Editor's remarks:

- Page 2, line32-33: please, consider checking the expression "an improved specificity of the ensemble forecast". It is not fully clear to me what you mean by that.

'Specificity' is the term that is used by Hamlet and Lettenmaier (1999) for 'the relative size of the area between the upper and lower bound of the ensemble'. We prefer to use the same terminology when referring to their results.

- Page 4, line 10: "loosely based on…" Can you be more precise about the differences between the original proposal and your adaptation of the method referenced? As it is, the sentence is very vague and not much informative for the reader.

We added: "… loosely based on a method for daily rainfall resampling developed by Brandsma and Buishand (1998)". This should make clear that our resampling technique is similar to the method referenced, but the sampling interval (daily vs monthly) and the variable (rainfall vs climate index) are different.

- Page 4, lines 15-16: please, check the use of the English language in this part of the sentence: "... represent realistic and equally likely representations of…"

Changed to: "It is assumed that the resampled time series are realistic representations of future weather patterns and that they are equally likely to occur as the full historical years in the original ESP."

- Page 4, line 30-31: If I understand well, the year of reference (which is the year associated with the time of forecast) is used in the re-sample. This would however not be feasible operationally (as one would not have the data available for the time steps after the time of forecast). I am however not sure I understood it correctly because on page 8, line 24, you mention that "The year of hindcast was excluded…". Could you please check these two sentences and make then clearer for the reader?

The reference date is not the time of forecast/hindcast. The reference year changes with every resampling round, except when the same year is resampled twice. We rephrased the description of step 4 in the stepwise description of the algorithm and added an example:

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

   …

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

- Page 7, line 3: when you mention CHPS, are you making reference to the CHPS used by NOAA too? http://www.nws.noaa.gov/ohd/hrl/chps/ The way it is mentioned in the paper, it makes the reader think that CHPS is a specific system developed at BPA. Maybe if you could add a reference here, the ambiguity would disappear.

The NWS-CHPS and BPA-CHPS use the same application software (CHPS) but they are not the same system. The implementation of CHPS at BPA is specific to BPA. We rephrased the sentence and added a reference to a paper that describes the CHPS application. Hope this makes it more clear.

- Page 7, line 8: can you explicitly indicate what the period (years) is used for calibration?

Done

- Page 8, line 5: consider changing "in the next year" to "in the following year"

Done

- Page 8, line 6: consider changing "in three sub-basins of interest" to "in the three sub-basins of this study"

Done

- Page 8, line 10: considering the period 1871-2013, is it a stationary period? Aren't there trends observed during this period? How does non-stationarity impact your study and method? I think that maybe a sentence or two about the dependence (or not) on the hypothesis of stationarity of historic time series used in the methodology should be added to the paper (for instance, in the "Discussion" section).

Stationarity is assumed in standard ESP, which may indeed be an issue for some applications. However, it is not the aim of our method to address or resolve this. Our method aims to improve the ESP forecast skill by taking into account climate mode information, not climate change.

Furthermore, the period 1871-2013 is only used to calculate the graphs in Figs 3 and 4. The ESP and resampling/subsampling method use data from 1949-2003. Non-stationarity should be less of a problem for this shorter period.

- Figure 3 and Figure 4: I guess these results pertain to a specific sub-basin of the study area. Is that right? In that case, could you indicate that in the caption? Also what is the difference between "ENSO-MEI difference", "ENSO-MEI signal" and "MEI difference"? You use these in the figures. If they refer to the same thing, maybe it would be better to choose one unique term to refer to the same thing. Please check terminology (see also remark below)

MEI is a global indicator. It is not bound to a specific region. We changed "ENSO-MEI" into "MEI".

- Captions of Figure 1, Figure 2, and Figure 5: you use three different terminologies for the same object: "test-sites", "test-basins" and "test location". Could you please check throughout the paper and make sure that you use the same terminology for the same objects that you make reference to? This makes it much easier for a reader to follow the reasoning and the results.

We replaced 'test-sites', 'test/sub-basins' and 'test locations' by 'forecasting stations' throughout the text.

- Page 9, line 7 and Page 11, line 1: please, consider mentioning also in the text the sub-basin to which Figure 5 and Figure 6 refer to. This information should also be added in the caption of Figure 6.

Added ‘at forecasting station Dworshak’ to the text. Figure 6 is not specific for any forecasting station. The subsampling/resampling is done for all stations collectively.

- Figure 6: consider changing the caption to “Number of…”. Also, please, check for the use of a unique term here too: “ensemble members” or “ensemble traces”. It is better to choose one and stick to it all over the text.

Changed the caption to “Number of…”.

Changed ‘member’ to ‘trace’ throughout the text.

- Page 11, line 19: check the use of punctuation (comma) here.

Removed the comma.

- Page 11-12: explain what you call “relative reduction in RMSE” (introduced on Page 12, line 7) already in the sub-section 3.3.

Added a sentence to section 3.3: “Likewise, the Continuous Ranked Probability Skill Score (CRPSS) and the relative improvement in RMSE are evaluated.”

