Response to Reviewer#3

The authors want to thank Reviewer#3 for the valuable comments, which will help us to enhance our paper. We provide below our answers to the comments.

Reviewer 3

This paper deals with an interesting topic, the effect of bias corrections on the quality of seasonal streamflow forecasts. It is mostly well written and the overall structure of the paper is good. However, I found the following issues that need revision before I could recommend the paper to be published.

Reviewer's comment (RC):
Main points:
1) My most important point is that the paper is too long. I suggest to set a hard (!) reduction requirement of at least 25% (number of words). It is up to authors to decide which parts they remove or shorten. Just a few suggestions from my side: discuss fewer bias correction methods, remove almost completely page 13 line 3 – page 14 line 7, remove third and fourth sentence of section 3.2.1.

Authors' reply (AR): Reviewer 2 and the editor also recommended shortening the paper. This issue will be addressed when producing the revised version.

RC: 2) In the paper sharpness is discussed with the assumption that quality increases with sharpness. Mason and Stephenson (2008) write that "in the extreme case of no predictability, the forecast probability should always be equal to the climatological probability". So, forecasts can be too sharp, which should be a conclusion from e.g. Figure 2, where sharpness for longer lead times is larger than that of the reference. So, the sharpness results and conclusions should be reconsidered.

AR: We agree with the reviewer that sharpness in itself, as any other forecast quality attribute, is not necessarily an indicator of a perfect forecast. In our study we adopted the paradigm of Gneiting et al. (2007): « maximizing the sharpness of the predictive distributions subject to calibration ». This means that for two systems with equal levels of reliability, the best one is the sharper one (i.e., lower IQR score in our study). The evaluation of sharpness is thus complementary to the evaluation of reliability. That is the reason why we adopted the scores PIT diagram and IQR in our study. We will make sure that this is clear throughout the text in the revised version.

RC: 3) A better (and longer) introduction to PIT diagrams is needed. Since these diagrams are not well explained in the paper, I was not able to understand the PIT results. I suggest at least to write much more clearly how these diagrams are constructed, to show a figure like Figure 2 from Laio and Tamea, to clarify what PIT values (vertical axis of figures in paper) are and to add a text to the horizontal axis of the PIT diagrams displayed in the paper. How does the area in the diagrams measure reliability? Is the area also sensitive to bias? Is that acceptable? In Section 3.3.1. the text mentions "concentration of points" but only lines are shown in the diagrams. So, what do you mean by "concentration of points"?

AR: The probability integral transform (PIT) histogram is used in forecast verification to evaluate if the empirical time series of PIT values (the PIT value is the value that the predictive cumulative distribution function associates with the observation at a given time step) has a uniform distribution (see also, Gneiting et al., 2005 [1], where it is also explained that "uniformity is usually evaluated in an exploratory sense, and one way of doing this is by plotting the empirical cumulative distribution function of the PIT values"). This is what we have done in our paper (Page 7, lines 6-8). In order to compare systems, we also evaluated the score defined as the "PIT area", as proposed in the reference cited in the paper (Renard et al., 2010). The further the PIT curve is from the 1:1 diagonal, the less reliable the ensemble is. Therefore, the smaller the area between the curve and the 1:1 diagonal, the more reliable the ensemble is. The rank histogram or Talagrand diagram, proposed independently in

the literature, is a similar measure. Gneiting et al. (2005) indicate that "If we identify the predictive distribution with the empirical cumulative distribution function of the ensemble values, this technique is seen to be equivalent to plotting a PIT histogram". The visual inspection of the PIT diagram can be a useful assessment (on systematic biases or spread deficiencies, as we mention on page 7, lines 8-13), but forecast deficiencies may still be hidden behind the assessment (deficiencies in sharpness, for instance). That's why we use (and recommend) the joint evaluation of other scores. We hope this clarifies our approach.
The paragraph on page 7, lines 5-15, as well as the x and y axes of the PIT diagrams in Figures 3, 5, 11 and 14 will be revised to clarify the construction and representation of the PIT diagram. We have linked our points with lines for a better visualization of the results of the 16 catchments in a unique PIT diagram, so we will also clarify the term "concentration of points" in the text. We would like, however, to avoid adding a figure that is already presented in another easy-to-access paper that we are referencing (Laio and Tamea, 2007). This is especially important since we also need to reduce the length of the paper.

[1] Probabilistic Forecasts, Calibration and Sharpness, Tilmann Gneiting Fadoua Balabdaoui and Adrian E. Raftery. Available here: https://www.stat.washington.edu/research/reports/2005/tr483.pdf

---

RC: 4) PIT area, MAE and CRPS are all sensitive to bias, as far as I can see. This should be mentioned in Sections 3 and 7 and discussed in Section 7.
AR: These scores may indeed inform on biases, and we will make sure it is clearly mentioned in Sections 3 and 7.

RC: 5) Section 2.2 mentions that observations are used to initialize streamflow. What about the initialization of snow and soil moisture? These form important contributions to predictability.

AR: The GR6J model is a conceptual, reservoir-based hydrological model (Page 4, lines 20-22). Its inputs are daily precipitation and potential evapotranspiration. These data are used to run the model and initialize its states, including the state of its reservoirs, prior to the forecast date. The upper reservoir of the model can be assimilated (although it is not equivalent to, as it is not a physically-based model) to a "soil moisture accounting" reservoir. Therefore, in a sense, this is also initialized. As for snow modelling, it is not represented in the version of the model used in our study. The catchments chosen have little or no snow-influenced runoff (Page 4, lines 17-19). On lines 25-27, page 4, we mention the forecast updating of the model, which is a different procedure from the initialization. After initialization, the model goes through an "updating procedure", common in hydrological forecasting, which, in our case, is based on the last observed discharge. We will check section 2.2 to make the difference clearer in the revised version.

