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

Anonymous Referee #1
Received and published: 7 March 2016

Summary:

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 will be 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 remove 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 will change '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 will add 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.
Will be corrected.

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

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 add average elevation, mean runoff, precipitation and runoff ratio and remove 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.
Agreed. We will provide better quality figures.

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 will be combined, as suggested and two additional panels with CRPS and RMSE results will be 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 will be 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 will be 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 rephrase 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
17. P2, L27: “case study” -> “case study basin”. Agreed
18. P2, L26: “weigh” -> “weight”. Agreed
19. P3, L19-21: “Sect.” -> “Section”. We thought this is HESS-style
20. P5, L13: “needs” -> “need”. Agreed
21. P7, L9: “of e.g.” -> “with”; “into the states” -> “into model states”. Rephrase to: ‘… blending in recent snow pack and streamflow gauge data into model states’
22. P8, L1: “parameter tuning” -> “parameter calibration”. Agreed
23. P12, L18: “the most variation” -> “the largest variation”. Agreed