Hydrol. Earth Syst. Sci., 20, 3277–3287, 2016
www.hydrol-earth-syst-sci.net/20/3277/2016/
doi:10.5194/hess-20-3277-2016

Hydrology and Earth System Sciences
Open Access
EGU

ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction

Joost V. L. Beckers[1], Albrecht H. Weerts[1,2], Erik Tijdeman[3], and Edwin Welles[4]

[1]Deltares, Delft, the Netherlands
[2]Department of Environmental Sciences, Wageningen University, Wageningen, the Netherlands
[3]Department of Hydrology, University of Freiburg, Freiburg, Germany
[4]Deltares USA Inc, Silver Spring, Maryland, USA

Correspondence to: Joost V. L. Beckers (joost.beckers@deltares.nl)

Received: 12 February 2016 – Published in Hydrol. Earth Syst. Sci. Discuss.: 19 February 2016
Revised: 13 May 2016 – Accepted: 5 July 2016 – Published: 12 August 2016

Abstract. Oceanic–atmospheric climate modes, such as El Niño–Southern Oscillation (ENSO), are known to affect the local streamflow regime in many rivers around the world. A new method is proposed to incorporate climate mode information into the well-known ensemble streamflow prediction (ESP) method for seasonal forecasting. The ESP is conditioned on an ENSO index in two steps. First, a number of original historical ESP traces are selected based on similarity between the index value in the historical year and the index value at the time of forecast. In the second step, additional ensemble traces are generated by a stochastic ENSO-conditioned weather resampler. These resampled traces compensate for the reduction of ensemble size in the first step and prevent degradation of skill at forecasting stations that are less affected by ENSO. The skill of the ENSO-conditioned ESP is evaluated over 50 years of seasonal hindcasts of streamflows at three test stations in the Columbia River basin in the US Pacific Northwest. An improvement in forecast skill of 5 to 10 % is found for two test stations. The streamflows at the third station are less affected by ENSO and no change in forecast skill is found here.

1 Introduction

The ensemble streamflow prediction (ESP) forecasting method is a common way to produce seasonal outlooks of river volumes. It is used by River Forecasting Centers of the National Weather Service (NWS-RFC) and other US agencies (Druce, 2001; Pica, 1997; McEnery et al., 2005). The ESP uses historical time series of mean areal precipitation (MAP) and mean areal temperature (MAT) and considers these as representative of the local climate (Twedt et al., 1977; Day, 1985). The historical MAP and MAT series are used as meteorological forcings to a hydrologic model to generate an ensemble of streamflow forecasts. The number of ensemble traces is equal to the number of historical years because every trace corresponds to a particular historical year. The initial model state is the current state of the watershed of interest, which is obtained from an update run with data assimilation of recent gauge data. Depending on the type of watershed and the time of year, the initial conditions can affect the streamflows for several months ahead (Wood and Lettenmaier, 2008; Li et al., 2009; Shukla and Lettenmaier, 2011; Yossef et al., 2013). This gives the ESP predictive ability over a climatological forecast, i.e. a distribution of historical streamflows (Franz et al., 2003).

Despite the great improvements in general circulation model (GCM)-based seasonal forecasting over the past decades (Leung et al., 1999; Hamlet and Lettenmaier, 1999; Wood et al., 2002, 2005; Clark and Hay, 2004; Wood and Lettenmaier, 2006; Yuan et al., 2015), the ESP method is still the current practice at most NWS-RFC. One of the reasons for this is that ESP uses the same type of meteorological input, i.e. historical MAP and MAT, as is typically used for calibration of the hydrologic models (Pica, 1997). GCM input typically needs to be downscaled and bias-corrected before it can be applied to hydrological modelling at the sub-basin scale. A second reason is that the ESP allows for a sampling of non-meteorological variables, such as water demand, from

the same historical years as the meteorological inputs. The fact that all variables are taken from the same historical year automatically preserves any cross-correlation between them, which is important for water resources planning.

In the original ESP, the historical MAP and MAT series represent the average climate; that is, every historical year is treated as an equally likely future scenario. In many regions, however, the local climate is known to be teleconnected to inter-annual to decadal fluctuations in oceanic–atmospheric circulation patterns, such as the El Niño–Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO) (Ropelewski and Halpert, 1986, 1996; Kiladis and Diaz, 1989; Halpert and Ropelewski, 1992; Diaz and Markgraf, 2000; McCabe and Dettinger, 2002). These fluctuations, or climate modes, affect the streamflow regime in US rivers (Redmond and Koch, 1991; Kahya and Dracup, 1993; Dracup and Kahya, 1994; Piechota and Dracup, 1996; Piechota et al., 1997; Mantua et al., 1997; Beebee and Manga, 2004; Tootle et al., 2005, 2007; Tootle and Piechota, 2006; Abudu et al., 2010; Lü et al., 2011; Gedalof et al., 2012; Sagarika et al., 2015).

The phase of most climate modes is quantified by climate indices that are evaluated and published monthly. Taking this information into account in streamflow forecasting could enhance its skill. Several methods have thus been developed to incorporate climate index information into the ESP. They can be classified into pre- and post-processing schemes (Werner et al., 2004; Kang et al., 2010). In the pre-processing approach, the MAP and MAT ESP inputs are modified to match the predicted climate anomalies (Perica, 1998). Hay et al. (2009) applied a climate-mode-dependent adjustment of hydrologic model parameters. Another pre-processing alternative is to generate synthetic input time series by random resampling of monthly MAP and MAT from historical years that have similar climate index values (Werner et al., 2004). Although some improvement of forecast skill was reported, Werner et al. (2004) concluded that these pre-adjustment techniques are computationally cumbersome and less suited for operational usage than post-processing techniques. Kang et al. (2010) also found the post-processing schemes more effective than pre-processing schemes in a Korean case study.

In the post-processing approach, the ESP output, i.e. the ensemble of hydrographs, is transformed to incorporate climate mode information. One technique is to weight the ensemble traces according to the similarity between climate indices in the historical year and the year of forecast (Croley II, 1996, 1997; Stedinger and Kim, 2010; Madadgar et al., 2012; Najafi et al., 2012; Bradley et al., 2015). Instead of a weighting scheme, Hamlet and Lettenmaier (1999) used a selection of ESP traces according to a classification of historical years based on ENSO and PDO climate indices. Although their results showed an improved specificity of the ensemble forecast, the classification leads to a reduction of ensemble traces, because the number of historical years in each class is obviously less than the original number of ensemble traces. A reduction of ensemble size generally leads to a degradation of the statistical properties of the ensemble forecast and to a reduction of forecast skill (Richardson, 2001; Ferro, 2007).

