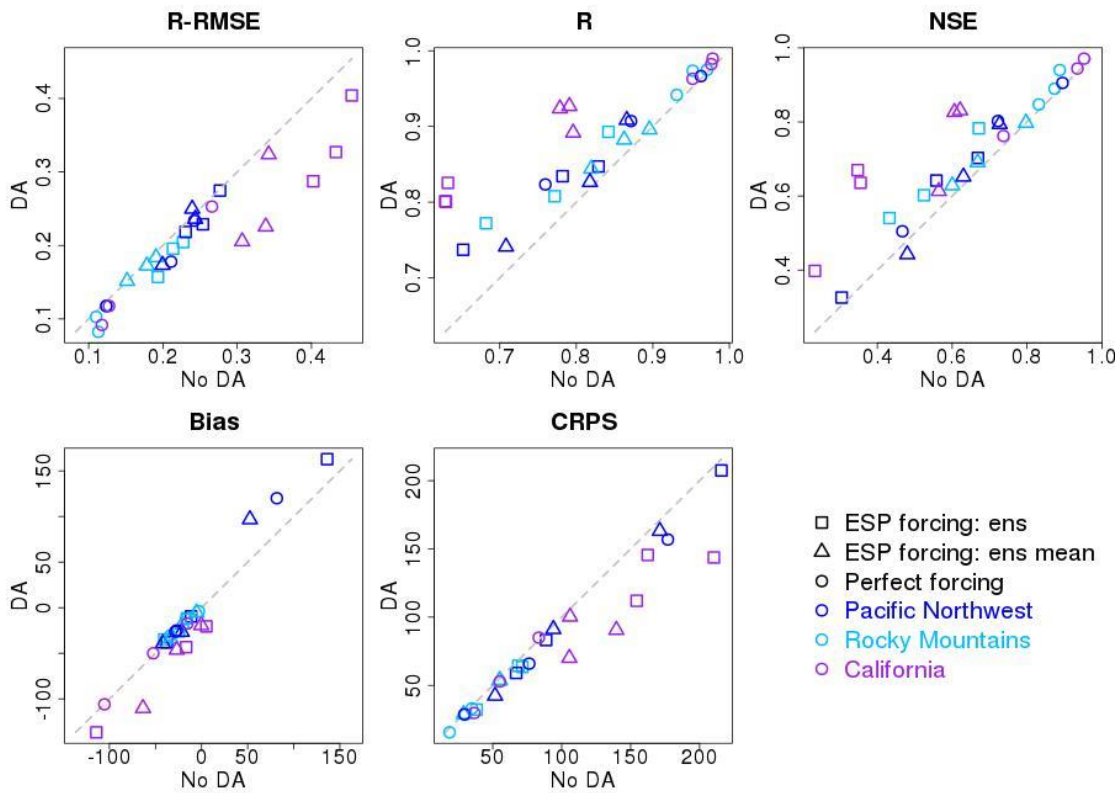
520

521



523

524 **Figure 4. Evaluation metrics for April-July ensemble mean streamflow from the Z-score interpolation for the nine case**  
 525 **basins using perfect forcing. The verification metrics from upper left to lower right are: R-RMSE is the relative**  
 526 **(normalized) root mean squared error, R is the linear (Pearson) correlation coefficient, NSE is the Nash-Sutcliffe**  
 527 **Efficiency, bias is the same as mean error, and CRPS is the continuous ranked probability skill scores.**

