Review of the paper entitled "Skill of a global forecasting system in seasonal ensemble streamflow prediction" by Candogan Yossef et al., submitted to HESS in October 2016

This manuscript evaluates the benefits of using ECMWF S3 bias-corrected seasonal ensemble forecasts for predicting high and low monthly flows compared with the Ensemble Streamflow Prediction (ESP) flow ensembles for 20 large river basins in all continents. The evaluation of the quality of ensemble forecasts is based on the Brier Score (BS) and the Brier Skill Score (BSS) for 2 binary events defined by the 75% and 25% probability thresholds, for forecast lead times up to 6 months. The relative contribution of the meteorological forcing uncertainty and the total contribution of both the meteorological and hydrological uncertainties are evaluated by verifying with simulated flow and observed flow. This is an interesting research topic area since probabilistic seasonal forecasts could potentially support various critical applications of global hydrometeorological ensemble prediction systems.

The paper is clearly organized and generally well written, although the terminology regarding forecast verification should be improved. It includes appropriate references and substantial evaluation results. However the process to define the benchmark forecasts in this study and the verification scores being used need to be clarified. The presentation of the evaluation results in the different figures and tables needs to be more synthetic and more easily understandable.

I recommend this paper to be published after the authors have addressed the following general comments and specific comments to help improve the quality of the manuscript.

General comments

The verification terminology used in the paper should be improved. When referring to forecast *skill,* in most part of the paper, the authors mean forecast *quality*, one aspect of the forecast quality being the forecast skill when using a given verification metric and a specific reference forecast. In this paper, the authors should refer to the evaluation of the forecast *accuracy* with one metric, the Brier Score (which does include different aspects of the forecast quality attributes, see the decomposition of BS for example). Then the authors used its associated skill score, the BSS, the ESP flow forecasts being the reference forecast, to evaluate the gain *in terms of the Brier Score* by integrating the seasonal forecasts.

Also the terminology used by the authors regarding the “theoretical skill” vs. the “actual skill” of the flow forecasts is not widely used in the literature and may be misleading (again, the *skill* term should only be used when a benchmark forecast is defined). It needs to be clarified in the paper (and in the interpretation of the results) that, when comparing forecast flows with simulated flows, the hydrological errors are cancelled out, which leads to assess the contribution of the forcing uncertainty only to the forecast flows; the verification with the observed flows leads to assess the total contribution of the forcing uncertainty and the hydrologic uncertainty. It may be more appropriate to refer to MF uncertainty vs. total uncertainty in the BS notations and interpretation of results. The equation 2 on page 7 should be corrected (BS_{theo}*100/BS_{act}) and the obtained ratio results should be discussed in terms of the relative contribution of these 2 sources of uncertainty.

Besides, the reforecasting process and the benchmark forecasts used in this study need to be clarified. As the authors mentioned, one has to first run the hydrological model with the observed forcing values during a spin up time period to define appropriate initial states to then start reforecasting. This spin up period should be excluded from the analysis of the reforecast dataset. In the paper, it is not clear whether the spin up period covers the period of 1979-1984 (p. 5) or 1979-1980 (p. 6) and has been excluded from the forecast dataset for the verification analysis. The authors then integrated retroactively the ECMWF S3 seasonal forecasts and the historical observed forcing values to produce, respectively, the S3-based flow forecasts and the ESP flow forecasts from 1981 to 2010. In the ESP flow ensembles, based on observed forcing values from all years in 1981-2010, one member corresponds to the simulated flow (e.g., the run with forcings from 1981 initiated in 1981). This member should not be part of the ESP flow ensembles since, in real-time forecasting, all ESP ensembles use past historical years of forcing as possible future outcomes (considering that the climate is stationary and repeated itself). Including the simulated flow in the ESP ensembles will lead to artificially increase the accuracy of these forecasts in the BS values, which will then decrease the skill of the S3-based flow forecasts in the BSS values. It seems that the authors did include the simulated flow (or control run) in the ESP members, which may explain why there is almost no gain in using the seasonal forecasts. If they didn't include that member in the ESP ensembles, this should be clarified in the paper.

The evaluation results are presented for 20 basins, using too many tables (10 tables for each basin). The BS values do not seem to be essential for the evaluation study since the main point of the authors concerns the benefits of seasonal forecasts compared to ESP. Having a common benchmark with the BSS score for all 20 basins makes it easier to compare the results among the different river basins. The authors could include only the BSS results (even if the BS values are mentioned in the text for specific aspects), as well as the ratio describing the relative contribution of the 2 sources of uncertainty. Also, all the tables could be turned into grid figures, using a color scale with more color categories than the ones currently used in the tables, to facilitate the interpretation of the results. To further reduce the number of figures, some of them could be included in appendices if the results for basins from similar climatic zones are the same.

