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]

5   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)

10   **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. model state uncertainty). Finally, we also examine the impact of varying the observation and model state uncertainty on forecast skill. We use basins

15   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 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

20   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 (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; 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. 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

55   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

60   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, the Ensemble Streamflow Forecast (ESP) approach is used to test the impact of SWE DA on subsequent streamflow forecasts.

We apply the EnKF system to nine river basins in the Western US that have a range of basin features and environmental

65   conditions. This is distinct 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., 2011; Moradkhani et al., 2005). 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 which used more arbitrary distributional assumptions. 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 conclusions.

70

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

75   water resource management. Note we do not examine rain driven low-lying basins as 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.

80   Two sources of SWE observations are used in this study: (1) the widely used Snow Telemetry (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.

85

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;

90   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

95   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

100   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 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

105   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. 2015),

we also include a sensitivity analysis to explore the sensitivity of DA results to variations in the estimated observation and

model uncertainty magnitudes.

110   #### 3.2 Generating ensembles of estimated observed watershed SWE

Since the SWE 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 SWE 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-

115   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. The SWE values are transformed into percentiles or Z-score 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

120   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 at the same site within a time-window of +/- n days centered on the date of the observation.

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

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

130   ### 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

135   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 $j = 1,\dots, 100$ represents ensemble member. Finally we use the following equation to transform Z-score to back to SWE values:

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

140   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 the time-window of +/- n days centered on the DA date 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 $j = 1,\dots, 100$.

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

145   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 focusing on assessing variations in methodological aspects of the

150   DA approach, we simplify this configuration by applying 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) :

155    $$x_{i+1}^t = M(x_i^t) + \eta_i, \tag{3}$$

$$y_i^o = h(x_i^t) + \varepsilon_i, \tag{4}$$

where i is the time step, M is the coupled Snow17 and SAC-SMA model system, x is the state variable and y is the observation

variable (in this study both x and 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, η and ε are the model and observation

160   errors respectively, and h is the observation operator that maps the model states to the observation variable. In this study, h is

simply the identity vector as we regard the SWE estimates that have been transformed to model space as observation 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

165   initial states x_0 taken from the retrospective control runs, producing the ensemble forecast states x_i^f. The superscript f

denotes forecast.

2) Calculate the ensemble analysis states:

$$x_i^a = x_i^f + s_i h_i^T (h_i s_i h_i^T + o_i)^{-1} d_i, \tag{5}$$

where superscript a means analysis, o and s are the observed and model simulation error variances (estimated by the variance

170   of ensemble observations and model states respectively) respectively, and the innovation vector (residual) is calculated as:

$$d_i = y_i^o - h_i(x_i^f), \tag{6}$$

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

175   from the retrospective (No DA) run at the start of every WY (October 1), when there is no SWE.

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

180   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

185   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

190   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).

195   **3.6 Verification metrics**

In this study, five frequently used statistics are calculated for evaluating the two DA approaches. The Bias, Correlation

coefficient (R), Relative root mean squared error (R-RMSE), Nash-Sutcliffe efficiency (NSE) are based on the ensemble

averages. And Continuous Ranked Probability Score (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

200   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 respectively. Smaller CRPS means more accurate forecasts.

4 Results

205   **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

210 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 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). Therefore, a 3-month window is recommended for both approaches..

215 The evaluation statistics for ensemble SWE observations estimated by percentile and Z-score interpolation approach are shown in Table 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 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

220 percentile and Z-score interpolation approach exhibit enhanced DA performances among the case basins, indicating that both these approaches are effective in adding observation based information to the model simulations. The percentile interpolation and Z-score interpolation methods vary in performance across the basins with both performing better in some basins and not others. 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

225 observations more and improves the DA performance. This indicates that the model error variance tends to be underestimated or our observation uncertainties tend to be overestimated.

The evaluation statistics of Table 2 and 3 are also presented as scatter plots in Figures 3 and 4 respectively. The metrics R-RMSE, R and NSE indicate that both approaches bring enhancements to the ensemble mean streamflow in most basins, although the DA does not help correct forecast biases. Post-processing procedures (e.g. bias correction) could be used to further

230 enhance the system performance, but is not a focus of this study. These figures also show that No DA forecasts with relatively better performance, mostly due to better simulations of forecast initial conditions, benefit less from DA.

Figure 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

235 CRPS. Again, higher skill No DA watersheds saw smaller DA improvements. The DA evaluation metric improvement increment versus the corresponding No DA evaluation metric score for the case basins are shown in Figure 6. It can be seen

that the DA improvements in all evaluation metrics have a general weak negative correlation with No DA performance, which again highlights that better simulated basins benefit less from SWE DA.

4.2 Case study analyses

240   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 7, 8 and 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

245   improves daily runoff forecasts in addition to seasonal volume forecast improvements, although this is not true in every watershed (e.g. Tolt River).

Figures 10, 11 and 12 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 No DA and DA simulated seasonal runoff vs observed runoff for perfect (top row) and ESP ensemble forcing

250   (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 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

255   overestimated), we 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 2nd (upper left) and 4th (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.