Although less obvious, this problem also arises in other ensemble post-processing schemes. The effective ensemble size is reduced by applying weights to ensemble traces. To be effective, the information that is added to the ensemble by the weighting should be in balance with the reduction of the forecast uncertainty (Weijs and van der Giesen, 2013). However, to obtain a coherent forecast for a large watershed, the forecasting must be done using a single set of weights for all forecasting stations, although the influence of the climate modes may differ per station. A weighting scheme that produces good results for stations that are influenced by a particular climate mode may not perform well for stations that are less affected by this climate mode. The forecast skill for these latter stations may be compromised by the weighting scheme. This problem has been underexposed in previous studies. Najafi et al. (2012) mentioned the loss of forecast skill for smaller ensemble size and used a modified skill score to remove the effect (Weigel et al., 2007). This conceals the negative effect that a weighting scheme could have on quantile estimates for stations that are less affected by climate modes.

In this study, an ESP conditioning method on climate mode information is described that produces a gain in forecast skill at stations that are affected by climate modes, while avoiding a loss of skill at other stations. The method is a combination of pre- and post-processing. The post-processing involves a selection of traces from the original ESP. In a pre-processing step, a number of new ensemble traces are generated by a monthly weather resampler. The newly generated traces augment the ensemble up to the original number of traces and all ensemble traces are weighted equally. This preserves the statistical properties of the ESP ensemble and avoids loss of forecast skill due to reduction of (effective) ensemble size.

The method is explained in detail in Sect. 2. The study region and the data used are described in Sect. 3. Section 4 includes the results obtained applying the method to the study area and a forecast skill assessment relative to the standard ESP. Section 5 discusses the results.

2 Method

The proposed method consists of two parts: a *subsampler*, which selects ensemble traces from the original ESP, and a *resampler*, which generates additional ensemble traces.

2.1 Subsampler procedure

The subsampler procedure is a k-nearest neighbour (k-NN) type scheme, similar to the schemes used by Werner et al. (2004) and Najafi et al. (2012). The selection is based

on similarity between the climate index value at the time of forecast and the value on the same day of a historical year. The selection can be based on a single climate index or on multiple indices. In the case of multiple indices the similarity criterion is the Euclidian distance in (multi-)index phase space. Weights can be applied to each index dimension to represent the relative importance of each index. The choice of indices and their optimal weights will depend on the region of interest. A correlation analysis of climate index versus historical streamflows is a straightforward way to find the strongest teleconnections.

The number of ESP traces to be selected by the subsampler needs to be optimized. By selecting fewer traces, the forecast becomes more specific, as only the historical years most similar to the present year are included in the forecast. However, there is a trade-off between specificity and sampling error. With fewer years, the resolution of the ensemble decreases and the sampling error increases. This reduction of skill can be overcome by adding more ensemble traces as is done in this study by using a resampler.

2.2 Resampler procedure

The resampler generates new ensemble traces to augment the dismissed traces in the subsampler scheme. The new traces are generated by a monthly weather resampler that is loosely based on a method for daily rainfall resampling developed by Brandsma and Buishand (1998). The resampler generates synthetic time series of precipitation and temperature by sampling from the historical record. Instead of using full historical years, as in the standard ESP, individual months from different historical years are sampled and assembled into new meteorological time series. The selection of historical months is conditioned on similarity between climate indices. A monthly resampling period is chosen to preserve the within-month temporal correlations and because most climate indices are also defined on a monthly timescale. It is assumed that the resampled time series are realistic representations of future weather patterns and that they are equally likely to occur as the full historical years in the original ESP.

The resampling procedure is as follows.

1. To initiate the sampling, the reference date is set to the time of forecast.

2. A historical year is selected by probability sampling, where the probability of selecting year y is a function of the weighted Euclidian distance between the climate index values on the reference date $m_{i,r}$ and on the same day of a historical year $m_{i,y}$. A Gaussian-type distribution is adopted for this probability:

$$P_\mathrm{y} = \frac{1}{N}\exp\left(-\sum_i w_i\left(m_{i,y} - m_{i,r}\right)^2\right), \quad (1)$$

where w_i is a factor that represents the importance of climate index i. N is a normalization factor so the sum of all P_y equals one.

3. From the selected historical year y, a month of climate indices and MAP and MAT values is added to the newly generated time series.

4. A new reference date is set by advancing 1 month and replacing the year by the selected historical year. For example, if the first reference date was 1 January 2016 and the selected historical year is 1997, the new reference date will be 1 February 1997. Subsequently, we proceed with the next resampling round and search for a historical year that is similar to the new reference date (step 2).

When going through the selection procedure, the same historical year can be selected several times in consecutive resampling rounds. The year of the reference date even has the highest probability of being reselected because it has the greatest similarity to the reference climate index. However, other historical years also have a non-zero probability of being selected. Therefore, the resampled time series typically consist of resampled months from several historical years. The resampling procedure can be repeated with different random seeds to generate an ensemble of synthetic weather time series.

The weights w_i in Eq. (1) can have any positive value (also larger than 1). Their values determine not only the relative importance of the climate indices i but also the stringency of the similarity criterion. The probability of selecting a historical year with a similar climate index becomes larger for large w_i. This increases the persistence of the climate phase signal and its effect on the streamflow forecast. For small values of w_i, historical months that have quite different climate indices will be selected. Consequently, the climate phase signal is lost after a few resampling rounds.

A stringent similarity criterion will lead to the same historical years being selected every time. This will produce many similar or even identical traces that resemble full historical years. In order for the ensemble to accurately describe the uncertainty distribution, more variation in the ensemble traces is needed, which is achieved by setting a less stringent similarity criterion. The choice for an appropriate similarity criterion is thus a trade-off between conservation of the climate phase signal and generating sufficient variation in the ensemble traces.

The weights w_i for each index need to be tuned to produce the required persistence of the climate signal and variation of ensemble traces at the relevant forecast lead times. Criteria that can be used for persistence are for example the difference between climate indices in consecutive months and the autocorrelation function. By adjusting w_i and comparing the autocorrelation and month-to-month differences for the resampled time series, the optimal value is determined.

Table 1. Forecasting stations and sub-basin properties. The mean flows, precipitation and runoff ratios are based on observations from 1949 to 2003.

