ENSO-Conditioned Weather Resampling Method for Seasonal Ensemble Streamflow Prediction

Joost V.L. Beckers, Albrecht H. Weerts, Erik Tijdeman and Edwin Welles

This memo includes a point-by-point response to the comments by the Editor and two Anonymous Referees. For each point, we indicate the changes that were made to the manuscript. At the end of this document is a marked-up manuscript version showing all changes.

RESPONSE TO EDITOR'S COMMENTS

Editor Decision: Reconsider after major revisions (02 May 2016) by Maria-Helena Ramos

Comments to the Author:

The authors are acknowledged for the answers to the two referees and are encouraged to submit a revised version of their manuscript. Please, specifically, consider the major remarks of the referees, namely:

- include omitted results to show if the method is actually robust;

This has been done, see Figure 7 and 8.

- provide more evidence that "several climate mode indices and combinations of indices for ensemble member selection and conditioning of the subsampler were evaluated" as mention in Section 3.2 by the authors;

The aim of this paper is to explain the method and demonstrate its use in a simple test case: a limited number of test basins and a single climate index (MEI). It is shown how the method performs for two test basins where the streamflow is correlated with MEI and for a test basin that is not strongly affected by the climate signal, in this case Libby. An optimization of the method for a larger number of test basins using other climate signals (including PDO) will be done by BPA. This is mentioned in the discussion. Results of that optimization study may be published separately at a later time.

- clarify how parameters were tuned or determined, indicating when cross validation was performed or not (and explaining the reasons behind the choices made);

The parameter setting is explained in Sect. 3.2. We add the following additional information:

- The climate index MEI was chosen from several candidates (SOI, ENSO3.4, PDO) based on a correlation analysis. The correlation analysis was done between the index values in December and the annual flow volume in the next year for the period 1948-2008. The MEI has the highest correlation (around 0.5 for DWR and HHW, 0.35 for LYD) with the historical streamflows for the three sub-basins and was therefore selected for this case study. PDO has the second highest correlation (around 0.35 for all three sub-basins). The correlation analyses were repeated for different months (lead times) and for different historical periods and the results showed a consistent picture that MEI is the strongest teleconnection at these lead times and PDO is the second strongest.
- It is also explained in section 3.2 how the value of the weight w was determined: based on the autocorrelation of the MEI signal (Figures 3 and 4). It is explained why a value of 25 was found appropriate for forecasting at the seasonal time scale.
- The choice for the number of ESP sub-samples is based on heuristic techniques. Several values were tested (see Figure 7). From this Figure, a combination of 10 subsampled members gives the best skill, but larger values also produce a positive skill. In Sect. 5, the optimal value of 10 sub-samples is

compared to values found in comparable methods from literature: “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) 5 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.”

A cross-validation using split datasets was not done because the resampling method relies on a large historical dataset. Splitting the dataset into a calibration and validation set would increase the uncertainty and reduce the skill. The results from a split dataset would therefore not be comparable to the results for the full dataset. Also note that none of the studies mentioned above (Werner et al., 2004; Najafi et al, 2012; Bradley et al., 2015) includes a cross validation of parameter settings.

- provide details on the methodology, as stressed by Referee 2;

More details were given where requested by Referee 2.

- provide enhanced figures for the results, such as these can be better evaluated.

This was done.

RESPONSE TO COMMENTS FROM REFEREE #1

Anonymous Referee #1
Received and published: 7 March 2016

In this paper, the authors propose a technique that combines a post-processing step – i.e., sub-sampler of raw ensemble streamflow prediction (ESP) outputs based on climate index similarity – with a pre-processing step that generates synthetic precipitation and temperature time series via resampling, based on climate index similarity, to force hydrologic model simulations and re-populate the previously sub-sampled ensemble forecast. The method is applied in three catchments located in the Pacific Northwest, using the SAC-SMA and Snow-17 models, for seasonal (May-June) streamflow forecasting. The authors conclude that their framework is an improvement in skill (RMSE, Brier Score and Continuous Ranked Probability Score) over both standard ESP and climate-based subsampling.

The paper is in general well written and well organized, the proposed technique is scientifically sound and the results are quite interesting. Further, the connection with the existing literature on this topic is nicely conducted. In my opinion, the manuscript has a lot of potential for publication in HESS, but the authors need to clarify some methodological choices, revise some statements, and include omitted results to show if the method is actually robust.

Major comments:

1. Why didn't the authors include the results for improvement in skill (as in Figure 9) for Libby and Hungry Horse? I think that showing the results at these locations is critical to demonstrate that the proposed technique is an advance over raw ESP and climate-based subsampling (see comment #14 for more details on this).

Figures for Hungry Horse and Libby have been added as additional frames in Figure 8 (formerly Figure 9). For Hungry Horse the improvement in skill is smaller than for Dworshak. For Libby, there is no gain or loss in forecast skill.

2. P4, L29: It is inferred from this paragraph that the reference date is set to the day when the forecast is initialized. Further, it is also mentioned that "the year of the reference date even has the highest probability of being re-selected". However, later in the paper the authors mention that "the year of reforecast was excluded from the subsampling and resampling schemes" (P8, L24). These statements are confusing, so the authors should clarify what was actually done. In my opinion, the year of the reference date (or initialization time) should NOT be included in the subsampling/resampling procedures, since that year is the one forcing the forecast.

The year of hindcast is excluded from the resampling to be able to assess the forecasting skill. At the start of the resampling procedure, the reference date is equal to the forecast date. That year is excluded from the resampling, so a different historical year must be selected in the first resampling round. In the next resampling round, however, the reference date is set to a date in the historical year that was selected in the first round. That year is not excluded from the resampling, so it can be selected again.

3. P8, L2: The authors state that "several climate mode indices and combinations of indices for ensemble member selection and conditioning of the subsampler were evaluated". However, from the same paragraph it is implied that MEI was selected because it provided the highest correlation with historical streamflow. Did the authors actually test several combinations of climate indices? Moreover, it has been shown that PDO strongly affects interannual variability of runoff in this region (e.g., McCabe, G.J., Wolock 2014; Sagarika et al. 2015). Did the authors perform any experiments including both MEI and PDO in the subsampling process? I think this manuscript would greatly benefit if - at least for the subsampler method - additional experiments showing the use of PDO were included. My guess is that the poor results obtained at Libby may be related to this issue.