RC: 6) Sections 3.2.1. and 3.2.2. about the bias correction methods need references. EDM and GDM seem to be have strange effects: a specific amount of daily precipitation is corrected differently for different years, depending on the monthly amount of precipitation. What is the motivation to possibly employ these two methods? Perhaps some of the investigated methods should not be considered at all, see point 1 about shortening the paper. I found LS-m and EDMD-m the most interesting methods.
AR: Our motivation is to evaluate if EDM brings additional value regarding LS, notably in correcting bias for extreme precipitation, and whether the use of a fitted distribution (here, GDM) enhances performance or not. We also found LS-m and EDMD-m more interesting, but this comes from the progressive analysis of all the other methods too. We think it is important to show all the methods as they have different levels of complexity. When shortening the length of the paper, however, we will pay attention to check if we can cut some text from this part of the paper.

RC: Minor points:
page 1, line 16: "contributes" instead of "contribute".

page 2, line 7: "widespread use of" instead of "the widespread of"
page 2, line 21: remove "rather than by initial conditions"
page 3, line 13: "varied between" instead of "derived from"
AR: All minor points will be considered in the revised version.

RC: The hydrological model also needs temperature as input to compute potential evapotranspiration. Write clearly how this input is constructed.
AR: The calculation of the evapotranspiration was done following the Oudin formulation. This formulation can be found in Equation (3) of Oudin et al. (2005). It was computed based on the daily temperature from the SAFRAN reanalysis. The reference is cited in the paper. We will make sure that this is clear in the text.

RC: page 3, line 18: add "heavily" before "influenced"
AR: This will be corrected in the revised version.

RC: page 3, line 23: replace "interannual" by "long-term mean". Over which years? On a monthly basis? Also for hindcasts?
AR: For a given day of the year, the estimated PET on this day is the mean of all PET computed for this day of the year, over all available years (with exception for the targeted year). Reviewer 1 and 2 also pointed out that the PET used in the article should be better explained (please, refer to the answers to their reviews). This will be clarified in the revised version.

RC: page 3, section 2.2: motivate why the focus is solely on the influence of precipitation input.
AR: This is a choice we made as we were focusing on catchments with a pluvial-dominated hydrological regime.

RC: page 6, section 3.3: So, do the evaluations for lead week 1 for the winter include all the hindcasts made on December 1, January 1 and February 1? These are then 15 members issued in December and January and 52 members issued in February. How do you deal with this inequality? And do the evaluations for lead week 6 for the winter include all the hindcasts made on November 1, December 1 and January 1? Explain this clearly.
AR: The reviewer's understanding is correct on the way we aggregated the forecast values to compute the evaluation scores for the four seasons. We can thus have seasonal-based scores that involve forecasts with 15 or 51 members. This comes from the data setup of ECMWF. We only handled inequality when comparing ensemble of different sizes with the CRPS (as explained on page 8, line 16-21). Despite the inequality in the seasonal aggregation of scores, we note that this should not impact comparisons between seasons (since all seasons have a month with 51 members), and comparisons between raw and bias corrected forecasts (since aggregation is considered equally in both systems).

RC: page 7, line 8: "coinciding with" instead of "superposed with"
page 7, line 24: "Ranked" instead of "Rank"
AR: These will be corrected in the revised version.

RC: page 8, line 6: What is the observation period?
RC: page 8, line 14: From which period are the observations?
AR: Observed precipitation data were available for the period running from August 1958 to July 2010. Observed streamflow data were available for different time periods, ranging from 51 years to 35 years, according to the catchment. This will be clarified in the revised version.

RC: page 8, line 23: "caused" instead of "brought"

AR: This will be corrected in the revised version.

RC: page 8, line 28: "becomes negative". What is done if there is more than one transition from a positive to a negative score?
AR: If there are several transitions, the lead time of the first transition is considered. We will add "first" before "lead time beyond which" in line 27, page 8, to make this clearer.

RC: page 9, line 28: "this is observed in the majority of catchments". This does not seem to be the case. There is roughly an equal number of curves below and above zero.
AR: We will revise the sentence in the revised version.

RC: page 13, figure 6: I would expect no bias at all in the lower right and left panel. What is the cause of these biases? Are the remaining biases caused by the one-year-leave-out method? If so, I would expect them to vary randomly around zero.
AR: We also believe that they may be mainly due to the one-year-leave-out approach, especially when differences among the validation (target) year and the calibration period exist (e.g. for the wettest or driest years of the data period, which may not be of equal intensity). Depending on the "distance" between the target year and the calibration period this may cause a divergence from zero.

RC: page 13, line 13: "in the easternmost part" instead of "at the most eastern part"
page 14, line 30: add "cumulative" before "probability"
AR: These will be corrected in the revised version.

RC: page 17, figure 8: "Fraction of catchments" instead of "Number of catchments"
AR: This will be changed in Figure 8 and in Figure 9.

RC: page 18, last line: As far as I can see the CRPS in not lower after bias correction.
AR: We will review the sentence and separate IQR and CRPS analysis in the revised version.

RC: page 19, line 3: replace "in regards to" by "with respect to"
AR: This will be changed in the revised version.

RC: I recommend to combine figure 2 with figure 10 into one figure, and figures 3 with figure 11 into one figure, etc. The reader now has to turn over pages to compare the figures.
AR: We will consider if we can have the figures closer together in the revised version.

RC: Figure 15: how are seasons represented?
AR: Strong blue is used for winter, lighter blue for autumn, red for summer and lighter red for spring. We realized that the legend for the four seasons was missing in the figure and we will add it in the revised version.