Station	River	Drainage area (km^2)	Mean elevation (m)	Mean flow ($m^3\ s^{-1}$)	Mean annual precipitation (mm)	Runoff ratio
Libby	Kootenay	23 270	811	310	851	0.49
Hungry Horse	Flathead	4145	239	100	1174	0.63
Dworshak	Clearwater	6320	363	160	1283	0.62

3 Example application

3.1 Study area

As a case study, the method was applied to seasonal streamflow forecasting at three forecasting stations (dams) on Columbia River tributaries in the US Pacific Northwest (PNW), listed in Table 1. The watersheds are located in the Cascade Range (see Fig. 1), where runoff is dominated by snowmelt. The typical annual pattern displays a build-up of snowpack in winter and snowmelt and runoff in spring. Figure 2 shows the median and variation of the monthly streamflows for the three stations. The flows are highest and have the most variation in the snowmelt season (May–June).

One of the forecasting centres that use ESP for seasonal streamflow forecasting is Bonneville Power Administration (BPA). BPA is a self-financing federal agency based in Portland, Oregon, that markets the hydroelectric power from 31 dams in the Columbia River basin (Bonneville Power Administration et al., 2001). The dams are operated following often competing needs and legal constraints, including hydropower production, supply of irrigation water, support of aquatic life and keeping the risk of undesirable peak flows and flooding at a minimum. Seasonal streamflow forecasting plays an important role in the dam operation planning and hydropower marketing. The high stakes on the energy market make even the smallest possible improvement in forecast skill worth pursuing.

BPA uses an operational forecasting system based on NWS-CHPS (Community Hydrologic Prediction System, Gijsbers et al., 2009) with ESP functionality for their seasonal streamflow outlooks (4- to 8-month lead time). The Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al., 1973; Burnash, 1995) and SNOW-17 snow accumulation and ablation model (Anderson, 1976) are used for simulating and forecasting the hydrologic processes per sub-basin at a 6 h time step, taking mean areal precipitation (MAP) and mean areal temperature (MAT) per subarea as inputs. The conceptual sub-basin models were calibrated on 55 years (1949–2003) of observational data. Initial (warm) states for the ESP forecasts are generated by running the models in operational mode, continuously blending in recent snowpack and streamflow gauge data into model states.

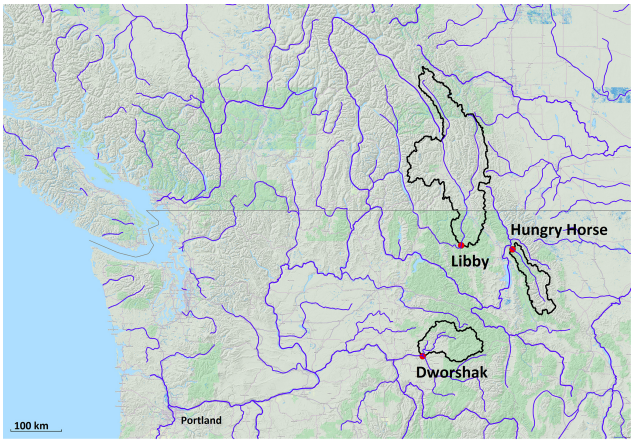

Figure 1. Study area with the three test stations and extent of subbasins. Portland is at 45.5231° N, 122.6765° W.

The PNW climate is teleconnected with ENSO (Philander, 1990). The warm phase of ENSO (El Niño) is associated with warm and dry winters, whereas the cold phase (La Niña) has the opposite effect with colder and wetter than average winters (Ropelewski and Halpert, 1986; Redmond and Koch, 1991). Other climate phenomena have also been shown to influence the climate in the PNW (Lau and Sheu, 1988; Knight et al., 2006). The different climate modes may amplify or counteract each other, but each is considered to contain unique information that might have additional value for the streamflow predictions. The influences of these climate phenomena make the PNW an interesting case study for the climate-conditioned ESP.

Historical weather time series for the three sub-basins (6 hourly MAP and MAT) covering a period from 1949 to 2003 were provided by BPA. Historical values for a range of indices describing various climate modes were obtained from NOAA-CPC (http://www.cpc.ncep.noaa.gov/data/indices/).

3.2 Experimental setup and parameter calibration

Several climate mode indices and combinations of indices for ensemble trace selection and conditioning of the subsampler were evaluated, including the Pacific Decadal Oscilla-

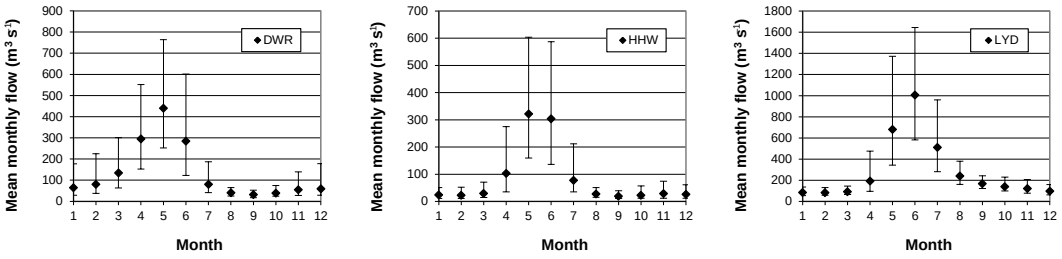

Figure 2. Median monthly streamflow and 10 and 90 % percentiles for the test locations Dworshak (DWR), Hungry Horse (HHW) and Libby (LYD).

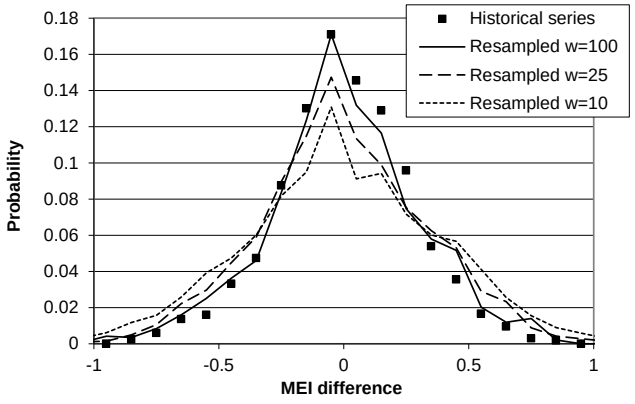

Figure 3. Distribution of MEI differences between consecutive months; historical series and three resampled time series with w values of 10, 25 and 100.