260

5 Discussion and Conclusions

SWE observations generally contain 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 is a challenge that is often met by empirical solutions. Two approaches of constructing SWE ensemble

265   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). 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

270   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.

Sensitivity analyses of model uncertainty impacts on DA performance suggest that either the forcing-alone based estimation

275   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. 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.

280   In this study, a 3-month window of SWE observations generally gives the best performance. However, in some basins a different window length may bring larger improvements. Generally, longer windows mean that more SWE values are used for transformation, and the transformation tends to be more statistically representative of the long-term model-observation climatology. However, shorter time windows mean 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

285   represent the relationship between model and observation climatologies. The window length must be a balance between these two considerations.

The ESP-based assessment of 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 this has been a long-standing challenge in operational forecasting. We show at least minor improvement in seasonal runoff forecasts in all nine basins. A notable finding is also that

290   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.

We chose 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

295   US. As a result, this study did not shed light on how to address additional challenges that may be associated with using SWE DA 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 SWE DA has been implemented in distributed models in limited experimental contexts (e.g., Wood and Lettenmaier, 2006), a systematic examination of EnKF

DA in spatially distributed hydrological models, coupled with a thoughtful accounting for model parameter and structural

300   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:

305   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 observations are available at: www.wcc.nrcs.usda.gov/snow/ while the California SWE observations are available at: cdec.water.ca.gov/snow.

Acknowledgements

310   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

315   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

320   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.

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.

325   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.

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

330   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.

335   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

340   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.

345   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

350   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.

Harrison, B., Bales, R., 2015. Skill Assessment of Water Supply Outlooks in the Colorado River Basin. Hydrology, 2(3): 112-131.

355   He, M., Hogue, T., Margulis, S., Franz, K., 2011. 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

360   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.

365   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
370   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.

375   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-
380   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.

385   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:
390   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, AW 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)	DA date	Basin area (km²)	Slope	Forest percent	Basin name
14	09081600	3092.15	April 1	436.88	150.58	0.6136	Crystal River
14	09352900	3459.15	April 1	187.74	156.09	0.5199	Vallecito Creek
17	13023000	2468.57	March 1	1163.72	98.51	0.6753	Greys River
17	12147600	998.25	April 1	16.07	159.37	1	SF Tolt River
17	13235000	2077.16	April 1	1158.47	126.25	0.8604	SF Payette River
17	14158790	1210.48	March 15	40.76	116.44	1	Smith River
16	10336645	2180.92	April 1	20.09	118.27	0.7136	General Creek
16	10336660	2188.08	April 1	32.46	83.46	0.7908	Blackwood Creek
18	11266500	2576.54	April 1	836.15	140.18	0.6741	Merced River

410

Table 2 Evaluation statistics for percentile interpolation before and after DA.

Basin name	Forecast error variance (P)	R-RMSE	R	NSE	Bias	CRPS
Crystal River	No DA	0.124	0.931	0.832	-16.64	34.92
	P	0.118	0.942	0.847	-12.34	33.12
	1/2·P	0.12	0.94	0.842	-14.29	33.39
	2·P	0.115	0.944	0.856	-9.21	32.7
Vallecito River	No DA	0.11	0.971	0.874	-33.89	29.95
	P	0.103	0.975	0.89	-30.69	28.37
	1/2·P	0.106	0.974	0.883	-32.11	28.87
	2·P	0.099	0.975	0.898	-28.34	27.85
Greys River	No DA	0.113	0.952	0.887	-2.98	18.67
	P	0.082	0.974	0.94	-3.91	15.55
	1/2·P	0.09	0.969	0.929	-3.8	16.3
	2·P	0.077	0.977	0.948	-3.83	15.06
SF Tolt River	No DA	0.242	0.76	0.467	81.64	176.98
	P	0.234	0.823	0.505	120.04	156.87
	1/2·P	0.234	0.812	0.503	111.55	164.73
	2·P	0.234	0.837	0.504	132.76	150.56
SF Payette River	No DA	0.124	0.963	0.895	-26.77	29.43
	P	0.118	0.967	0.905	-24.91	28.81
	1/2·P	0.117	0.968	0.906	-25.88	28.07
	2·P	0.119	0.963	0.903	-23.41	30.02
Smith River	No DA	0.211	0.872	0.722	-28.5	76.42
	P	0.178	0.907	0.803	-26.25	65.92
	1/2·P	0.187	0.898	0.781	-28.04	69.24
	2·P	0.168	0.916	0.824	-22.91	62.07
General Creek	No DA	0.118	0.978	0.952	-15.54	36.64
	P	0.092	0.99	0.971	-17.36	29.76
	1/2·P	0.097	0.988	0.968	-18.19	30.29
	2·P	0.089	0.99	0.973	-15.22	30.86
Blackwood Creek	No DA	0.127	0.976	0.935	-52.44	55.12
	P	0.118	0.982	0.944	-49.95	52.8
	1/2·P	0.119	0.982	0.943	-53.23	51.11
	2·P	0.117	0.98	0.945	-44.53	55.04
Merced River	No DA	0.266	0.952	0.737	-105.4	83.41
	P	0.253	0.963	0.762	-105.95	84.98
	1/2·P	0.258	0.961	0.752	-107.76	86.62
	2·P	0.245	0.964	0.776	-102.65	81.87

Table 3 Evaluation statistics for Z-score interpolation before and after DA.