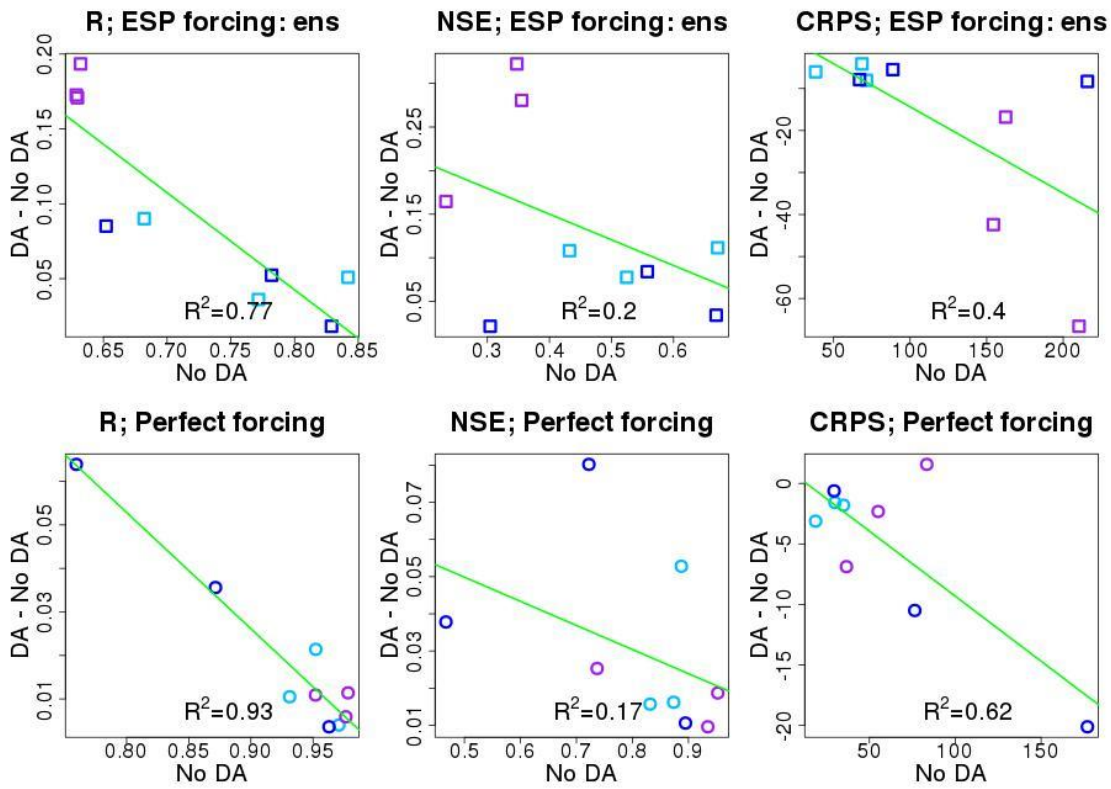


528

529 **Figure 5. Evaluation statistics of percentile interpolation for the nine case basins with the two variations of Ensemble**  
 530 **Streamflow Prediction (ESP) and with perfect forcing data (ens in the legend denotes ensemble). The verification**  
 531 **metrics from upper left to lower right are: R-RMSE is the relative (normalized) root mean squared error, R is the linear**  
 532 **(Pearson) correlation coefficient, NSE is the Nash-Sutcliffe Efficiency, bias is the same as mean error, and CRPS is the**  
 533 **continuous ranked probability skill scores.**