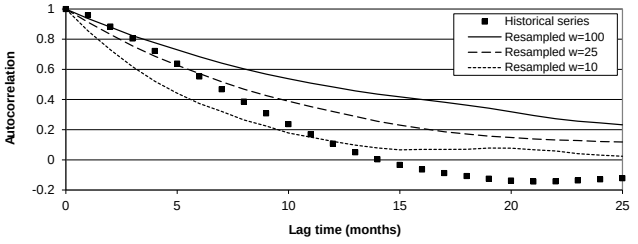

Figure 4. Autocorrelation of MEI signal for the historical and three resampled time series with w values of 10, 25 and 100.

tion (PDO), multivariate ENSO index (MEI), El Niño index NINO3.4 and Southern Oscillation Index (SOI). A correlation analysis was done between the index values in December and the annual flow volume in the following year. The MEI, as defined by Wolter and Timlin (1998), showed the highest correlation with the historical streamflows at the three test stations in this study and was therefore used for conditioning of the case study forecasts. The MEI combines several meteorological observables in a single metric and is issued monthly as a 2-month value.

To tune the parameter w for this case study, several values were evaluated. Figure 3 shows the distribution of differences between climate indices in consecutive months for the historical MEI series (1871–2013) and three resampled time series with w values of 10, 25 and 100. From this figure, a value of $w = 100$ seems optimal. However, the autocorrelation function (Fig. 4) shows that the $w = 100$ series has a higher autocorrelation than the historical time series. This can be explained by the fact that the historical series has a 2–3-year quasi-biannual frequency (Barnett, 1991). The autocorrelation turns negative after 15-month lag time, indicating that a positive ENSO phase is most likely followed by a negative ENSO phase in the succeeding year and vice versa. This periodic behaviour cannot be reproduced by the basic lag-1 resampling method. The autocorrelation of the resampled time series simply decays to zero.

In order to approximate the persistence of the historical climate index series, a weight w of 25 is chosen, which reproduces the autocorrelation of the historical MEI series at the relevant lead times for the seasonal forecasts, i.e. between 4 and 6 months.

The method was implemented as a module in Delft-FEWS, a hydrological forecasting and data management platform (Werner et al., 2013) upon which CHPS is built. The subsampler–resampler module was run from CHPS to generate meteorological forecasts with lead times up to 12 months for every month in the period 1949–2003. Next, ensemble streamflow hindcasts (reforecasts) were produced by running the hydrologic models, taking the subsampled and resampled MAP and MAT series as input. The year of hindcast was excluded from the subsampling and resampling schemes.

Figure 5 shows example hindcasts of (from top to bottom) climate index, monthly mean areal precipitation (MAP), monthly mean areal temperature (MAT) and monthly mean streamflow ensembles at forecasting station Dworshak, starting from reference dates 1 December of 1973 (La Niña year), 1978 (neutral) and 1997 (El Niño year). The historical values are shown in red. Except for the shortest lead times in a few cases, the historical traces fall within the range of the ensemble. The MEI, precipitation and temperature ensembles for the three starting dates differ due to the conditioning of the resampler. As a result, the streamflow ensembles have less spread than the original ESP and a better forecast skill, as will be shown in Sect. 4.

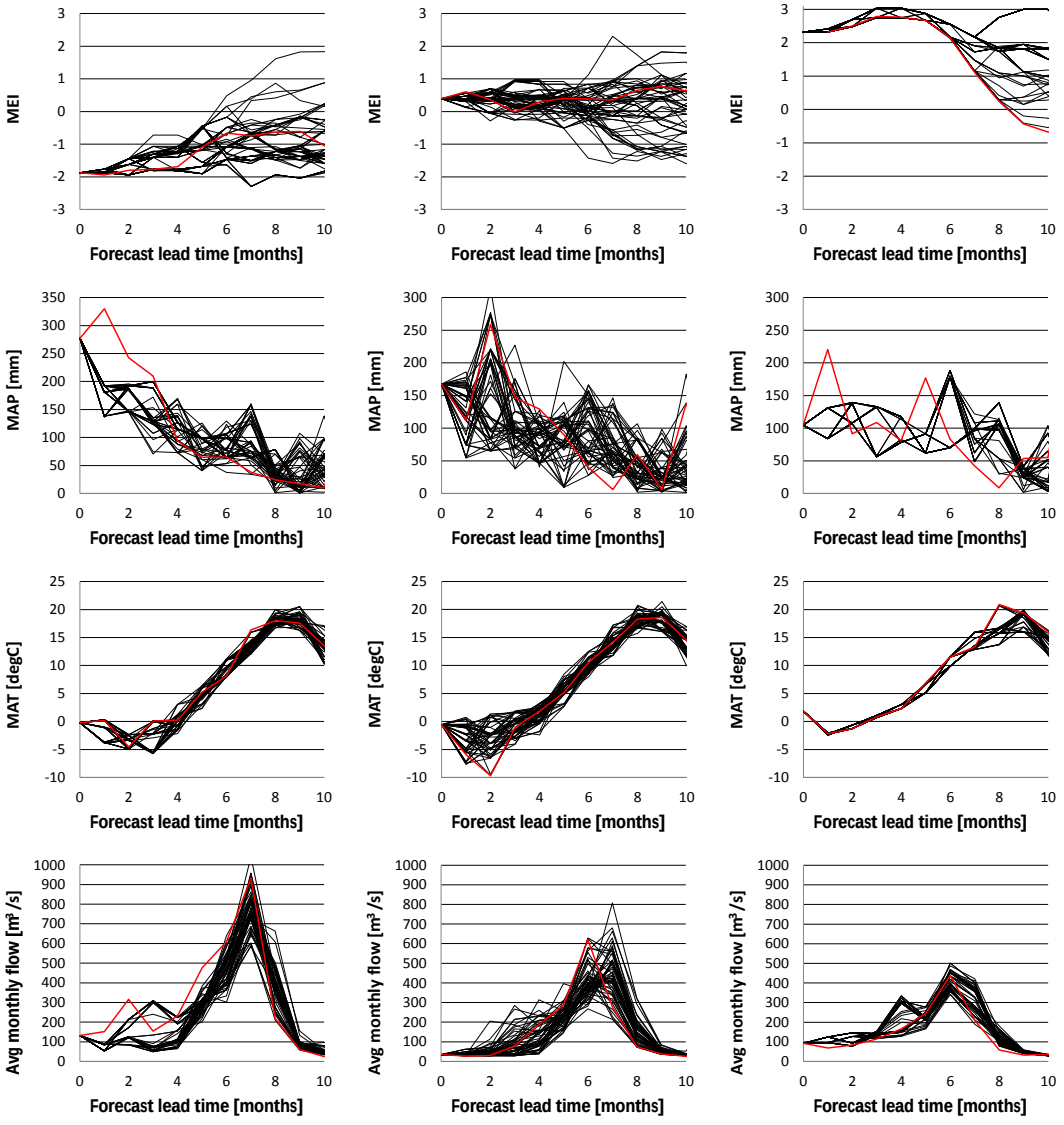

Figure 5. Resampled ensemble forecasts of (from top to bottom panels) MEI, MAP, MAT and streamflow at forecasting station Dworshak. Forecast dates are 1 December 1973 (left panels), 1978 (middle panels) and 1997 (right panels). The historical runs are shown in red.

