Interactive comment on "ENSO-Conditioned Weather Resampling Method for Seasonal Ensemble Streamflow Prediction" by J. V. L. Beckers et al.

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. This is explained in Sect. 3.2. The MEI had the highest correlation with the streamflow data and was therefore selected.
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 will change 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 will be replaced with 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 will change 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 will change 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

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

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

17) Improve the figure quality. Many are blurry.
Agreed

18) Is Figure 5 one figure or four? There are four captions.
Figure 5 will be change into 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.