Basin name	Forecast error variance (P)	R-RMSE	R	NSE	Bias	CRPS
Crystal River	No DA	0.124	0.931	0.832	-16.64	34.92
	P	0.119	0.942	0.845	-11.22	33.94
	1/2·P	0.12	0.939	0.841	-13.4	34.05
	2·P	0.117	0.945	0.85	-8.57	33.56
Vallecito River	No DA	0.11	0.971	0.874	-33.89	29.95
	P	0.1	0.975	0.896	-29.84	28.7
	1/2·P	0.103	0.975	0.89	-31.15	29.06
	2·P	0.098	0.974	0.9	-28.27	28.19
Greys River	No DA	0.113	0.952	0.887	-2.98	18.67
	P	0.092	0.969	0.926	-3.5	16.22
	1/2·P	0.098	0.965	0.916	-3.41	16.8
	2·P	0.087	0.972	0.933	-3.48	16.02
SF Tolt River	No DA	0.242	0.76	0.467	81.64	176.98
	P	0.182	0.879	0.7	102.55	127.9
	1/2·P	0.193	0.857	0.661	97.22	137.26
	2·P	0.174	0.896	0.725	109.53	121.26
SF Payette River	No DA	0.124	0.963	0.895	-26.77	29.43
	P	0.117	0.965	0.907	-22.69	28.69
	1/2·P	0.115	0.967	0.909	-22.99	27.89
	2·P	0.122	0.961	0.897	-23.36	30
Smith River	No DA	0.211	0.872	0.722	-28.5	76.42
	P	0.173	0.911	0.813	-30.55	63.69
	1/2·P	0.183	0.902	0.792	-31.22	67.56
	2·P	0.163	0.921	0.835	-29.45	59.99
General Creek	No DA	0.118	0.978	0.952	-15.54	36.64
	P	0.096	0.99	0.968	-19.96	29.97
	1/2·P	0.1	0.988	0.966	-18.65	30.95
	2·P	0.096	0.992	0.969	-21.17	30.4
Blackwood Creek	No DA	0.127	0.976	0.935	-52.44	55.12
	P	0.118	0.984	0.944	-53.46	51.82
	1/2·P	0.118	0.983	0.944	-53.17	50.62
	2·P	0.12	0.983	0.942	-53.52	53.74
Merced River	No DA	0.266	0.952	0.737	-105.4	83.41
	P	0.24	0.967	0.787	-99.26	79.81
	1/2·P	0.246	0.964	0.775	-101.37	83.11
	2·P	0.234	0.969	0.797	-96.91	76.03

415

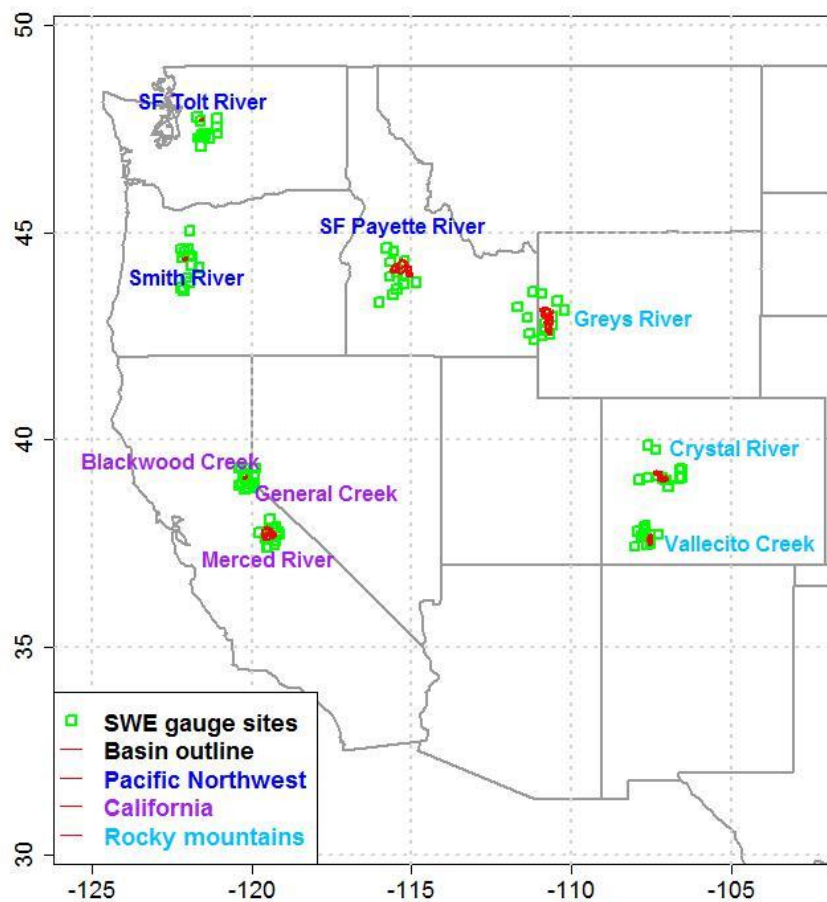



Figure 1. Location of nine case basins in the Western US and SWE gauge sites.