Figure 6 shows the number of unique ensemble traces as a function of lead time. Different behaviour is found for the three forecasts. The 1997 forecast starts off from a rather extreme positive MEI. The probability of resampling a different historical year depends on the difference in MEI. Since the number of historical years that have such extreme MEI values is limited, a small set of historical years gets resampled multiple times and the number of unique ensemble traces after 5 resampling rounds is only 17. In contrast, the 1978 forecast starts off from an average MEI value, with many historical years with similar MEI values to resample from. As a result, each of the 50 ensemble traces is unique after 5 resampling rounds.

3.3 Forecast evaluation

The skill of the forecasts was assessed in terms of root mean square error (RMSE) of the ensemble mean, Brier score (BS) and continuous ranked probability score (CRPS). The RMSE is a direct measure of the accuracy of the mean forecast, but it does not account for ensemble spread. The BS and CRPS are integral measures of ensemble forecast quality (Jolliffe and Stephenson, 2003; Wilks, 2006). The Brier score was computed for a threshold level at 80 % exceedance probability of the monthly flow for each test station.

The subsampler–resampler method was run in parallel to the original ESP method within CHPS to enable a comparison. The skill metrics for the two methods were compared through relative skill scores, for example the Brier skill

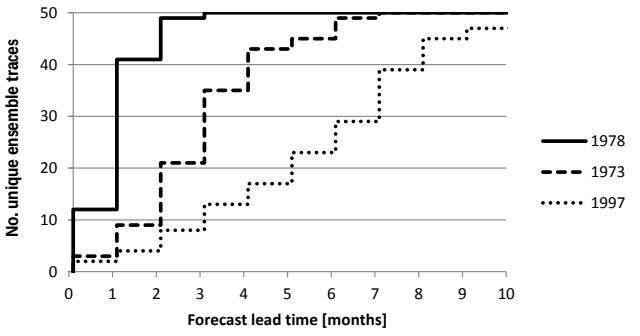

Figure 6. Number of unique ensemble traces in 50-member ensembles of resampled time series ($w = 25$), starting from 1 December 1978, 1973 and 1997.

score (BSS):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{model}}}{\mathrm{BS}_{\mathrm{reference}}}, \quad (2)$$

where the $\mathrm{BS}_{\mathrm{reference}}$ is the Brier score of the standard ESP method. Likewise, the continuous ranked probability skill score (CRPSS) and the relative improvement in RMSE are evaluated. The skill metrics were calculated using the Ensemble Verification System (Brown et al., 2010). The next section focuses on forecast skill for streamflows in May and June. These months have the largest variation (see Fig. 2), which makes the effect of an improved forecast more pronounced.

4 Results

The performance of the subsampler selecting historical years from the original ESP based on climate mode similarity was first evaluated without the addition of resampled time series. Figure 7 shows the BSS, CRPSS and relative reduction in RMSE of the resampler method in red as a function of the number of ESP ensemble traces. Skill scores reported here refer to May and June monthly flows and are averaged over forecast lead times between 3 and 12 months. The 50-year ensemble is identical to the original ESP and has a skill score of 0 by definition. Upon reducing the ensemble size, the forecast skill increases for two of the three test stations (Dworshak and Hungry Horse) as a result of dismissing historical years with dissimilar MEI values. This indicates that the climate mode conditioning is shifting the ensemble forecast towards the most probable outcome.

For one test station (Libby), the forecast skill decreases for smaller ensemble sizes. The reduction of the number of ensemble traces has an adverse effect on its statistical properties. The sampling uncertainty increases, which counteracts the gain in forecast skill from the climate mode information. The dashed lines represent the general behaviour of the forecast skill for a randomly reduced ensemble size, as described by Ferro (2007) for BSS. The analytical results for CRPSS and RMSE were derived from Ferro et al. (2008), Eq. (22), and Ho et al. (2013), Eq. (1), respectively. The streamflow at Libby has the weakest correlation with MEI. Apparently, the MEI information has little additional value for the Libby streamflow forecasts and their skill follows this trend. For the other two stations the skill also drops below zero for ensemble sizes less than 10.

Next, the forecast skill of the combined subsampler–resampler method was computed (black lines in Fig. 7). The ensemble traces that were dismissed in the subsampler are now replaced by resampled traces. The ensemble size is thus 50 in all cases. The forecast skill is still a function of the number of original ESP traces (full historical years), but in contrast to the subsampler forecasts, the subsampler–resampler produces a generally positive skill over the full range. The marginal loss of skill for Libby is probably due to statistical uncertainty of the skill score calculation (which could be verified by bootstrapping, but this is left for future studies). This demonstrates that the loss of skill from the reduction of ensemble size can be neutralized by additional ensemble traces from the resampler method. A mix of 10 historical years from the subsampler ESP and 40 additional resampled traces produces in general the best result for the three test stations in this case study.

Figure 8 shows the forecast skill as a function of forecast lead time. A combination of 10 historical and 40 resampled traces is used for all lead times. Three different skill metrics are shown for the May and June monthly flow from the three test stations. A positive skill is found up to 12 months of forecast lead time for Dworshak and Hungry Horse. This confirms the persistent nature of the ENSO climate mode. Because of this persistence, the conditioning of the subsampler and resampler on the climate phase at the time of forecast produces a positive skill over several months up to a full year in the future. For Libby, no gain in forecast skill is found.

Figure 8 shows that for lead times of 1 or 2 months, the skill is negative. This is due to a small effective ensemble size of the resampled traces for the shortest lead times, as discussed in Sect. 3.2. In order to maintain climate mode information on the seasonal timescale, the similarity criterion was set fairly stringent ($w = 25$). This produces good results for the 4- to 6-month lead times, but it causes the same small set of historical years to be selected in the first resampling rounds every run. Although the absolute number of ensemble traces is 50, a small subset of historical years keep reappearing in the resampled time series at the shortest lead times. This has a negative effect on the statistical properties of the ensemble and on the forecast skill. For longer lead times, this effect vanishes (see Fig. 6).