The aim of this paper is to explain the proposed method and demonstrate its use in a simple test case: a limited number of test basins and a single climate index. This proof of concept includes demonstrating how the method

performs for a test basin that is not strongly affected by the climate signal, in this case Libby. An optimization of the method for other locations in the Columbia River basin and using other climate signals (including PDO) will be done by BPA. This is mentioned in the discussion. Results of that optimization study may be published separately at a later time.

Minor comments:

4. P1, L23: The authors should note that the hydrologic model does not necessarily have to be conceptual in ESP frameworks.

Agreed. We removed the word 'conceptual'.

5. Throughout the manuscript: the authors refer to "reforecasts" or "forecasts in retrospect" when reporting results, but it might be better to use the word "hindcasting" (Beven and Young 2013).

We changed 'reforecasts' into 'hindcasts' (4 instances). The first time that the term 'hindcast' is mentioned, we add '(reforecasts)' in brackets for clarity. The term 'reforecasts' is also used in literature, e.g. by Werner (2004) and Wood (2002).

6. P2, second paragraph: the text may be enriched by adding a few more references (Hamlet and Lettenmaier 1999; Tootle et al. 2007; Abudu et al. 2010; Sagarika et al. 2015).

Thanks for this suggestion. We added these references.

7. P2, L18: Several studies recommend developing custom climate indices for the basin(s) of interest using reanalysis datasets (e.g., Grantz et al. 2005; Regonda et al. 2006; Block et al. 2009; Opitz-Stapleton et al. 2007; Bracken et al. 2010; Mendoza et al. 2014), instead of using standard climate indices for predicting seasonal runoff volumes. This point could be made in the introduction.

We feel that these custom climate indices should be part of the optimization of the method for a specific area and lead time of interest. Our paper focuses on explaining the basic method and demonstrating its use in a simple test case of three locations and a single climate index. Optimization of the method for a larger study area using multiple indices and/or custom climate indices would be a separate study. BPA is currently carrying out the optimization and results of that may be published at a later time (see also our answer to point 3).

8. P2, L21: The reference is missing here.

This was corrected.

9. P5, L17: A better title for section 3 would be "Example Application".

Agreed, we changed the title.

10. P7, Table 1: It would be more informative to add mean basin elevation (or elevation range), mean annual runoff and mean annual precipitation (mm/yr), and runoff ratio. I think that powerhouse capacity is not relevant here.

Agreed. We added average elevation, mean runoff, precipitation and runoff ratio and removed powerhouse capacity.

11. I strongly encourage the authors to improve the quality (resolution) of Figures 1, 4, 5, 7 and 8. This is critical to enhance the readability of the paper.

The figures were improved.

12. Figures 7 and 8: The authors could merge the results displayed here into a single figure, using different colors for different methods (for instance, red for subsampler, and black for combined subsampler-resampler), and keeping the title of x-axis label as "Number of historical years in ensemble". This would allow a direct comparison

between the proposed method and the benchmark technique (i.e. only sub-sampling). I also think that the authors should add two additional panels (similar to the one described) with results of CRPSS – which is in my opinion a much more interesting score to assess the skill of ensemble systems – and RMSE. Further, it should be mentioned in the caption that results are averaged over lead times of 1-12 months.
Figures 7 and 8 were combined and two additional panels with CRPS and RMSE results were added.

Results are averaged over lead times 3 to 12 months, because the skill for 1 and 2 months is poor. The fact that the skill scores are averaged is mentioned in the caption.

13. Figures 7-9: The captions indicate that results are for May-June flows, but the text refer to June flows. What is actually being presented? If results are for May-June flows, are these aggregated (i.e. how many values are used for computing the scores, Nyears or 2 x Nyears)? Is the 80% flow computed from all monthly streamflow values, or only from May and June historical flows?
What is shown are the verification scores for forecasts of monthly streamflows for May and June. This was clarified in the text.

14. Figure 9: As pointed in comment #1, the authors are encouraged to add and discuss results for Libby and Hungry Horse in this figure. This could be done by or adding two panels (b and c, for instance), or extra lines with different colors for each basin. The improvement in skill could also be compared to that obtained from using only subsampling (the benchmark method) to understand the added value of re-populating the ensemble.
Additional panels were added to Figure 8 (formerly Figure 9) for Libby and Hungry Horse and results are discussed in the text. The gain in forecast skill for these subbasins is less than for Dworshak. For Libby there is no gain in forecast skill.

15. P13, L10-16: The authors might want to re-word or delete a couple of sentences. For instance, they point for Figure 8 that "in contrast to Fig. 7, the BSS for all test basins are now positive over the full range", which is NOT true for the Libby reservoir (there are still negative BSS values). Moreover, the authors mention that "a mix of 10 historical years from the subsampler ESP and 40 additional resampled traces produces the best result for these sub-basins", which is inaccurate again when looking at Libby (higher BSS is obtained using five historical years).
The small negative score for Libby and the positive skill for five historical years are attributed to uncertainty/noise in the calculation, i.e. statistical uncertainty related to the limited number of hindcasts. We rephrased these sentences to:
"in contrast to the skill of the subsampler forecasts, the subsampler-resampler produces in general a positive skill over the full range. The marginal loss of skill for Libby is attributed to statistical uncertainty of the skill score calculation."
"a mix of 10 historical years from the subsampler ESP and 40 additional resampled traces produces in general the best result for these sub-basins"

Suggested minor edits:
16. P1 L23: "forcing" -> "forcings". Agreed, changed.
17. P2, L27: "case study" -> "case study basin". Agreed, changed.
18. P2, L26: "weigh" -> "weight". Agreed, changed.
19. P3, L19-21: "Sect." -> "Section". No change made, we think this is HESS-style
20. P5, L13: "needs" -> "need". Agreed, changed.
21. P7, L9: "of e.g." -> "with"; "into the states" -> "into model states". Rephrased to:
'… blending in recent snow pack and streamflow gauge data into model states'
22. P8, L1: "parameter tuning" -> "parameter calibration". Agreed, changed.
23. P12, L18: "the most variation" -> "the largest variation". Agreed, changed.