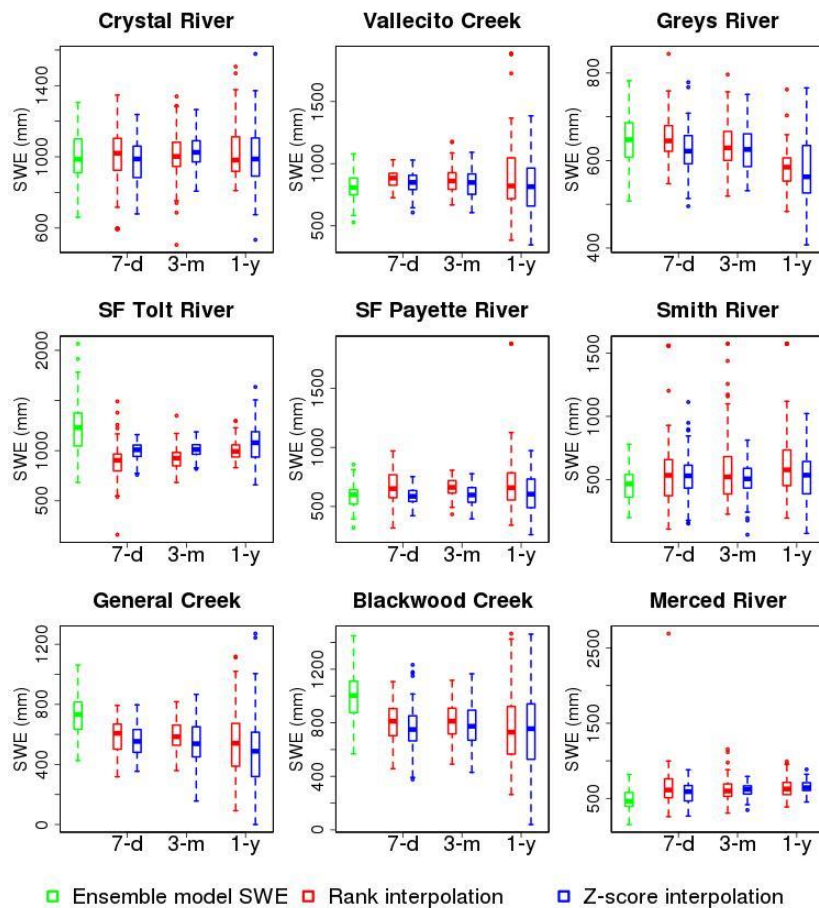



Figure 2. Boxplots of ensemble model SWE and estimated ensemble SWE observations for the nine case basins on the DA date in 2004, for three window lengths – 7 days, 3 months, and 1 year.

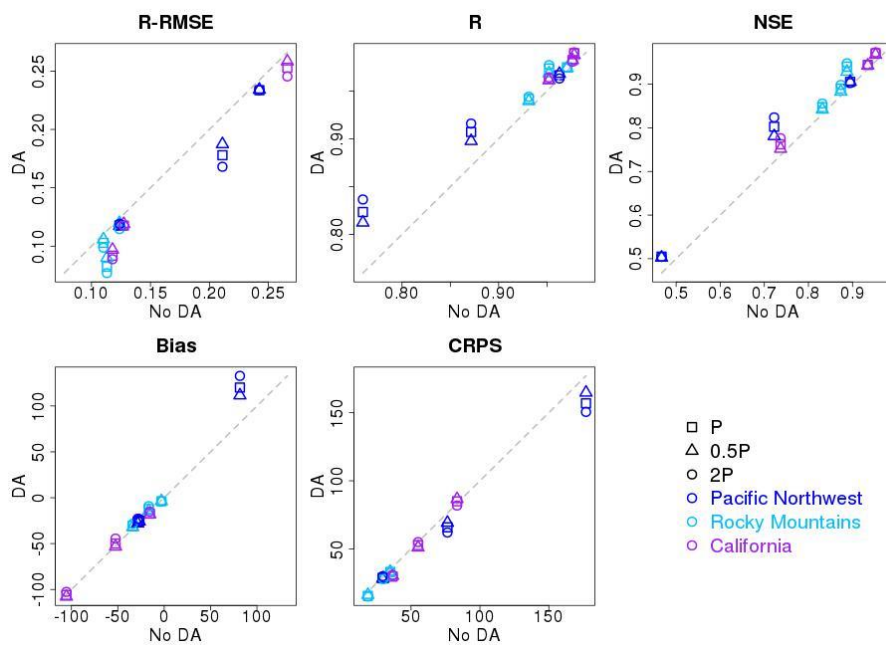



Figure 3. Evaluation metrics for April-July ensemble mean streamflow from the percentile-based interpolation method