534





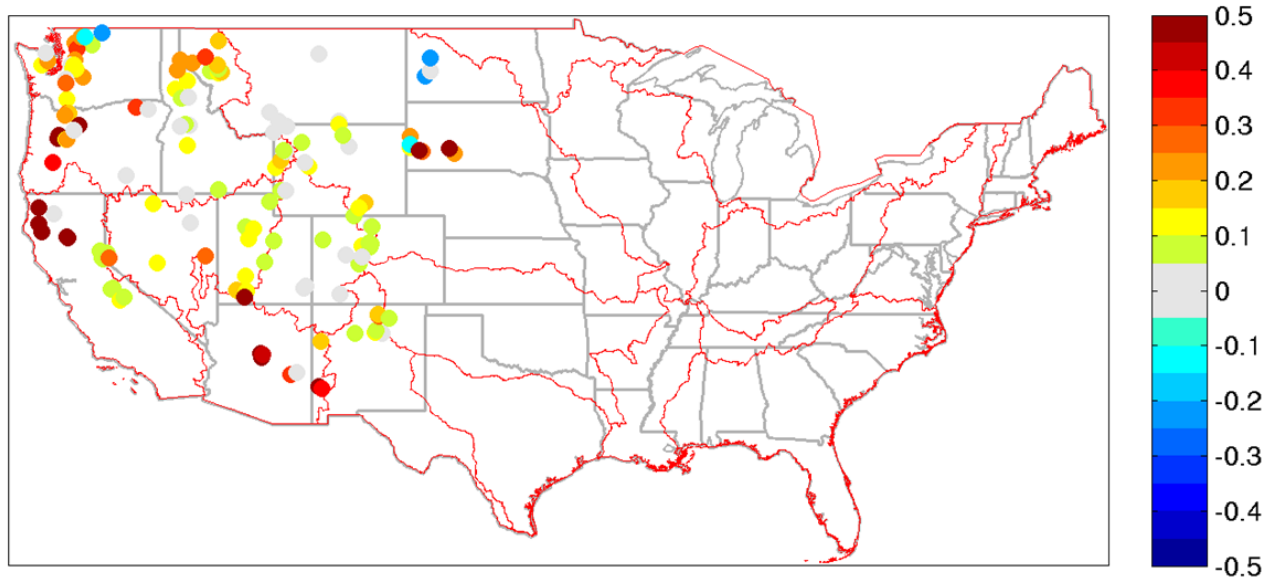
535

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

539

540

### Best Snotel - Model SWE Flow Correlation Difference



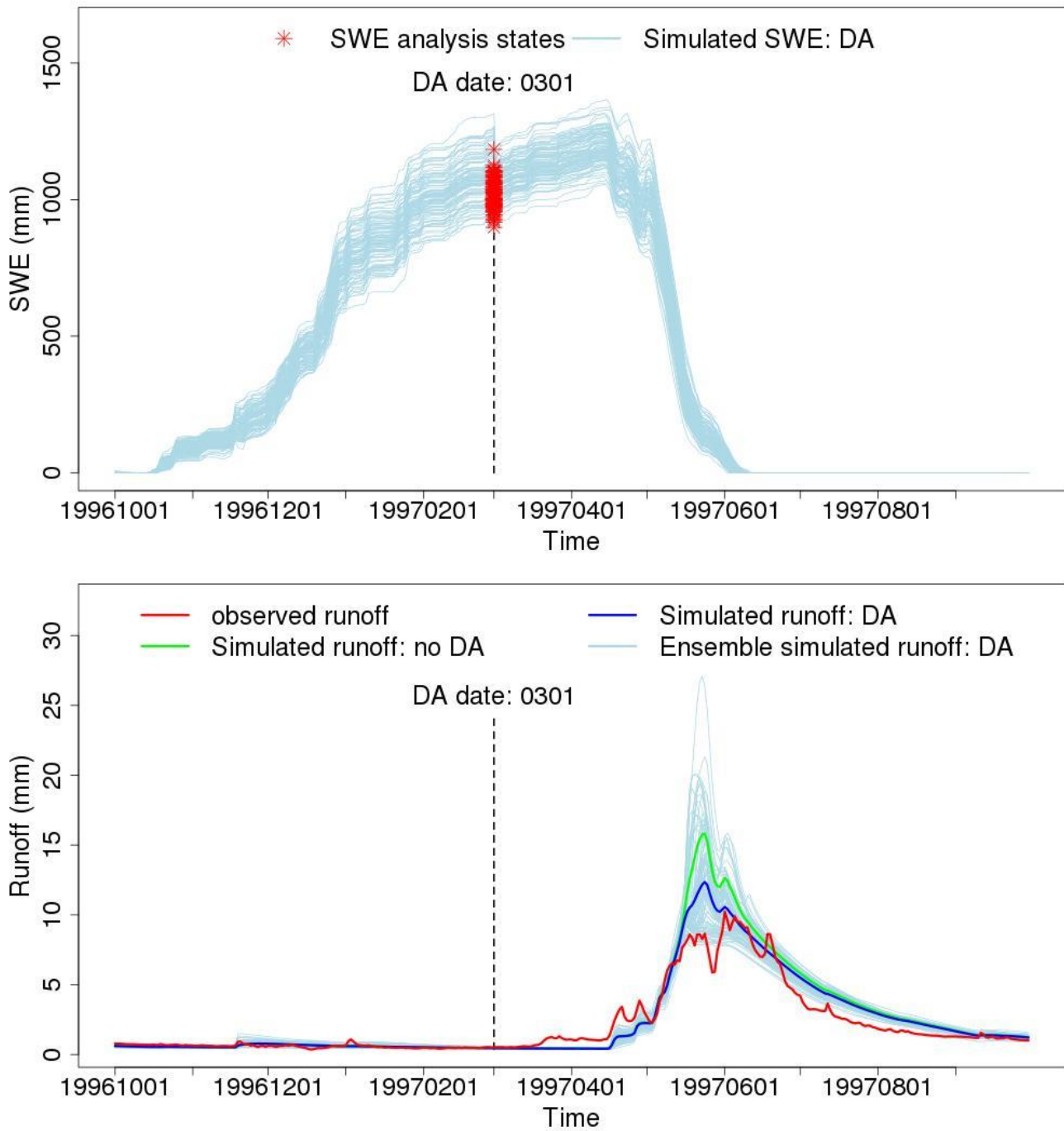