RESPONSE TO COMMENTS FROM REFEREE #2

Anonymous Referee #2
Received and published: 14 March 2016

This paper presents a two-pronged approach for conditioning ESP forecasts on ENSO conditions. In the first step, a sub-sample of ESP forecasts are selected from an ensemble (e.g. of size 50) by conditioning on a climate index. This reduces the number of ensemble members. In the second step, the ensemble is augmented to the original size by sampling precipitation and temperature from the historical record, conditioned on the climate index, and thereafter producing additional ESP forecasts. I think the paper presents a pragmatic approach to incorporating climate information into ESP forecasts and for enlarging the ensemble size. These types of technique are of wide interest in the hydrologic ensemble forecasting community. The writing is generally of publication quality but several figures need improvement.

I have some issues with the clarity, execution and explanation of the science. If the authors can thoroughly address the issues, some of which are not simple, my opinion is the paper should eventually be published in HESS. General comments:

1) A number of parameters are tuned on the basis of subjective analysis for the whole period of interest. Because this is a forecasting paper, the parameter values ought to be determined from an objective analysis that can then be cross-validated using a leave-out scheme. If the results are not cross-validated then the results are potentially inconclusive. Given that the results are marginal, and perform best for the period tuned to (4–6 months lead time), I suggest this is quite important.

Ideally, the following elements would be cross-validated:
a. The climate index selection
b. The number of optimal ESP sub-samples selected
c. The "weight", w. If cross-validation isn't used, justification is required.

We believe that the parameter setting is not subjective. It is explained in the manuscript how the climate index selection was done and how the weight *w* was determined on statistical analysis of the climate signal before the actual hindcasts were made, i.e. without using hindcast information:

a. The climate index MEI was chosen from several candidates (SOI, ENSO3.4, PDO) based on a correlation analysis with historical streamflow data. We added a sentence "A correlation analysis was done between the index values in December and the annual flow volume in the next year." to clarify how this was done. The MEI has the highest correlation (around 0.5 for DWR and HHW, 0.35 for LYD) with the historical streamflows for the three sub-basins and was therefore selected for this case study. PDO has the second highest correlation (around 0.35 for all three sub-basins). These correlation analyses were repeated for different months (lead times) and for different historical periods and the results showed a consistent picture that MEI is the strongest teleconnection at these lead times and PDO is the second strongest.
b. See below.
c. It is explained in section 2.3, page 8 how the value of the weight *w* was determined, based on the autocorrelation of the MEI signal (Figures 3 and 4). It is explained why a value of 25 was found suitable for forecasting at the seasonal time scale.

For the third parameter: the number of ESP sub-samples selected, several values were tested and results presented in Figure 7. A consistent positive forecast skill is found for two sub-basins, except for less than 10 sub-samples. The reason for the poor scores for small numbers of sub-samples is explained in the text (see also

response to remark nr 2). The absence of a gain in forecast skill for the third sub-basin and the loss of forecast skill for short lead times are also explained.

Based on Figure 7, a combination of 10 subsampled members and 40 resampled members is chosen as optimal in this case, but larger values also produce a positive skill. In Sect. 5, the optimal value of 10 sub-samples is compared to values found in comparable methods from literature: "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) 5 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." Note that none of these studies included a cross validation of parameter settings.

A cross-validation on split datasets indeed could provide insight into the uncertainty of the results. However, the uncertainty of the calculated verification scores would increase for a smaller dataset, so we are not sure if this analysis would be conclusive. In general, we feel that we have shown that the results are rational and robust to the choice of parameter settings.

2) The results use the Brier score (for 80% exceedance probability forecasts) and CRPS as probabilistic measures. I think the paper would be much stronger if accuracy skill and reliability results were separated. Whether skill is attributable to accuracy or reliability or both may vary significantly with lead time. Also, it is stated repeatedly throughout the paper that a small effective number of ensemble members is associated with "degradation of the statistical properties" of the ensemble forecast. What exactly does this mean? I suggest be specific and explain exactly which properties are affected and how they are affected. This is particularly important in the results (P15 L6) and discussion (P15 L18–19).

In answer to the first remark, we use three different skill metrics that are quite common in forecasting. They are related to typical usage of a probabilistic forecast, namely the best estimate or mean forecast (the accuracy of which is measured by RMSE), the probability of exceeding a critical threshold (measured by Brier score) and the overall reliability of the forecast probabilities (measured by CRPS). Many other skill scores and measures of forecast quality are possible but we feel that these three cover the most important aspects of a probabilistic forecast.

In answer to the second remark, the effect of a reduction of ensemble size on verification scores is well-known. The effect of ensemble size on Brier score has been analysed extensively by Richardson (2001) and Ferro (2007). An extension to CRPS was done by Ferro et al (2008). The RMSE of the ensemble mean also increases with decreasing ensemble size (see e.g. Ho et al., 2013, Eqn (1) or Weigel 2007 for weighted ensembles). An ensemble of fewer members has a less accurate ensemble mean and is less well capable of accurately describing a probability distribution. In the manuscript, we explain this effect qualitatively and refer to existing literature where appropriate:

- Page 3: "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; Ferro et al, 2008)."
- Page 4: "…there is a trade-off between specificity and sampling error. With fewer years (ensemble members), the resolution of the ensemble decreases and the sampling error increases."
- Page 12, Line 20 and further: "The reduction of the number of ensemble members 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), Eqn. 22 and Ho et al. (2013), Eqn. 1 respectively."
- Page 15 (discussion): "It was shown that dismissing ensemble members from the ESP leads to a reduction of forecast skill for this sub-basin that is similar to the expected reduction for a randomly reduced ensemble, as described by Richardson (2001), Ferro (2007) and Ferro et al. (2008). "

More references (Ferro et al. 2008; Ho et al. 2013) were added for effects on CPRS and RMSE as these scores were added to Figure 7 following a suggestion from reviewer nr 1. We believe that the general description of the effect and references to literature are adequate for this manuscript.

3) The resampling approach performs poorly for short lead times. Particularly, as shown by Figure 9, the forecasts at short lead times are up to 16% worse. The resampler produces much too narrow forecasts for the first couple of months. This is a problem with the ad-hoc nature of the approach, the spread in the ensembles at any given lead time could be either too narrow or too wide or somewhere in between. What happens if the resampling begins several months prior to the forecast date (i.e. lag 2 or lag 3 MEI)? It's a hard sell to say that forecasts get worse as lead time shortens. At what point should the forecasts be ignored? I encourage a resolution.
To make full use of the information of the current climate signal we do not recommend starting the resampling several months prior to the forecast date. Instead, we describe a way to improve the performance at shorter lead times in Sect. 5: "The performance at short lead times can possibly be improved by introducing a random time shift in the historical resampling scheme. This would introduce more variability in the resampled traces without compromising the persistence of the climate phase signal. "

The poor performance at short lead times is not necessarily problematic if the ESP is used only for forecasting at longer lead times (4 months or longer) and other techniques (e.g. NWP weather input) are used for forecasting at short lead times.