The authors need to include more information about the selection of the 20 test basins, with a table describing the basins in terms of basin size, average flow, and a corresponding map of the rivers and outlet locations with names (see the material included in Candogan Yossef et al. 2012). Since the authors referred quite extensively to 2 past studies with the same hydrologic model and the analysis of the forcing uncertainty and initial conditions uncertainty (Candogan Yossef et al. 2012 and 2013), it would be necessary to mention whether the same 20 basins were used in all these studies. The impact of flow regulations needs also to be discussed here since this impact could dominate the hydrologic uncertainty as well as the forcing uncertainty (cf. discussion in the 2012 paper).

Specific comments

- Abstract, p. 1 line 12: please specify that the PCR-GLOBWB hydrologic model is distributed (maybe better to spell out the acronym of the model name). It would be better to mention “ensemble reforecasts” (as the forecasts are produced retroactively) and indicate the reforecast period.
- Abstract, p. 1 line 15: please remove “the skill from” (see general comment about the use of the term “skill”).
- Abstract, p.1 line 18: please change “skill” to “forecast accuracy”.
- Abstract, p. 1 line 19-22: consider clarifying that the analysis concerns the relative contribution of the forcing uncertainty and the hydrologic uncertainty to flow predictions when verifying with both simulated flow and observed flow (see general comment).
- Page 3, lines 2-6: please consider including more specifics about the 2012 study (period of evaluation, test basins, simulated and/or observed flows used for evaluation, possible differences with the presented work) since the authors referred to the study results quite extensively in the paper (see general comment).
- Page 3, line 18: consider adding a short description of the ESP and reverse ESP approach since the conclusions about the relative importance of the forcing uncertainty and the initial conditions uncertainty is one of the main points of discussion in this paper and results from the 2013 study are mentioned quite extensively in the results section.
- Page 4, line 27: clarify what DDM30 is (source of dataset? Please spell out the acronym).
- Section 2.2, page 5: please specify the spatial and temporal resolution of the forcings for both the observation/reanalysis and reforecast datasets (spell out the acronyms only if necessary; WCRP is not needed); is there any change of spatial resolution from the original dataset to the forcing inputs for the reforecast runs?
- Page 5, line 14: please explain why the authors used the ECMWF S3 seasonal forecasts when the S4 forecasts are operational since November 2011 (and being evaluated in the GLOWASIS project); how the S3 and S4 seasonal forecast datasets compare to each other (number of members; is the same bias correction procedure also included in the S4 forecasts?); see also the comment for the conclusion section.
- Section 2.3, page 5, starting at line 35: please clarify the reforecast process, which starts with the spin up run of the hydrologic model and needs to exclude the spin up period (1979-1980 to start reforecasting in 1981?) from the reforecasting and verification period (see general comment).
- Page 6, lines 5-9: please clarify the ESP members being used, excluding the member corresponding to the control run (see general comment).
- Section 2.4, page 6: please consider adding a reference to the forecast accuracy when using the Brier Score, and the forecast skill when using the BSS (see general comment).
- Page 6, line 21: please specify that GRDC is the source of the flow observations (spell out the acronym).

- Page 6, lines 24-29: please include that the BS is the mean squared error of probabilistic forecasts for a given dichotomous event; a probability threshold is used to define the binary event to be observed and forecasted. The authors should point out that the BS is a relevant metric for analyzing the performance of a forecast system for specific categories of flow (in this case, with a high flow threshold and a low flow threshold). The authors should clarify why they selected the 75% and 25% flow thresholds (user requirements? large enough sample sizes?)
- Page 6, lines 30-35: please clarify how the threshold values for the BS computation are defined using the simulated flow values vs. forecast flow values (using which forcing forecasts?).
- Page 7, lines 19-28: please refer to the relative contribution of the forcing uncertainty and the hydrologic uncertainty to flow predictions when verifying with both simulated flow and observed flow (see general comment). Please correct equation 2 and comment on the variations of the ratio (what if the denominator has a value close to 0?).
- Section 3.1, page 8: please see the general comment about including only figures with the BSS results (not the BS values) with a smaller number of color-coded figures.
- Page 10, lines 14-15: clarify what the authors mean by “the skill of the ESP is below the climatology” (referring to the unconditional climatological record of observed flow?).
- Section 5, Conclusion, page 13: please comment on the potential use of the S4 seasonal forecasts in a similar study and the potential gain in forecast quality due to the seasonal system enhancements. The authors should also mention that the use of a single verification metric (the BS and its associated skill score) could be complement by other verification metrics, such as the ROC score (to characterize the event discrimination of the forecasts) and the BS decomposition (to evaluate in more details the conditional and unconditional biases in the forecasts). The sampling uncertainty of the verification metrics should also be evaluated (for example with a bootstrapping technique), especially if the verification analysis is also conducted with higher probability thresholds for the BS computation. Finally the authors could include some comments about the user requirements for seasonal probabilistic flow forecast systems (and the evaluation the system performance) and collaborations between forecasters and end users (for example in the GLOWASIS project) to further improve the usefulness of such systems.
- Figure 1, pages 38-39: please clarify what the white circles mean; it would be better to use more different colors in the color scale. Please refer to the BSS and the ESP reference forecasts in the legend (forecast skill being specific to a given verification metric and a benchmark).