- Page 12, lines 6-8: check the use of the English language. I think some words (ex., “the”) are missing here.

Added “ … as a function of the number of ESP ensemble members.”

- Page 13, line 9: are you sure it is “… attributed to statistical uncertainty of the skill score calculation”? How did you compute that? Otherwise, please moderate it with “probably” or “might be due to”.

Changed to “…probably due to …” and added “(which could be verified by bootstrapping, but this is left for future studies).”

- Page 15, line 2: I think you mean “Sect. 3.2” instead of “Sec. 3.1”. Please, check.

Indeed. Corrected

- Page 15, line 10: consider changing to “… in two of the three sub-basins …”

Done

- Page 15, line 14: “… used a separate calibration of post-processing parameters per sub-basin”. I do not understand this sentence. Please check to make it clearer.

Changed to: “Werner et al. (2004) used a separate calibration of post-processing parameters to arrive at a different set of weights for each test station.”

- Page 15, line 21: consider changing to “…for very small sub-samples.”

Done

- Page 15, line 32: consider changing to “…as input do not need to be updated. Finally,…”

Done

- Page 16, lines 2-3: Is “water demand” stationary over historic periods? Can we use “historic data” to infer water demand of today? I think you could discuss some limitations also of using historic data for today’s inferences. Water demand may have changed with societal changes (for instance, with growing concerns towards water economy). Again, I think that “stationarity” issues and implications to the methods presented would deserve a few lines in the discussion section.

I can imagine that BPA makes a correction of historical water demand to today’s circumstances, but that is beyond the scope of our study. We simply note that parallel sampling of secondary variables is common practice in operational settings and that our method accommodates for this, in contrast to some other methods.

- Page 15-16: Discussion section: I miss a discussion on the issues of cross-validation, i.e., taking some years out from calibration and then using them for validation. Could you add a sentence or two in the discussion section about it? In your opinion, how it may play a role and be considered in further studies?

There is hardly any calibration needed in our method. The only parameter that is calibrated on hindcast results is the number of ESP sub-samples. Figure 7 shows that a positive forecast skill is found for a range of choices for this parameter (for Dworshak and Hungry Horse), except for less than 10 sub-samples. Based on Figure 7, a value of 10 subsampled members is chosen for subsequent experiments, but larger values also produce a positive skill.

A bootstrapping procedure could provide insight into the uncertainty of the verification scores. We added a suggestion to apply a bootstrapping to assess the uncertainty in a future study on Page 14, lines 11-12.

- Finally, I think it is strange to have a paper without a “Conclusion” section. Could you add a section with the main conclusions of the study for the sake of completeness of the paper and clarity of the message and highlights of your results? Maybe it could be derived from a re-organization of the “Discussion” section.

We renamed the last section to “Discussion and conclusions”.

Reviewer#2 remarks:

1. P2, L2-5: The papers by Tootle, Abudu and Sagarika should not be cited here, since they do not relate to GCM-based seasonal forecasting. Instead, these should be included in L15.

Done

2. Section 2.2: The authors may agree or not on this, but after reading this section several times, I still find difficult to understand how the resampler technique works. In my opinion, a comprehensive diagram would help a lot to clarify the method, and therefore help other readers to reproduce it.

We tried to clarify this by including an example in step 4: historical year selection and setting of a new reference date. Hope this helps.

3. Figure 1: I encourage the authors to add latitude, longitude and a scale bar.

Done

4. Table 1: Can the authors please add in the caption the period used to compute mean flow, precipitation and runoff ratio?

Done

5. P13, L8-9: The authors state that "the marginal loss of skill for Libby is attributed to statistical uncertainty of the skill score calculation". Can the authors clarify what does this mean? So far, no uncertainty (using bootstrapping, for instance) has been considered in the calculation of any score, so this statement seems misleading. Perhaps delete?

Changed to "…probably due to …" and added "(which could be verified by bootstrapping, but this is left for future studies)."

6. P15, L2: "as discussed in Sect. 3.1". Should it be Sect. 3.2?

Indeed, corrected.

7. P16, L18: Can the authors please clarify what they mean with "introducing a random time shift"?

Added an example for clarification: 'For example, instead of sampling a historical period April 1 – April 30, we shift 5 days back and sample March 27 – April 25. '

Suggested minor edits

8. P1 L12: "a number of… are" -> "a number of… is" (noun is singular, not plural).

Not adopted. From the dictionary: "Although the expression 'a number' is strictly singular, the phrase 'a number of' is used with plural nouns (as what grammarians call a determiner (or determiner)). The verb should therefore be plural."

9. P1 L17: "in the Pacific Northwest" -> "in the U.S. Pacific Northwest" (same in P5, L20).

Done

10. P3 L9: "latter" -> "later".

Not adopted. From the dictionary:
"Latter: relating to the second of two groups or things mentioned. "
"Later: at some time subsequent to a given time."

11. P3 L16: "In a pre-processing" -> "In a pre-processing step".

Done

12. P6 L4: Add a comma after "Oregon".

Done

13. P15 L24: “resampler” -> “resampled”.

Done