Specific comments

4) Abstract, last sentence: This needs to explicitly say when and where improvements of up to 10% are found and probably should also say that the results for short lead times are worsened.
The forecast skill improvement of 5 to 10% for two sub-basins is mentioned as well as the lack of improvement for the third sub-basin. We choose not to mention the poor performance for short lead times in the abstract because the method is meant for seasonal forecasting at longer lead times, as the title says. The poor performance at 1 and 2 month lead time is discussed extensively in the results and discussion sections. A possible solution is described on Page 16, Line 15-16.

5) P4 L1 suggests selecting climate indices based on correlations with MAT/MAP. But P8 4–5 reports MEI was selected on the basis of correlation with streamflow. Please make more consistent.
We changed page 4 line 1 to "historical streamflows". MAP/MAP would also be possible but that is not what was done here.

6) MEI is a two-month index. Were two-month values of the other indices considered?
Indeed two- and more-month averaged values of other indices were considered in the correlation analysis, but the results were no better than for the original indices.

7) Equation (1): The summation appears to be the squared Euclidean distance (no square root). Also, how are indices in different units handled (is it implicitly through scaling/weighting)?
The indices can be normalised or the weights w could carry units. In principle, the weights can have any positive value (as mentioned on Page 5, Line 3).

8) Figure 2: It might be better to show percentile intervals rather than statistics based on normal distributions (unless of course the data is very normal).
Agreed, the figure was changed to show median and 10% and 90% error bars.

9) P13 L10–13: The BSS is marginally negative for some cases for Libby Dam, so the statement saying BSS is positive for all cases needs correcting. Also, re the comment about Figure 8, the text says the BSS is a function of "number of the original ESP members", but I think it means the number of sub-sampled years (hence less than 50 on the x-axis is Figure 8).
The number of the original ESP members is equal to the number of sub-sampled years

10) The introduction states that section 5 summarises and concludes the paper, but section 5 is headed "Discussion". Suggest renaming.
We changed the outline in the introduction to: "Sect. 5 discusses the results."

11) P15 L10 should say in two of the test basins *at lead times greater than X*
We changed this sentence to: "… by 5 to 10% in two of the test basins for lead times greater than 2 months."

12) P15 L13–15: Operational applications should be flexible enough to adapt to different methods if there's a proven benefit. So this argument doesn't carry a lot of weight.
Operational applications typically require a coherent seasonal forecast over the entire basin. A separate calibration per sub-basin may affect the spatial correlations between the sub-basins.

13) P16 L13–14: I'm confused by this. PDO was apparently investigated already in this study and disregarded.
PDO was considered, but MEI was found to have a better overall correlation with the streamflows in the three sub-basins (see Sect. 3.2). Therefore, MEI was used in the single index example application, but PDO would be the first candidate in an extension to multivariate conditioning.

Technical corrections (typing errors, etc.):
14) Figure 2 and Table 1. Abbreviations do not match for Hungry Horse and Libby Dam.
Agreed, changed accordingly.

15) P12 L19 and elsewhere: Text refers to June flow instead of May–June flow.
Agreed, changed accordingly.

16) There are some instances of weigh and weighing instead of weight and weighting. Will be easy to find and correct.
Two instances found and changed accordingly.

17) Improve the figure quality. Many are blurry.
Agreed, changed accordingly.

18) Is Figure 5 one figure or four? There are four captions.

Changed Figure 5 to one figure and one caption.

References

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. Applic. 15: 19-24, 2008.

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

Richardson, D.S., Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, Q. J. R. Meteor. Soc., 127, 2473-2489, 2001.

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

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, 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)

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 method is proposed to incorporate climate mode information into the 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 forecast skill in sub-basins that are less affected by ENSO. The skill of the ENSO-conditioned ESP is evaluated over 50 years of seasonal ~~reforecasts~~hindcasts of streamflows in three sub-basins of the Columbia River basin in the Pacific Northwest. An improvement in forecast skill ~~up to 10% is found.~~of 5 to 10% is found for two sub-basins. The third sub-basin is less affected by ENSO and shows no improvement in forecast skill.

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 U.S. 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 ~~forcing~~forcings to a ~~conceptual~~ 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; Clark and Hay, 2004; Wood et al., 2005; Wood and Lettenmaier, 2006; Tootle et al., 2007; Abudu et al., 2010; Yuan et al. 2015; Sagarika 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 modeling 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 U.S. 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; Tootle and Piechota, 2006; Lu et al., 2011; Gedalof et al., 2012).

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; ~~an~~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 basin.

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 ~~weigh~~weight the ensemble traces according to the similarity between climate indices in the historical year and the year of forecast (Croley II, 1996, 2003; 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 members, because the number of historical years in each class is obviously less than the original number of ensemble members. 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 members. 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 sub-basins, although the influence of the climate modes may differ per sub-basin. A ~~weighing~~weighting scheme that produces good results for sub-basins that are influenced by a particular climate mode may not perform well for sub-basins that are less affected by this climate mode. The forecast skill for these latter sub-basins 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 sub-basins 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 in sub-basins that are affected by climate modes, while avoiding a loss of skill in other sub-basins. 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, 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. Sect. 4 includes the results obtained applying the method to the study area and a forecast skill assessment relative to the standard ESP. Sect. 5 ~~summarizes and concludes~~discusses the ~~paper~~results.

2 Method

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

2.1 Subsampler procedure

The subsampler procedure is a k-nearest neighbor (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 ~~MAP/MAT~~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 members as is done in this study by using a resampler.

2.2 Resampler procedure

The resampler generates new ensemble members 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 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 time scale. It is assumed that the resampled time series represent realistic and equally likely representations of future weather 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:

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

   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. The new reference date is set to the selected date in the historical year plus one month and a new search (step 2) is started.

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 re-selected 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 ~~needs~~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.