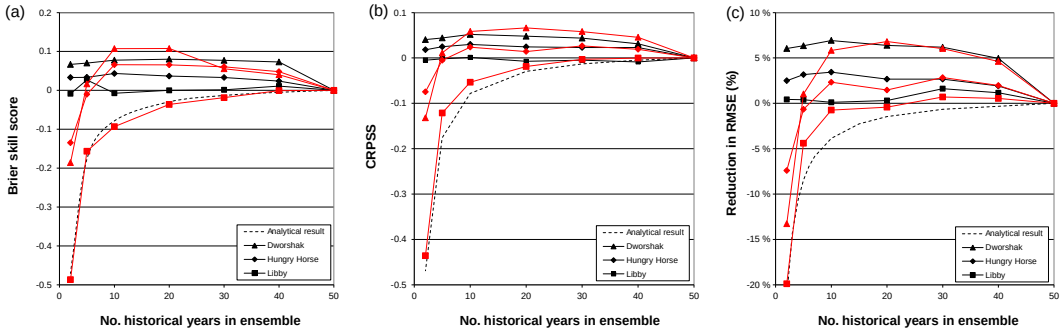

Figure 7. Forecast skill of the subsampler method (in red) compared to subsampler–resampler method (in black) as a function of number of historical ESP ensemble traces. **(a)** Brier skill score (80 % threshold), **(b)** CRPSS and **(c)** relative reduction of RMSE. Skill scores are averaged over 50 years of hindcasts for May and June monthly streamflow at lead times between 3 and 12 months.

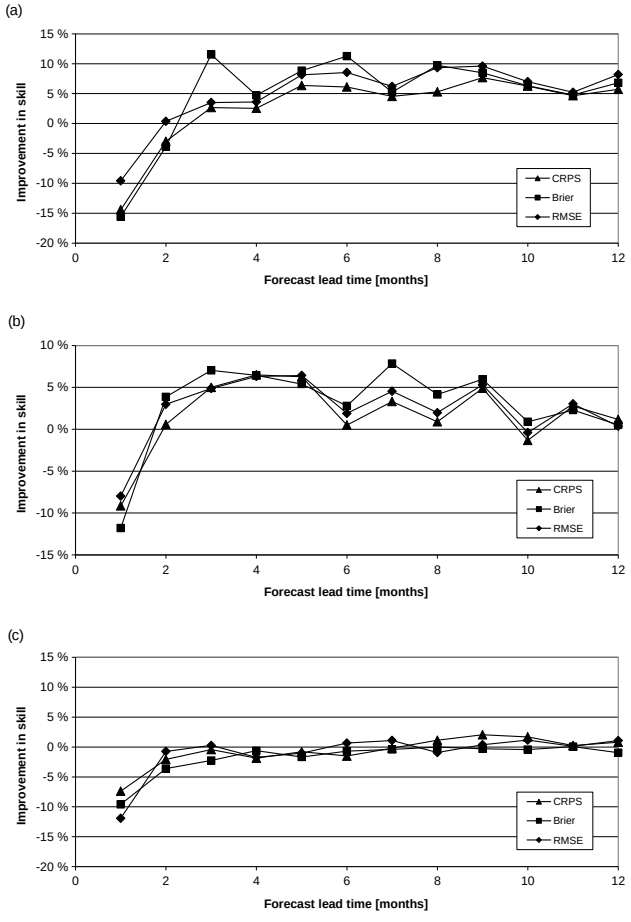

Figure 8. Improvement in May/June streamflow forecast skill of the subsampler–resampler ($w = 25$) method relative to the standard ESP as a function of lead time for **(a)** Dworshak, **(b)** Hungry Horse and **(c)** Libby test stations. Three different skill metrics: CRPS, Brier score and RMSE.

5 Discussion and conclusions

The results in the previous section show that the subsampler–resampler method is able to improve the ESP forecast skill by 5 to 10 % in two of the three test stations in this case study for lead times greater than 2 months. This improvement seems modest compared to the 28 % gain in forecast skill reported by Werner et al. (2004) and 27 % by Bradley et al. (2015), who used similar post-processing methods. We note, however, that the performance may vary considerably per forecasting station. Werner et al. (2004) found a much smaller skill improvement of 4 and 6 % for two other stations, which is comparable to the results found in this study. Moreover, Werner et al. (2004) used a separate calibration of post-processing parameters to arrive at a different set of weights for each test station. However, many operational applications require equally weighted ensembles for all forecasting stations in the area of interest. This requirement does not allow for a separate optimization of weights per station.

For the third test station in our case study, Libby, no improvement of skill was found. The streamflows at this station have the weakest correlation with MEI and the local climate is least affected by ENSO. It was shown that dismissing ensemble traces from the ESP leads to a reduction of forecast skill for this station that is similar to the expected reduction for a randomly reduced ensemble, as described by Richardson (2001), Ferro (2007) and Ferro et al. (2008). The same effect occurs for the other two stations for very small subsamples. A smaller ensemble has a less accurate ensemble mean and is less well capable of accurately describing a probability distribution. The subsample–resampler method resolves this issue. The additional traces from the resampler restore the forecast skill to that of the original ESP, and the adverse effect of the dismissal of ensemble traces by the subsampler is neutralized by the resampled traces. This is an important advantage of the subsampler–resampler method in operational settings, where avoiding loss of forecast skill anywhere is at least as important as improving the skill for a few forecasting stations.

The subsampler–resampler method also has some practical advantages over alternative approaches. Firstly, the subsampler–resampler produces an equal-likelihood streamflow ensemble, in contrast to the ensemble-weighting schemes. Also, the total number of ensemble traces can be set equal to the original number of ESP traces. This facilitates a comparison between the forecast skill of the conditioned ESP and that of the unconditioned ESP. Even more importantly, it facilitates the migration of an operational forecasting system from a standard ESP to a climate-mode conditioned ESP, since the downstream processes that use the streamflow ensemble as input do not need to be updated. Finally, the resampler method allows for a parallel sampling of non-meteorological variables from the historical record, with automatic preservation of cross-correlations. This is an important advantage for agencies like BPA that use these variables (e.g. power demand) in their water resources planning tools.

There are several parameters in the subsampling–resampling method that must be reconsidered or recalibrated if the method is applied to other regions or lead times of interest. Firstly, the relevant climate modes should be identified for the region of interest. To simplify the test case in this study, we have used only a single climate index: MEI. Next, the number of original ESP traces to be selected in the subsampler should be set. The optimal number of traces was found to be 10 in the current application, which is close to the values of 7 found by Werner et al. (2004), 12 by Najafi et al. (2012) and 9 by Bradley et al. (2015). Apparently, a selection of 15 to 20 % of original ESP traces gives the best performance for this type of ESP subsampling.