425 **for the nine case basins.**

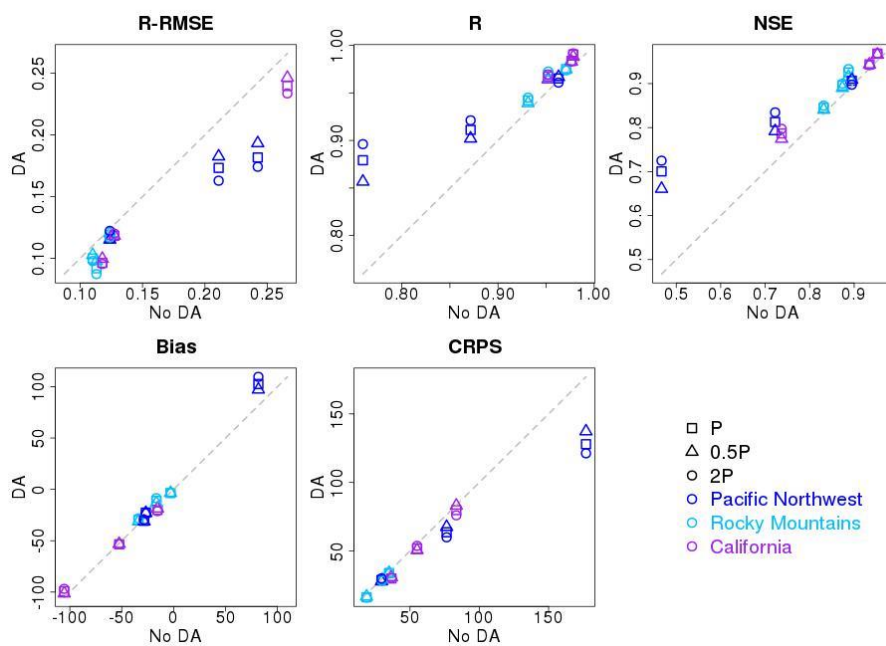



Figure 4. Evaluation metrics for April-July ensemble mean streamflow from the *Z*-score interpolation for the nine case