~~**3 Study area and data**~~

3 Example Application

3.1 Study area

As a case study, the method was applied to seasonal streamflow forecasting at three ~~projects~~locations (dams) on Columbia River tributaries in the 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 snow melt and runoff in spring. Figure 2 shows the ~~average~~median and ~~standard deviation~~variation of the monthly streamflows for the three ~~projects~~locations. The flows are highest and have the most variation in the snow melt season (May-June).

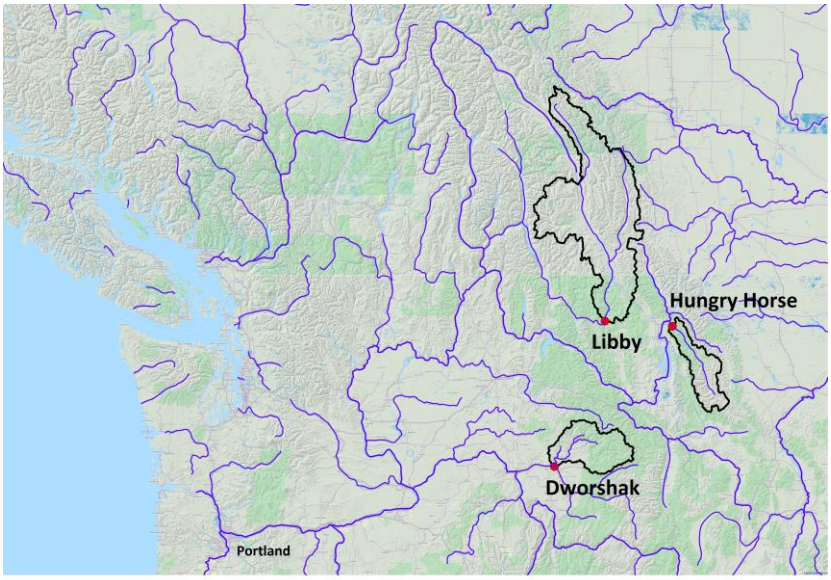

Figure 1: Study area with the three test-sites and extent of sub-basins.

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 projects 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.

Table 1: Case study ~~projects~~locations and sub-basin properties.

~~Project~~ Location	River	Drainage area (km^2)	Mean elevation (m)	Mean flow (m^3/s)	~~Powerhouse capacity (MW)~~Mean annual precip. (mm)	Runoff ratio
Libby ~~Dam (LYD)~~	Kootenay	23,270	811	310	~~600~~851	0.49
Hungry Horse ~~(HHW)~~	Flathead	4,145	239	100	~~428~~1174	0.63
Dworshak ~~(DWR)~~	Clearwater	6,320	363	160	~~400~~1283	0.62

Inserted Cells

Inserted Cells

BPA uses an operational forecasting system called the Community Hydrologic Prediction System (CHPS) with ESP functionality for their seasonal streamflow outlooks (4 to 8 months lead time). The Sacramento Soil Moisture Accounting model (SAC-SMA) (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-hour time step, taking mean areal precipitation (MAP) and mean areal temperature (MAT) per sub-area as inputs. The conceptual sub-basin models were calibrated on 30 years of observational data. Initial (warm) states for the ESP forecasts are generated by running the models in operational mode, continuously blending in recent ~~gauge data of e.g.~~ snow pack and streamflow gauge data into ~~the~~model states.

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/).

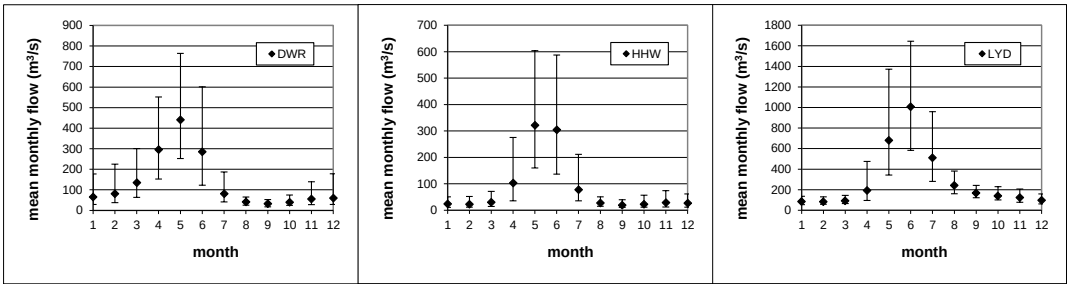

Figure 2: ~~Mean~~Median monthly streamflow and ~~standard deviation~~10% and 90% percentiles for the test-basins Dworshak (DWR), Hungry Horse (~~HGH~~HHW) and Libby ~~Dam (LIB~~(LYD).

3.2 Experimental Setup and Parameter ~~Tuning~~Calibration

Several climate mode indices and combinations of indices for ensemble member selection and conditioning of the subsampler were evaluated, including the Pacific Decadal Oscillation (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 next year. The MEI, as defined by Wolter (1998), showed the highest correlation with the historical streamflows in three sub-basins of interest 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 two-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 months 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 ~~(forecasts in retrospect~~) were produced by running the hydrologic models, taking the

subsampled and resampled MAP and MAT series as input. The year of ~~reforecast~~hindcast was excluded from the subsampling and resampling schemes.

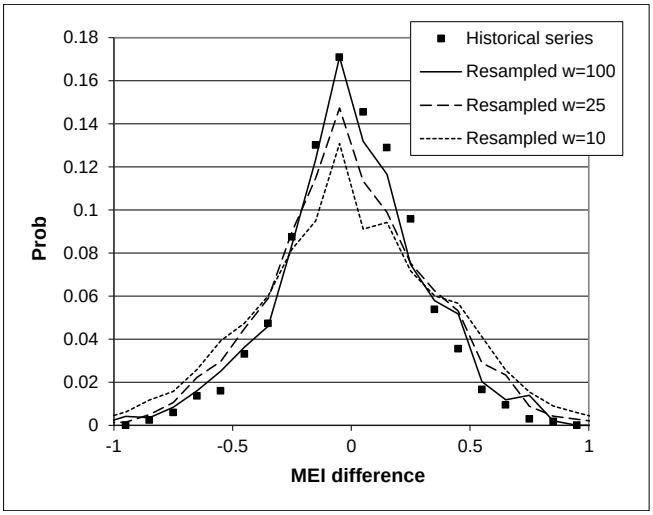

Figure 3: Distribution of ENSO-MEI differences between consecutive months; historical series and three resampled time series with w-values of 10, 25 and 100.