541

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

544

545

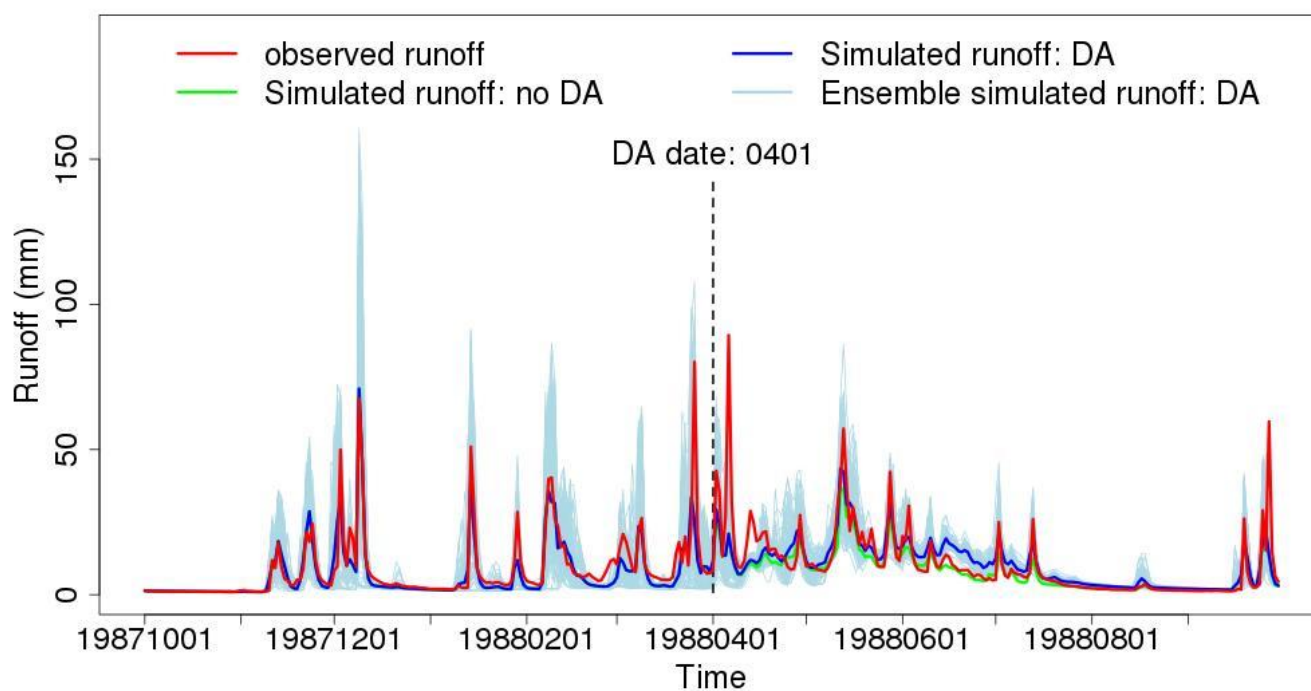
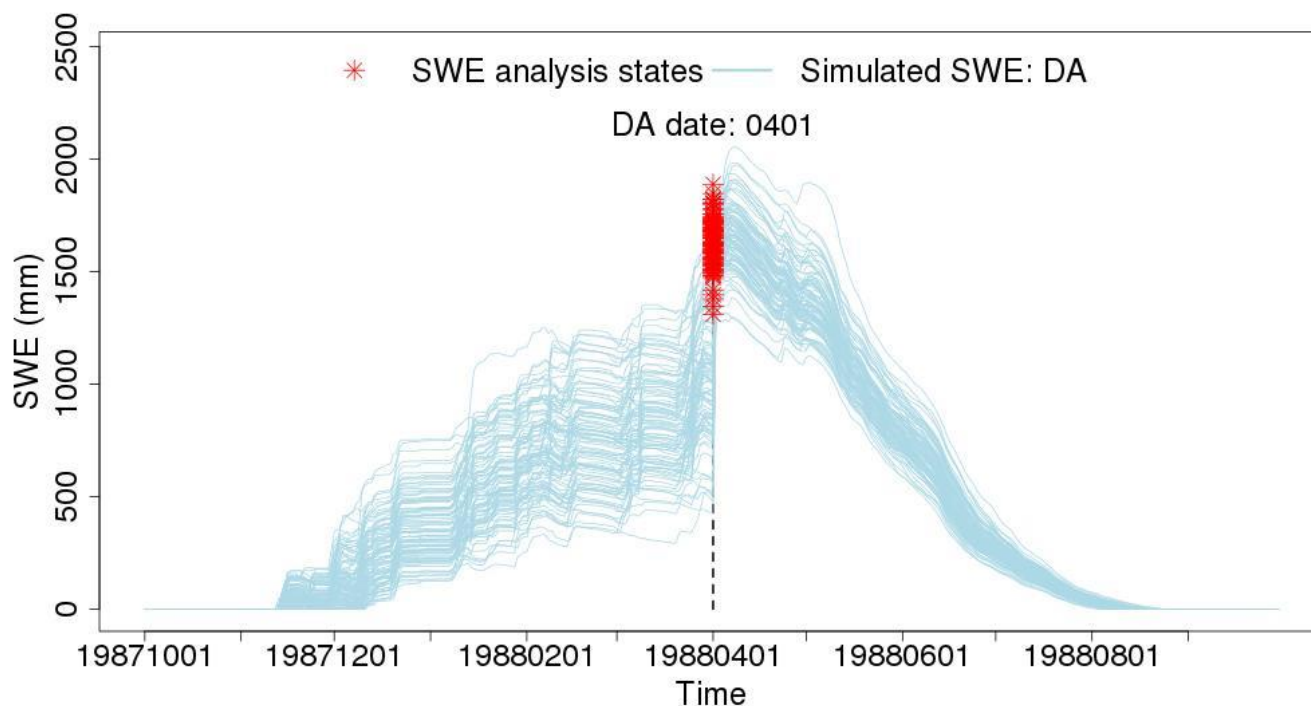
**Region: 17 Basin ID: 13023000 Name: Greys River**