430   **basins.**

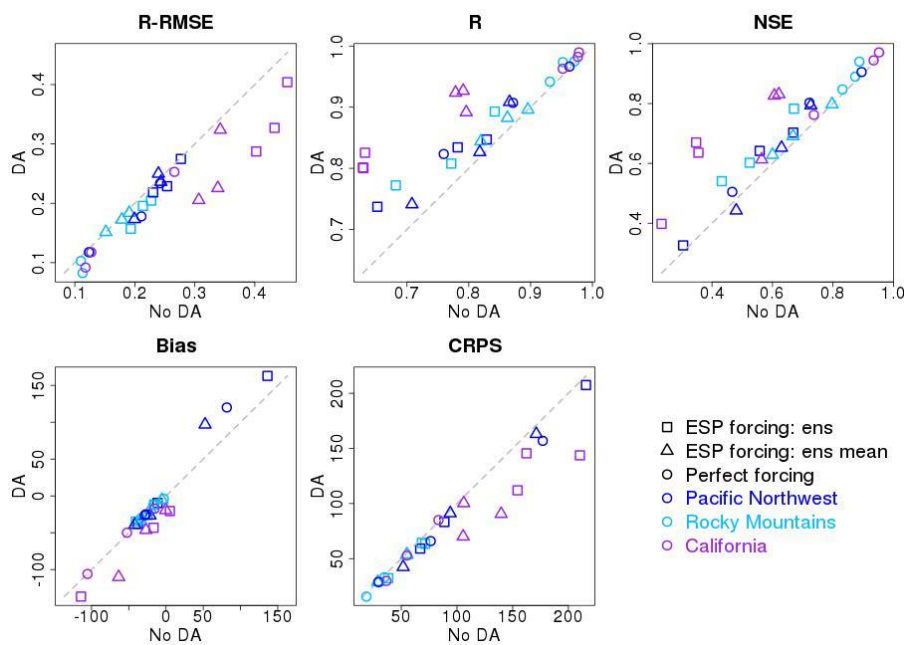
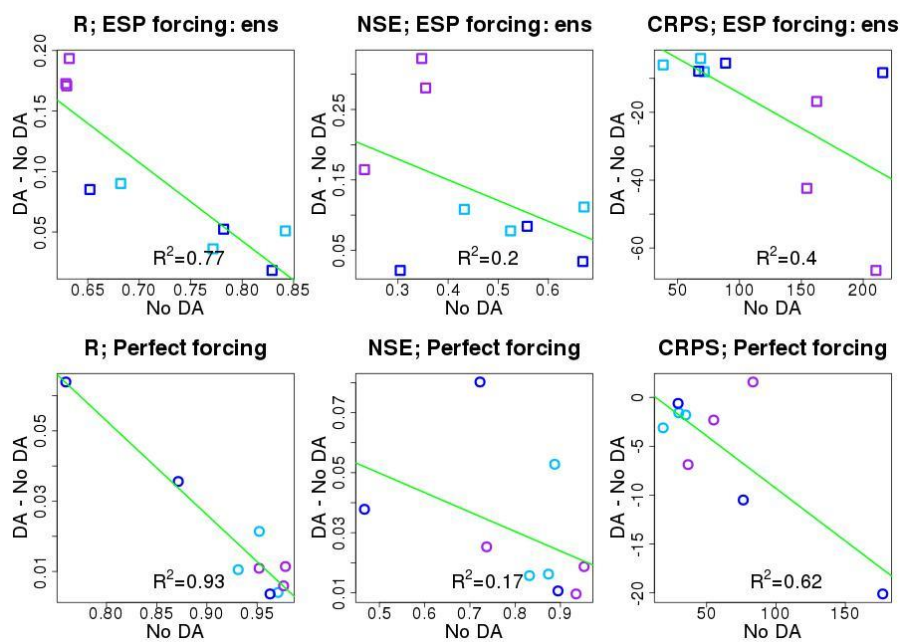
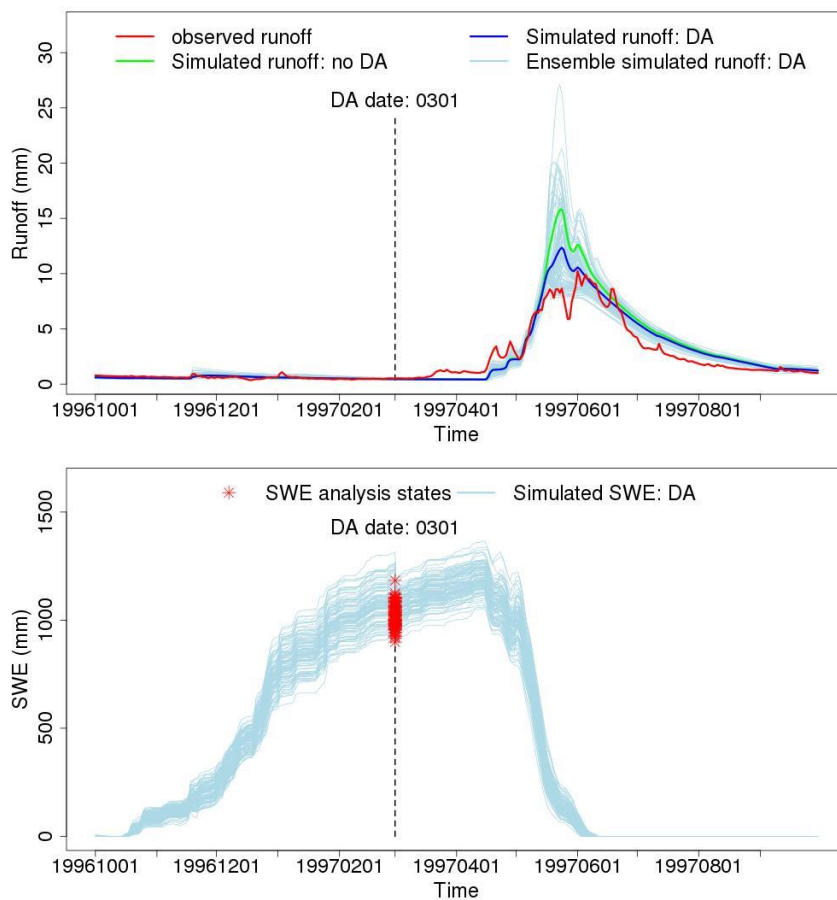
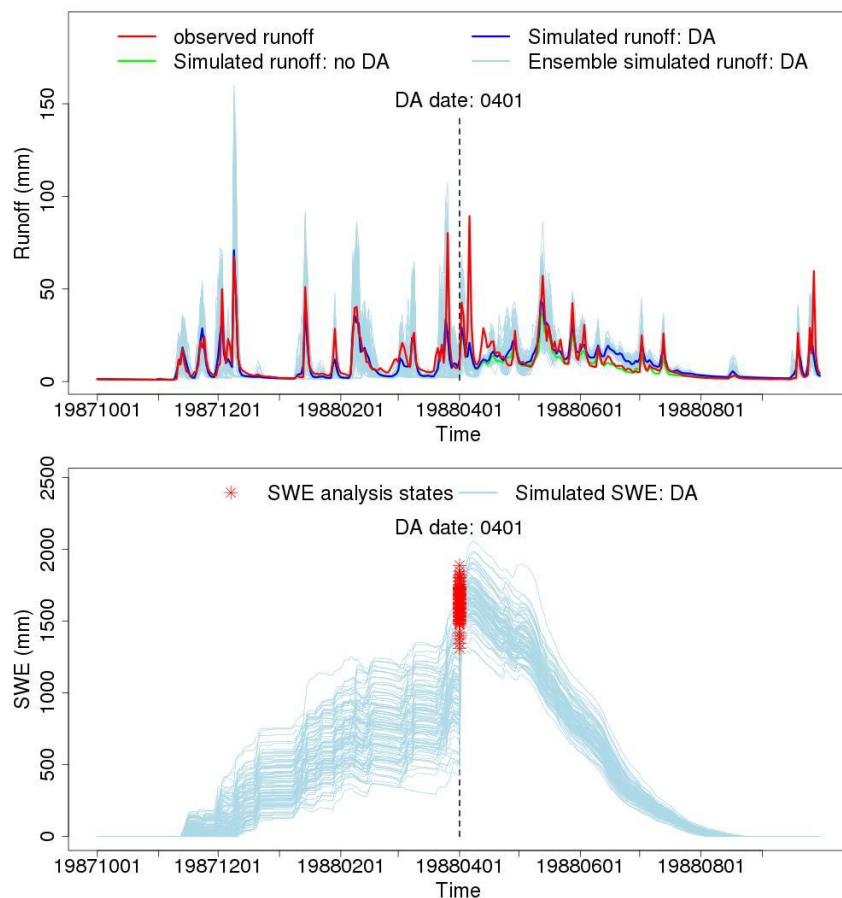