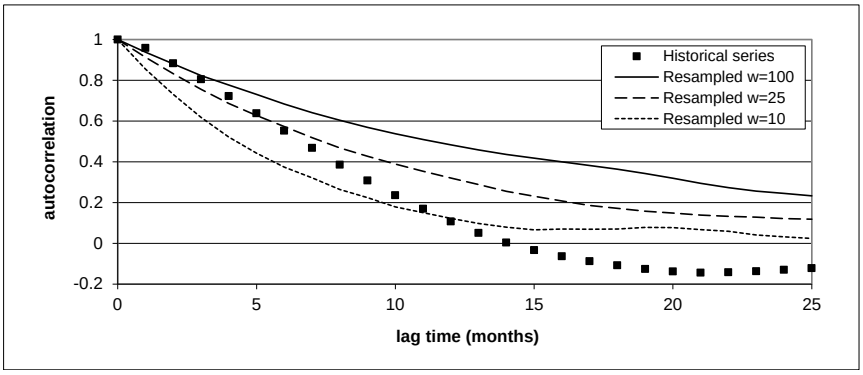

Figure 4: Autocorrelation of ENSO-MEI signal for the historical and three resampled time series with w-values of 10, 25 and 100.

Figure 5 shows example ~~reforecasts~~hindcasts of (~~a~~from top to bottom) climate index, ~~(b)~~ monthly mean areal precipitation~~,~~ ~~(c)~~ (MAP), monthly mean areal temperature (MAT) and ~~(d)~~ monthly mean streamflow ensembles, starting from reference dates December 1st 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.

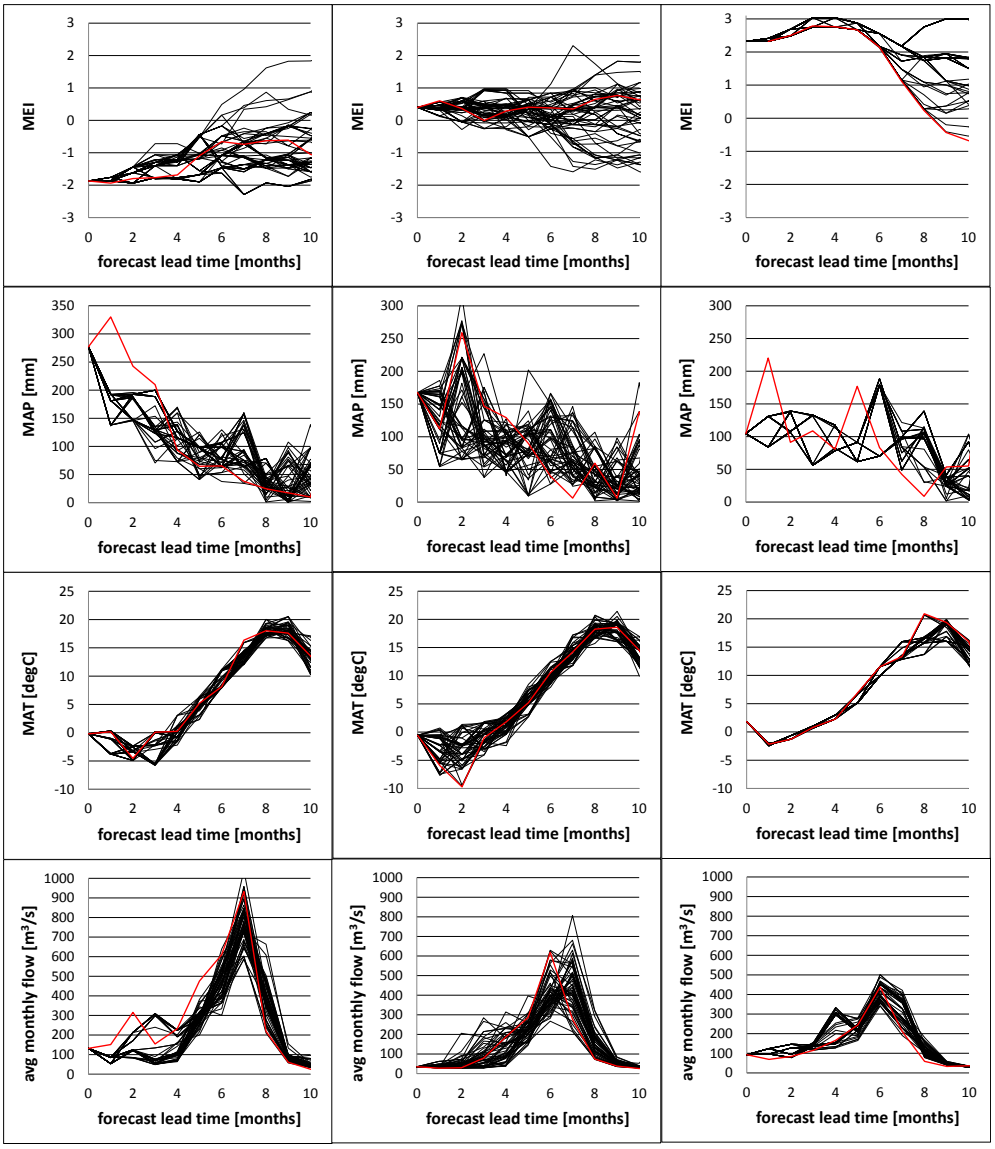

Figure 5: Resampled ensemble forecasts of (from top to bottom) MEI, MAP, MAT and streamflow at test location Dworshak. Forecast dates are December 1st 1973 (left), 1978 (middle) and 1997 (right). The historical runs are shown in red.

Figure 6 shows the number of unique ensemble members 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 re-sampled 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.

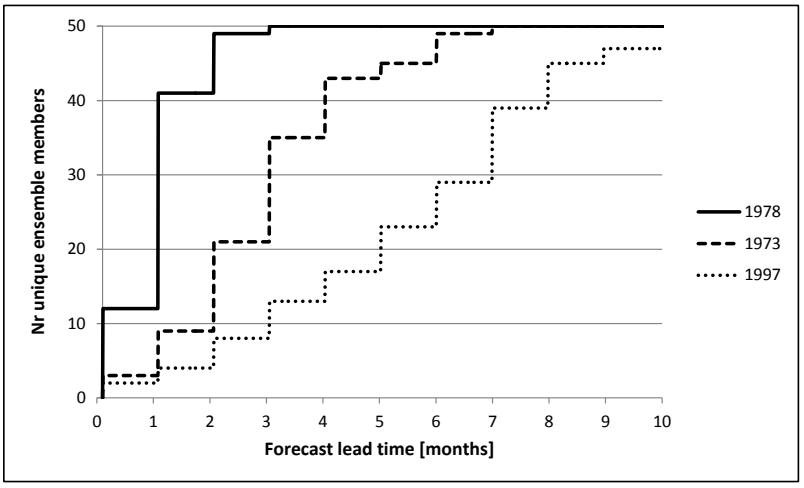

Figure 6: The number of unique ensemble traces in 50-member ensembles of resampled time series (w=25), starting from December 1st, 1978, 1973 and 1997.