546

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

549

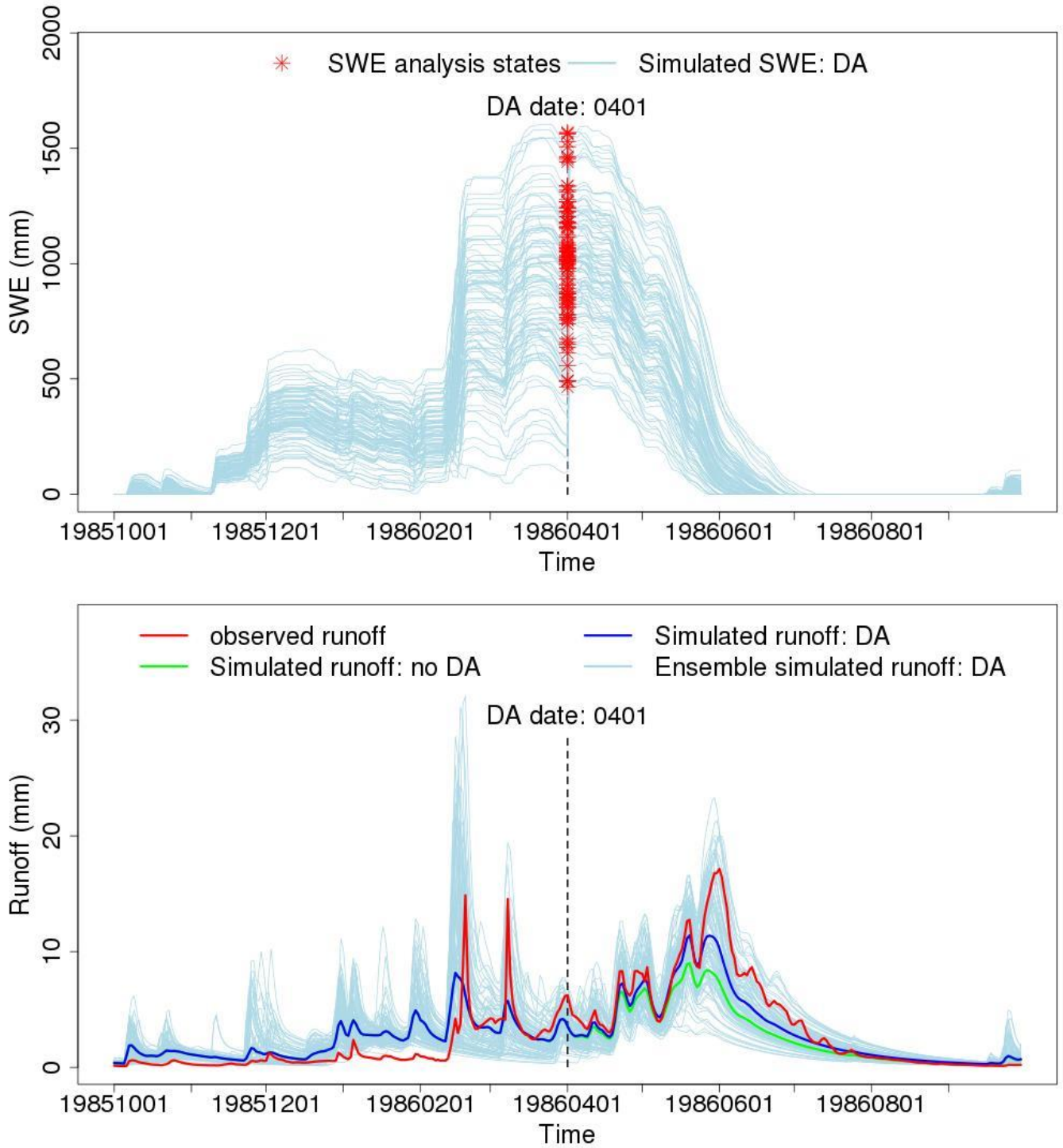
**Region: 17 Basin ID: 12147600 Name: SF Tolt River**