Another calibration parameter is the weight per climate index in the resampler procedure, which determines the persistence of the climate phase signal and the spread of the ensemble. It was found that a weight $w = 25$ gave the best results for the 4- to 6-month lead times of interest in this case study, although it leads to an underdispersed ensemble for the shorter lead times. A less stringent similarity criterion, i.e. a smaller w, would improve the spread for short lead times. However, this would lead to a less persistent climate phase signal and loss of forecast skill for the longer lead times.

There are several opportunities for further improvement of the method. For the Columbia Basin, a conditioning on other climate modes (e.g. PDO) could improve the forecast skill. This is being explored by BPA at the moment. The performance at short lead times can possibly be improved by introducing a random time shift in the historical resampling scheme. For example, instead of sampling a historical period 1–30 April, we shift 5 days back and sample 27 March–25 April. The time shifts would introduce more variability in the resampled traces without compromising the persistence of the climate phase signal. Another possible improvement is to employ GCM-based climate mode forecasts instead of the lag-1 resampling procedure described in Sect. 2.2. This is left for future research.

6 Data availability

MAP and MAT time series were provided by BPA (contact amcmanamon@bpa.gov). MEI historical data were obtained from http://www.esrl.noaa.gov/psd/enso/mei/table.html. Skill scores were calculated using the freely available EVS, see www.nws.noaa.gov/oh/evs.html. For other data, e.g. those underlying graphs 7 and 8, please contact joost.beckers@deltares.nl.

Acknowledgements. This work was supported by Bonneville Power Administration. The authors wish to thank Ann McManamon from BPA for valuable comments and providing test data.

Edited by: M.-H. Ramos
Reviewed by: two anonymous referees

References

Abudu, S., King, J. P., and Pagano, T. C.: Application of Partial Least-Squares Regression in Seasonal Streamflow Forecasting, J. Hydrol. Eng., 8, 612–623, 2010.

Anderson, E. A.: A Point Energy and Mass Balance Model of a Snow Cover', NOAA Technical Report NWS 19, NOAA, Silver Spring, MD, 1976.

Barnett, T. P.: The interaction of multiple time scales in the tropical climate system, J. Climate, 4, 269–285, 1991.

Beebee, R. A. and Manga, M.: Variation in the relationship between snowmelt runoff in Oregon and ENSO and PDO, J. Am. Water Resour. Assoc., 40, 1011–1024, 2004.

Bradley, A. A., Habib, M., and Schwartz, S. S.: Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach, Water Resour. Res., 51, 7382–7400, doi:10.1002/2014WR016811, 2015.

Brandsma, T. and Buishand, T. A.: Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling, Hydrol. Earth Syst. Sci., 2, 195–209, doi:10.5194/hess-2-195-1998, 1998.

Brown, J. D., Demargne, J., Seo, D.-J., and Liu, Y.: The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environ. Modell. Softw., 25, 854–872, 2010.

Burnash, R. J. C.: The NWS River Forecast System - catchment modeling, in: Computer Models of Watershed Hydrology, edited by: Singh, V. P., Water Resources Publications, Littleton, CO, 311–366, 1995.

Burnash, R. J. C., Ferral, R. L., and McGuire, R. A.: A Generalized Streamflow Simulation System – Conceptual Modeling for Digital Computers, US Department of Commerce, National Weather Service and State of California, Dept. of Water Resources, Technical Report, Sacramento, CA, 1973.

Clark, M. P. and Hay, L. E.: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, J. Hydrometeorol., 5, 15–32, 2004.

Croley II, T. E.: Using NOAA's new climate outlooks in operational hydrology, J. Hydrol. Eng., 1, 93–102, 1996.

Croley II, T. E.: Mixing probabilistic meteorology outlooks in operational hydrology, J. Hydrol. Eng., 2, 161–168, 1997.

Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Res. Pl.-ASCE, 111, 157–170, 1985.

Diaz, H. F. and Markgraf, V.: El Niño and the Southern Oscillation: Multiscale Variability and Global and Regional Impacts, Cambridge University Press, Cambridge, 2000.

Dracup, J. A., and Kahya, E.: The relationships between U.S. streamflow and La Niña events, Water Resour. Res., 30, 2133–2141, 1994.

Druce, D. J.: Insights from a history of seasonal inflow forecasting with a conceptual hydrologic model, J. Hydrol., 249, 102–112, 2001.

Ferro, C. A. T.: Comparing Probabilistic Forecasting Systems with the Brier Score, Weather Forecast., 22, 1076–1088, 2007.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 19–24, 2008.

Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin, J. Hydrometeorol., 4, 1105–1118, 2003.

Gedalof, Z., Peterson, D. L., and Mantua, N. J.: Columbia River flow and drought since 1750, J. Am. Water Res. As., 40, 1579–1592, 2012.

Gijsbers, P. J. A., Cajina, L., Dietz, C., Roe, J. M., and Welles, E.: CHPS – an NWS development to enter the interoperability era, Eos Trans. AGU, 90, Fall Meeting 2009 Suppl. Abstract IN11A-1041, 2009.

Halpert, M. S. and Ropelewski, C. F.: Surface temperature patterns associated with the southern oscillation, J. Climate, 5, 577–593, 1992.

Halpert, M. S. and Ropelewski, C. F.: Surface temperature patterns associated with the southern oscillation, J. Climate, 5, 577–593, 1992.

Hamlet, A. F. and Lettenmaier, D. L.: Columbia River Streamflow Forecasting Based on ENSO and PDO Climate Signals, J. Water Resour. Pl. Manage., 125, 333–341, 1999.

Hay, L. E., McCabe, G. J., Clark, M. P., and Risley, J. C.: Reducing streamflow forecast uncertainty: Application and qualitative assessment of the upper Klamath River basin, Oregon, J. Am. Water Res. Assoc., 45, 580–596, 2009.

Ho, C. K., Hawkins, E., Shaffrey, L., Böcker, J., Hermanson, L., Murphy, J. M., Smith, D. M., and Eade, R.: Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion, Geophys. Res. Lett., 40, 5770–5775, doi:10.1002/2013GL057630, 2013

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, Wiley, New York, 2003.

Kahya, E. and Dracup, J. A.: United-States streamflow patterns in relation to the El-Niño Southern Oscillation, Water Resour. Res., 29, 2491–2503, 1993.

Kang, T.-H., Kim, Y.-O. and Hong, I.-P.: Comparison of pre- and post-processors for ensemble streamflow prediction, Atmos. Sci. Lett., 11, 153–159, doi:10.1002/asl.276, 2010.

Kiladis, G. N. and Diaz, H. F.: Global climatic anomalies associated with extremes in the southern oscillation, J. Climate, 2, 69–90, 1989.