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 sub-basin.

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 Score (BSS):

$$BSS = 1 - \frac{BS_{model}}{BS_{reference}} \quad (2)$$

Where the $BS_{reference}$ is the Brier Score of the standard ESP method. 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 ~~most~~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, CPRSS and relative reduction in RMSE of the resampler method in red as a function of number of ESP ensemble members. Skill scores reported here refer to May and June ~~flow~~monthly flows and are averaged over forecast lead times between ~~1~~3 and 12 months. The 50-year ensemble is identical to the original ESP and has a ~~BSS~~skill score of 0~~.~~ by definition. Upon reducing the ensemble size, the ~~BSS~~forecast skill increases for two of the three sub-basins (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.

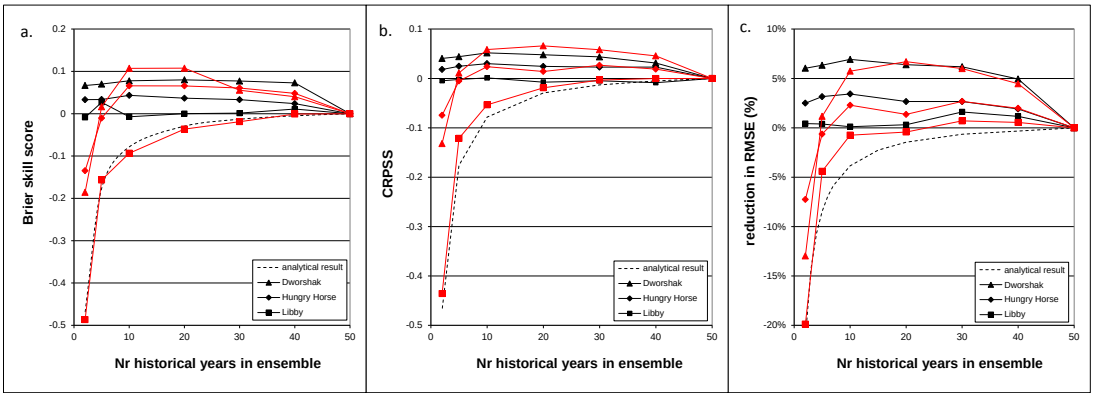

Figure 7: ~~Subsampler~~Forecast skill of the subsampler method~~: Brier Skill Score (80% threshold) of May-June streamflow forecasts~~ (in red) compared to subsampler-resampler method (in black) as a function of number of historical ESP ensemble members. a. Brier Skill Score (80% threshold) b. CPRSS 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.

For one sub-basin (Libby ~~Dam~~), the ~~BSS~~forecast skill decreases for smaller ensemble sizes. The reduction of the number of ensemble members has ~~a negative~~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 ~~line represents~~lines represent the general behaviour of the ~~BSS~~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), Eqn. 22 and Ho et al. (2013), Eqn. 1 respectively. ~~Streamflows~~The streamflow at Libby ~~Dam have~~has the weakest correlation with MEI. Apparently, the MEI information has little additional value for the Libby ~~Dam~~ streamflow forecasts and their skill follows this trend. For the other two sub-basins the ~~BSS~~skill also drops below zero for ensemble sizes less than 10.

~~Figure 8 shows~~Next, the ~~BSS for forecasts from~~forecast skill of the combined subsampler-resampler method~~, where the~~ was computed (black lines in Figure 7). The ensemble members that were dismissed in the subsampler are now replaced by resampled traces. The ensemble size is thus 50 in all cases ~~and the BSS is~~. The forecast skill is still a function of the number of original ESP members (full historical years~~). In~~), but in contrast to ~~Fig. 7, the BSS for all test basins are now~~ the subsampler forecasts, the subsampler-resampler produces a generally positive skill over the full range. The marginal loss of skill for Libby is attributed to statistical uncertainty of the skill score calculation. This demonstrates that the loss of skill from the reduction of ensemble size can be neutralized by additional ensemble traces from the resampler method. ~~The improvement of skill in terms of RMSE and CRPS was also investigated and found in agreement with the Brier skill score (results not shown).~~ A mix of 10 historical years from the subsampler ESP and 40 additional resampled traces produces in general the best result for ~~these~~the three sub-basins 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 ~~Dworshak sub-basin. The other two~~three test basins ~~show similar but less pronounced behaviour (results not shown).~~ 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.