551

552 **Figure 9. Time series plots for runoff and SWE for the South Fork (SF) of the Tolt River for water year 1988. Light**  
553 **blue lines indicate individual ensemble member traces. Vertical black dashed line denotes the data assimilation (DA)**  
554 **date.**

555

**Region: 18 Basin ID: 11266500 Name: Merced River**

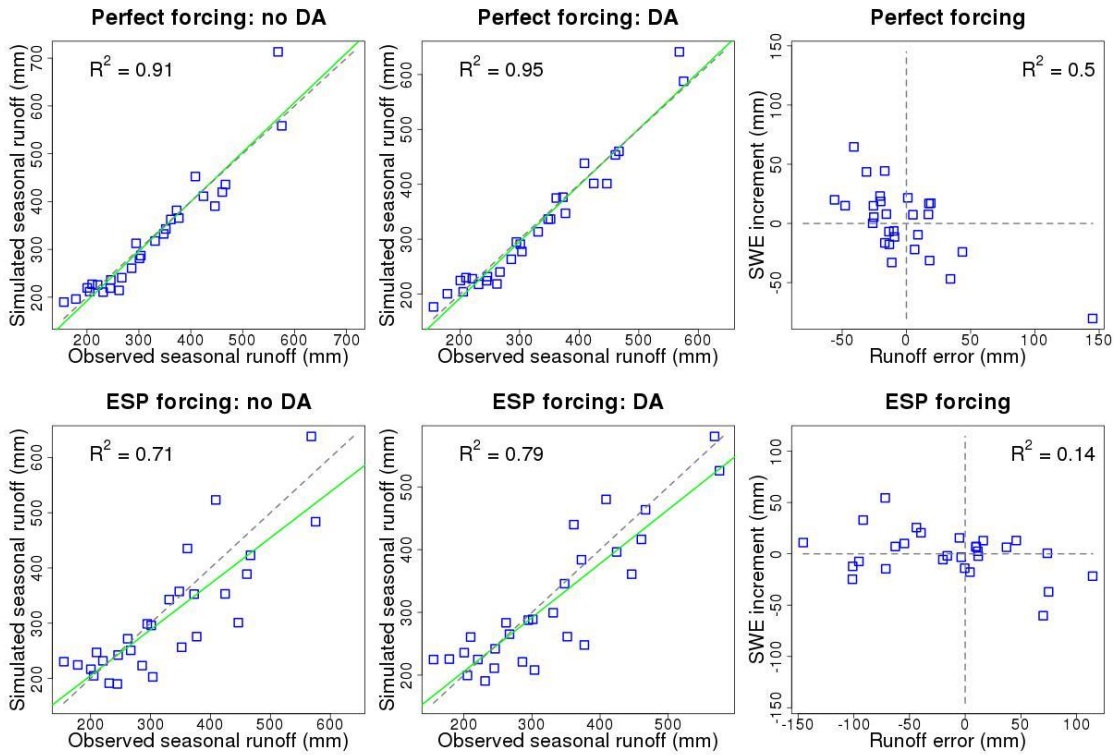


556

557 **Figure 10. Time series plots for runoff and SWE for the Merced River for water year 1986. Light blue lines indicate**

558 **individual ensemble member traces. Vertical black dashed line denotes the data assimilation (DA) date.**