Knight, J. R., Folland, C. K., and Scaife, A. A.: Climate impacts of the Atlantic Multidecadal Oscillation, Geophys. Res. Lett., 33, L17706, doi:10.1029/2006GL026242, 2006.

Lau, K. M. and Sheu, P.: Annual cycle, QBO and Southern Oscillation in global precipitation, J. Geophys. Res., 93, 10975–10988, 1988.

Leung, L. R., Hamlet, A. F., Lettenmaier, D. P., and Kumar, A. A.: Simulations of the ENSO hydroclimate signals in the Pacific Northwest Columbia River basin, B. Am. Meteorol. Soc., 80, 2313–2330, 1999.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114, doi:10.1029/2008JD010969, 2009.

Lü, A., Jia, S., Zhu, W., Yan, H., Duan, S., and Yao, Z.: El Niño-Southern Oscillation and water resources in the headwaters region of the Yellow River: links and potential for forecasting, Hydrol. Earth Syst. Sci., 15, 1273–1281, doi:10.5194/hess-15-1273-2011, 2011.

Madadgar, S., Moradkhani, H., and Garen, D.: Towards improved reliability and reduced uncertainty of hydrologic ensemble forecasts using multivariate postprocessing, Hydrol. Process., 28, 104–122, doi:10.1002/hyp.9562, 2012.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C.: A Pacific interdecadal climate oscillation with impacts on salmon production, B. Am. Meteorol. Soc., 78, 1069–1079, 1997.

McCabe, G. J. and Dettinger, M. D.: Primary modes and predictability of year to year snowpack variations in the western United States from teleconnections with Pacific ocean climate, J. Hydrometeorol., 3, 13–25, 2002.

McEnery, J., Ingram, J., Duan, Q., Adams, T., and Anderson, L: NOAA's advanced hydrologic prediction service: building pathways for better science in water forecasting, B. Am. Meteorol. Soc., 86, 375–385, 2005.

Najafi, M. R., Moradkhani, H., and Piechota, T. C.: Climate signal weighting methods vs. Climate Forecast System Reanalysis, J. Hydrol., 442–443, 105–116, 2012.

Perica, S.: Integration of Meteorological Forecasts/Climate Outlooks into an Ensemble Streamflow Prediction System, 14th Conference on Probability and Statistics in the Atmospheric Sciences, 78th AMS Ann. Meet., Phoenix, AZ, 130–133, 1998.

Philander, S. G.: El Niño, La Niña, and the Southern Oscillation, in: Int. Geophys. Ser. Vol. 46, Academic Press, San Diego, CA, 293 pp., 1990.

Pica, J. A.: Review of Extended Streamflow Prediction of the National Weather Service River Forecast System, CE505 Conference Course, Civil Engineering, Portland State University, Portland, 1997.

Piechota, T. C. and Dracup, J. A.: Drought and regional hydrologic variation in the United States: Associations with the El Niño Southern Oscillation, Water Resour. Res., 32, 1359–1373, 1996.

Piechota, T. C., Dracup, J. A., and Fovell, R. G.: Western US streamflow and atmospheric circulation patterns during El Niño Southern Oscillation, J. Hydrol., 201, 249–271, 1997.

Redmond, K. T. and Koch, R. W.: Surface climate and streamflow variability in the western United States and their relationship to large scale circulation indices, Water Resour. Res., 27, 2381–2399, 1991.

Richardson, D. S.: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, Q. J. Roy. Meteorol. Soc., 127, 2473–2489, 2001.

Ropelewski, C. F. and Halpert, M. S.: North American precipitation and temperature patterns associated with the El Niño Southern Oscillation (ENSO), Mon. Weather Rev., 114, 2352–2362, 1986.

Ropelewski, C. F. and Halpert, M. S.: Quantifying Southern Oscillation–Precipitation Relationships, J. Climate, 9, 1043–1059, 1996.

Sagarika, S., Kalra, A., and Ahmad, S.: Interconnections between oceanic–atmospheric indices and variability in the U.S. streamflow. J. Hydrol., 525, 724–736, doi:10.1016/j.jhydrol.2015.04.020, 2015.

Shukla, S. and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529–3538, doi:10.5194/hess-15-3529-2011, 2011.

Stedinger, J. R. and Kim, Y. O.: Probabilities for ensemble forecasts reflecting climate information, J. Hydrol., 391, 11–25, 2010.

Tootle, G. A. and Piechota, T. C.: Relationships between Pacific and Atlantic ocean sea surface temperatures and U.S. streamflow variability, Water Resour. Res., 42, W07411, doi:10.1029/2005WR004184, 2006.

Tootle, G. A., Piechota, T. C., and Singh, A. K.: Coupled oceanic-atmospheric variability and US. streamflow, Water Resour. Res., 41, W12408, doi:10.1029/2005WR004381, 2005.

Tootle, G. A., Singh, A. K., Piechota, T. C., and Farnham, I.: Long Lead-Time Forecasting of U.S. Streamflow Using Partial Least Squares Regression, J. Hydrol. Eng., 12, 442–451, 2007.

Twedt, T. M., Schaake, J. C., and Peck, E. L.: National weather service extended streamflow prediction, Proc. Western Snow Conference, Albuquerque, NM, 52–57, 1977.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Generalization of the discrete Brier and ranked probability skill scores for weighted multi-model ensemble forecasts, Mon. Weather Rev., 135, 2778–2785, 2007.

Weijs, S. V. and van de Giesen, N.: An information-theoretical perspective on weighted ensemble forecasts, J. Hydrol., 498, 177–190, 2013.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S.: Ensemble Streamflow Prediction: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts, J. Hydrometeorol., 5, 1076–1090, 2004.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., and Heynert, K.: The Delft-FEWS flow forecasting system, Environ. Model. Softw., 40, 65–77, 2013.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, London, 2006

Wolter, K. and Timlin, M. S.: Measuring the strength of ENSO – how does 1997/98 rank?, Weather, 53, 315–324, 1998.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res., 107, 4429, doi:10.1029/2001JD000659, 2002.

Wood, A. W. and Lettenmaier, D. P.: A testbed for new seasonal hydrologic forecasting approaches in the western US, B. Am. Meteorol. Soc., 87, 1699–1712, doi:10.1175/BAMS-87-12-1699, 2006.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, L14401, doi:10.1029/2008GL034648, 2008.

Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, J. Geophys. Res.-Atmos., 110, D04105, doi:10.1029/2004jd004508, 2005.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resour. Res., 49, 4687–4699, doi:10.1002/wrcr.20350, 2013.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, WIREs Water, 2, 523–536, doi:10.1002/wat2.1088, 2015.