Figure 5. Evaluation statistics of percentile interpolation for the nine case basins with the two variations on ESP and with perfect forcing data (ens in the legend denotes ensemble).

435

440   **Figure 6. Incremental change in evaluation statistics for ESP and perfect forcing forecasts using percentile-based interpolation for the nine case basins.**

445 **Figure 7. Time series plots for runoff and SWE for Greys River. Light blue lines indicate individual ensemble member**

traces. Vertical black dashed line denotes data assimilation date.

450   **Figure 8. Time series plots for runoff and SWE for SF Tolt River following Figure 7.**

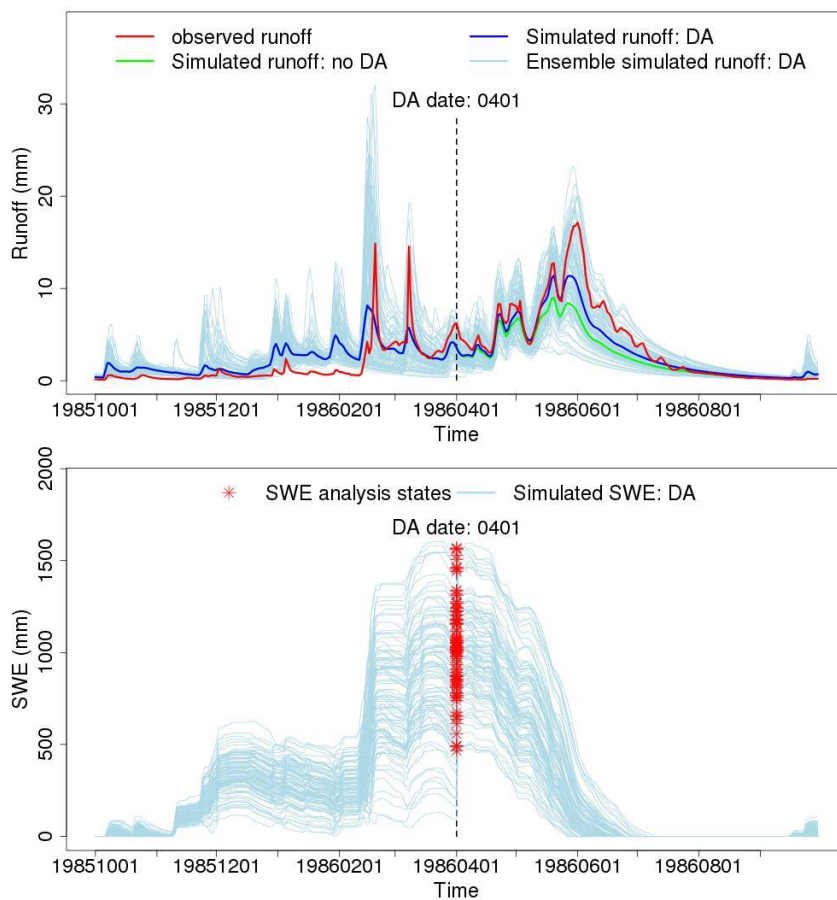



Figure 9. Time series plots for runoff and SWE for Merced River following Figure 7.