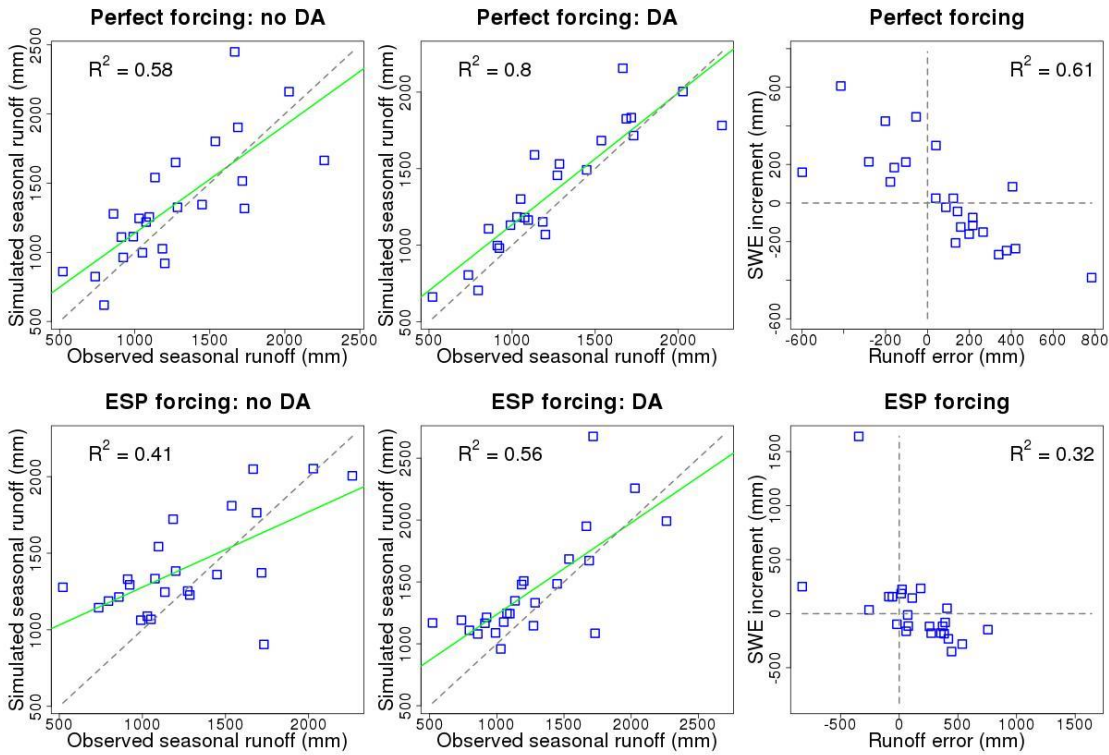
559



560

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

564



565

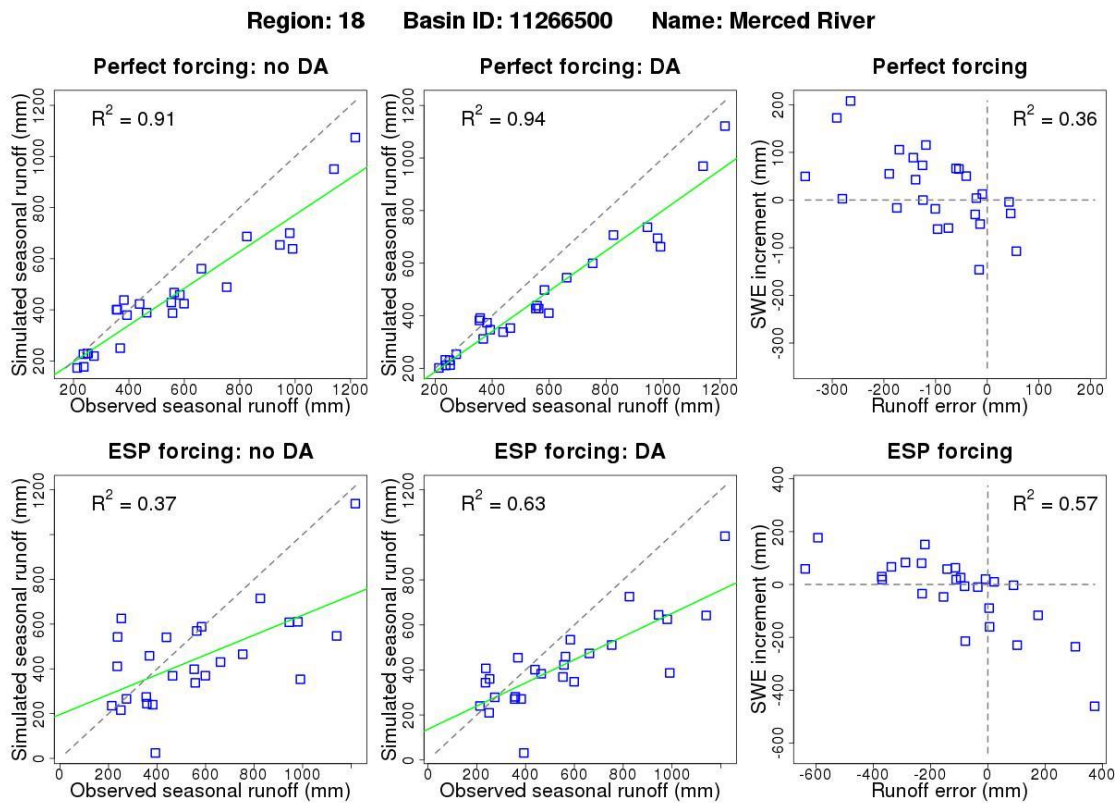
566

567

568

569

**Figure 12. Scatter plots for seasonal runoff and SWE on the data assimilation (DA) date for the South Fork of the Tolt River. Black dashed diagonal lines 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.**