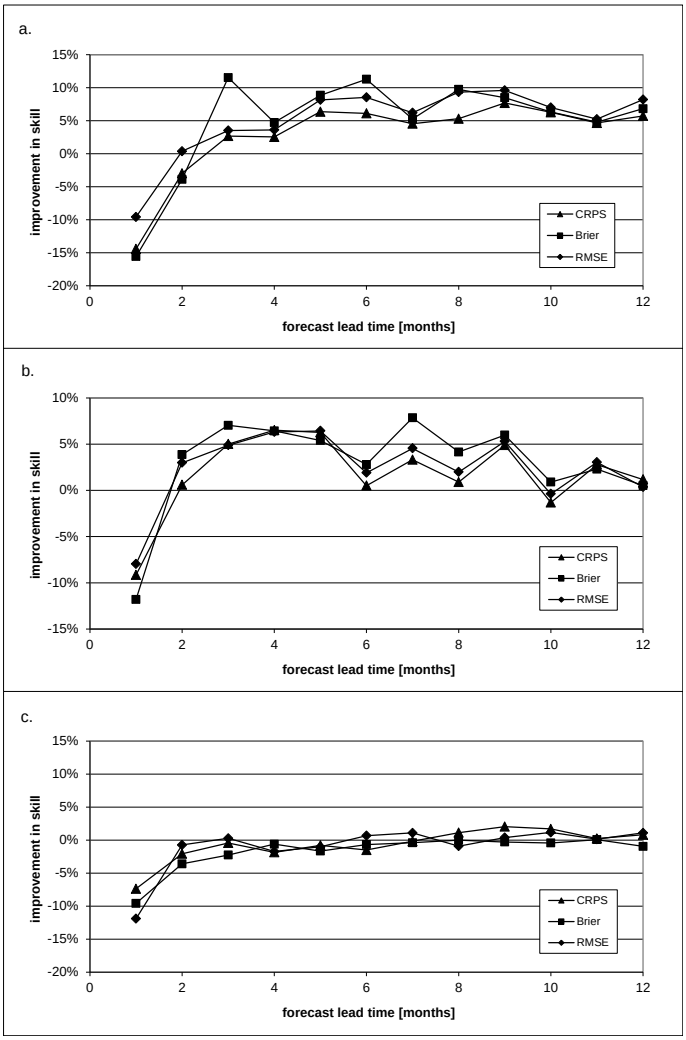

Figure ~~9~~8: Improvement in May-/June streamflow forecast skill of the subsampler-resampler (w=25) method relative to the standard ESP as a function of lead time for ~~DWR test site, three~~: a: Dworshak, b: Hungry Horse, c: Libby test sites. Three different skill metrics: CRPS, Brier Score and RMSE.

Figure ~~9~~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. 3.1. In order to maintain climate mode information on the seasonal time scale, 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 members is 50, a small subset of historical years keep re-appearing 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).

5 Discussion

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 basins.~~sub-basins 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 sub-basin. Werner et al. (2004) found a much smaller skill improvement of 4 and 6% for two other sub-basins, which is comparable to the results found in this study. Moreover, Werner et al. (2004) used a separate calibration of post-processing parameters per sub-basin. Many operational applications require equally weighted ensembles for all sub-basins in the area of interest. This requirement does not allow for a per-sub-basin optimization.

For the third sub-basin in our case study, Libby~~ Dam~~, no improvement of skill was found. The streamflows in this sub-basin have the ~~lowest~~weakest correlation with MEI and the local climate is least affected by ENSO. It was shown that dismissing ensemble members from the ESP leads to a reduction of forecast skill for this sub-basin ~~because of~~ that is similar to the ~~degradation of statistical properties of the~~ expected reduction for a randomly reduced ensemble~~. However, the~~, as described by Richardson (2001), Ferro (2007) and Ferro et al. (2008). The same effect occurs for the other two locations for very small numbers of sub-samples. An ensemble of fewer members 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~~. The~~ and the adverse effect of the dismissal of ensemble traces by the subsampler is neutralized by the resampler 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 sub-basins.

The subsampler-resampler method also has some practical advantages over alternative approaches. Firstly, the subsampler-resampler produces an equal-likelihoods 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 members. 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 need not 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. This 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.

5 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.

References

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

Anderson, E.A.: A Point Energy and Mass Balance Model of a Snow Cover', NOAA Technical Report NWS 19, 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(4), 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. Sc., 2, 195-209, 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(7), 854-872, 2010.

Burnash, R.J.C., Ferral, R.L. and McGuire, R.A.: A Generalized Streamflow Simulation System - Conceptual Modeling for Digital Computers ,U.S. Department of Commerce, National Weather Service and State of California, Dept. of Water Resources, 1973.

Burnash, R.J.C.: The NWS River Forecast System - catchment modeling. In: Singh, V. P. (Ed.). Computer Models of Watershed Hydrology, 311-366, 1995.

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

Croley II, T. E.: Using NOAA's new climate outlooks in operational hydrology, J. of Hydrol. Eng., 1 (3), 93-102, 1996.

Croley II, T. E.: Mixing probabilistic meteorology outlooks in operational hydrology, J. of Hydrol. Eng., 2 (4), 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, 2000.

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

Druce, D.J.: Insights from a history of seasonal inflow forecasting with a conceptual hydrologic model. J. Hydrology. 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. Applic. 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. Hydrometeor, 4, 1105–1118, 2003.

Gedalof, Z., Peterson, D.L. and Mantua, N.J.: Columbia River flow and drought since 1750, J. of the Am. Water Res. Ass. 40 (6), 1579-1592, 2012.

Halpert, M.S. and Ropelewski, C.F~~,~~.: Surface temperature patterns associated with the southern oscillation. J. Clim. 5:577–593, 1992.
Halpert, M.S. and Ropelewski, C.F~~,~~.: Surface temperature patterns associated with the southern oscillation. J. Clim. 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. Plan. Manag., 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. of the Am. Water Res. Ass., 45 (3), 705 580-596, 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 (8), 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. Let. 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. Clim 2(1):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 (11), 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, D04114, doi:10.1029/2008JD010969, 2009.

Lu, A., Jia, S., Zhu, V., 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, Hydrology and Earth System Sciences, 15 (4), 1273–1281, 2011.

Madadgar, S., Moradkhani, H. and Garen, D~~,~~.: Towards improved reliability and reduced uncertainty of hydrologic ensemble forecasts using multivariate postprocessing, Hydrol. Process., 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, Bull. 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. Hydromet. 3(1):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 (3), 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. Ac. Press. ISBN 0 12 553235 0. Int. Geophys. Ser. Vol. 46. 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, 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 (5), 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. of Hydrology, 201 (1-4), 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(9), 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. R. Meteor. 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.

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. of Hydrol., 391 (1-2), 11-25, 2010.

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

Tootle, G.~~A. and Piechota, T.C.,~~ 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., 442–451, 2007.

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

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

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

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. Wea. Rev. 135, 2778-2785, 2007.

Weijs, S.V. and van de Giesen, N., An information-theoretical perspective on weighted ensemble forecasts. J. of Hydrol. 498 (2013) 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. Hydrometeor., 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. Modell. Softw., 40, 65–77, 2013.

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(D20), 4429, doi:10.1029/2001JD000659, 2002.

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. of Geophys. Res. Atm., 110 (D4), doi:10.1029/2004jd004508, 2005.

Wood, A.W. and Lettenmaier, D.P., A testbed for new seasonal hydrologic forecasting approaches in the western US, B. Am. Meteorol. Soc., 87(12), 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(14), L14401, doi:10.1029/2008GL034648, 2008.

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 2015. doi: 10.1002/wat2.1088, 2015.