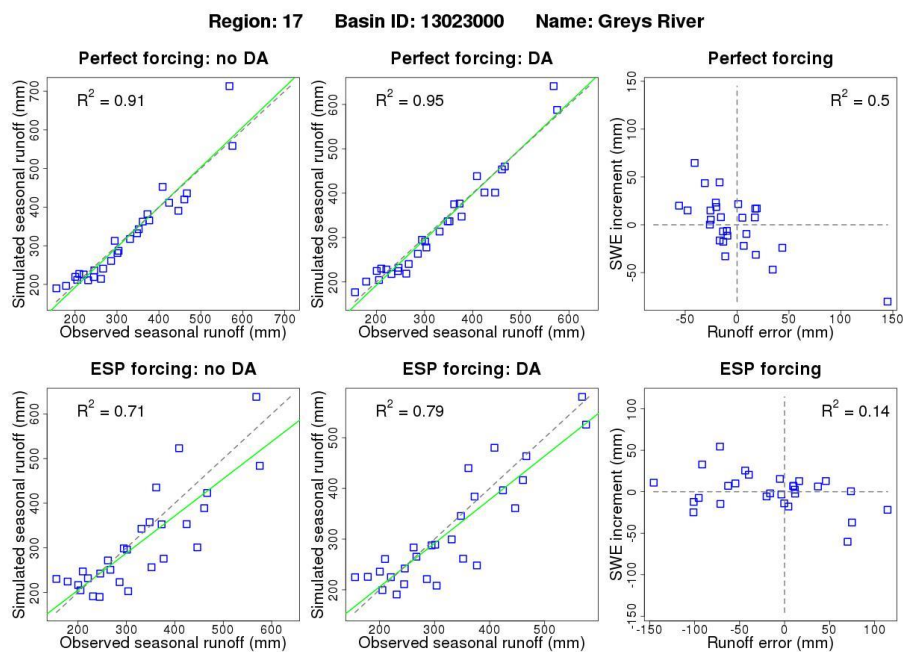



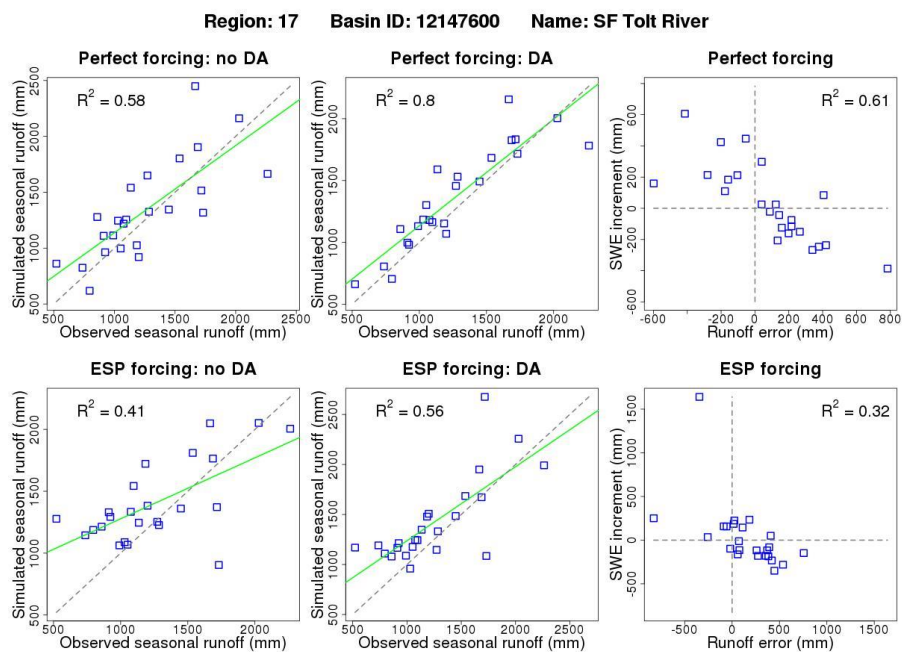
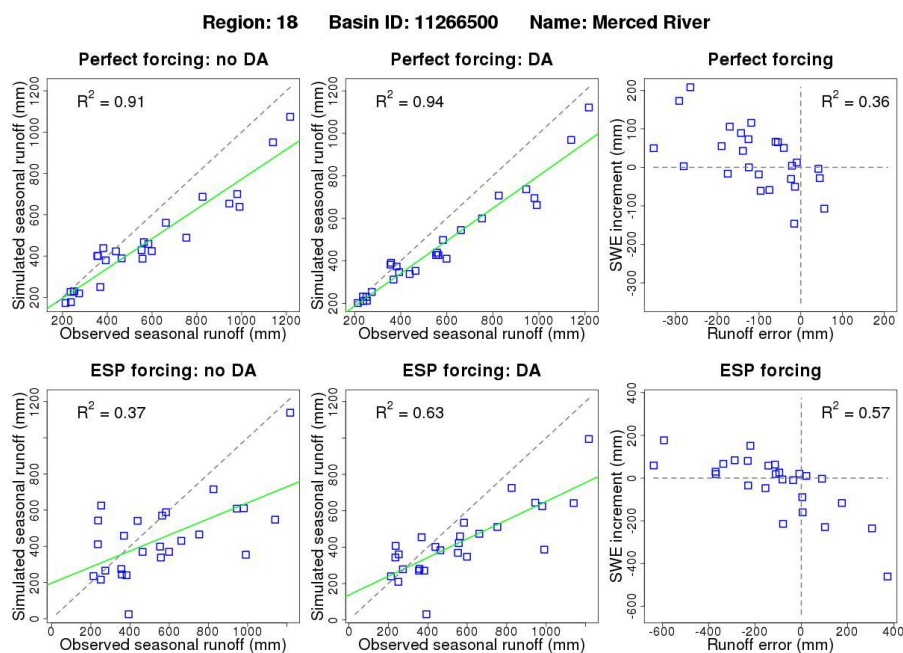
Figure 10. Scatter plots for seasonal runoff and SWE on DA date in the DA years for Greys River. Black dashed diagonal lines is the 1:1 line, while the green lines indicate linear regression fits to data.

460

Figure 11. Scatter plots for seasonal runoff and SWE on DA date in the DA years for SF Tolt River following Figure 10.

465 **Figure 12. Scatter plots for seasonal runoff and SWE on DA date in the DA years for Merced River following Figure 10.**