571

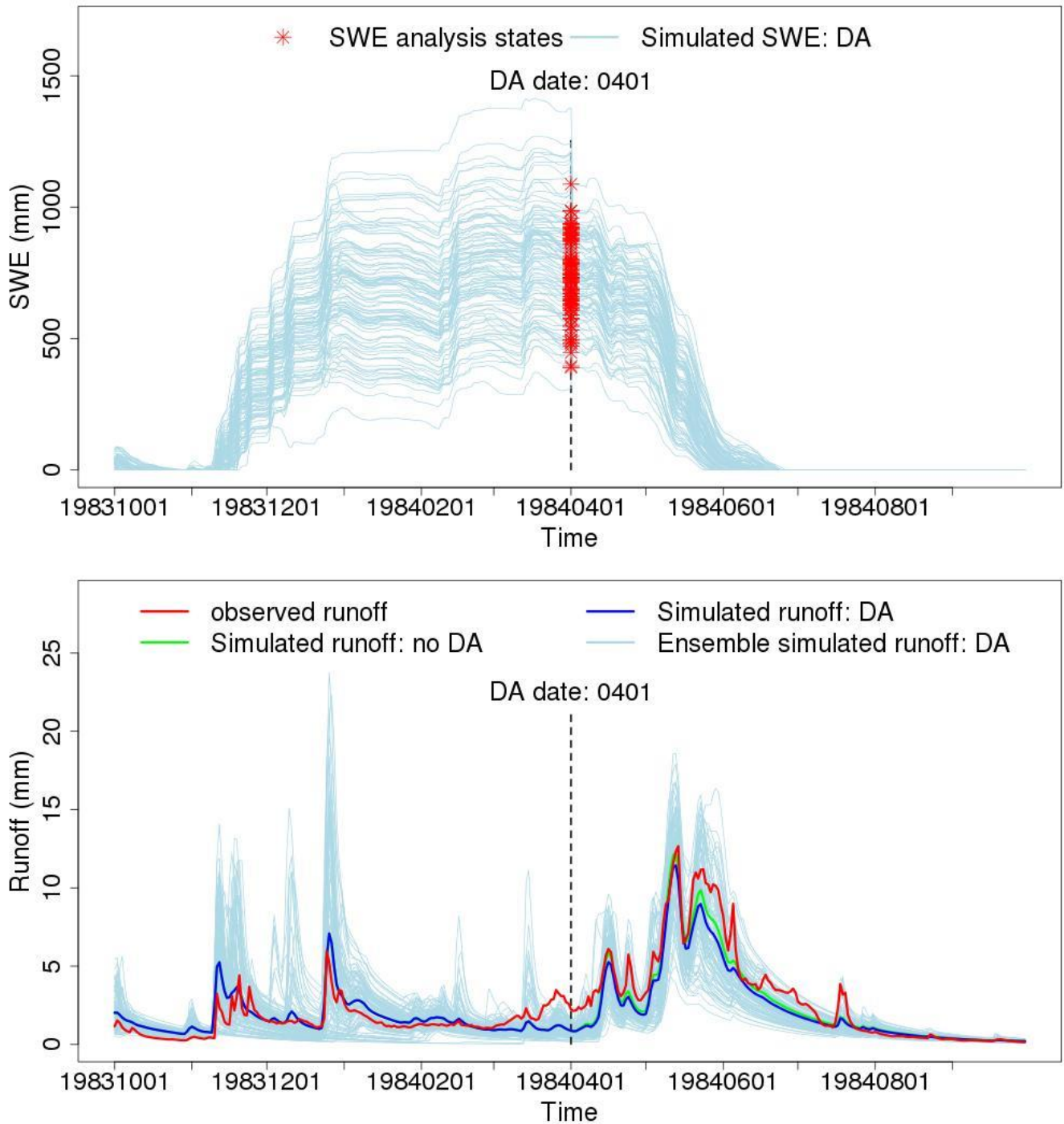
572 **Figure 13. Scatter plots for seasonal runoff and SWE on data assimilation date (DA) for Merced River. Black dashed**  
 573 **diagonal lines are the 1:1 line, while the green lines indicates linear regression fits to data. Perfect forcing results are**  
 574 **shown in the top row, while Ensemble Streamflow Prediction (ESP) results are in the bottom row.**

575

576



**Region: 18 Basin ID: 11266500 Name: Merced River**



577

578 **Figure 14. Time series plots for runoff and SWE for the Merced River for water year 1984. Light blue lines indicate**

579 **individual ensemble member traces. Vertical black dashed line denotes the data assimilation (DA) date.**

580