Response to Reviewer#2

The authors thank Reviewer#2 for his valuable review, which help us to enhance our paper. We provide below our answers to the reviewer's comments.

Reviewer 2

Summary

The study analyses the skill of ECMWF's System 4 seasonal forecasting system for precipitation and streamflow forecasts in 16 French catchments, using the hydrological model GR6J for transferring the meteorological forecasts into streamflow forecasts. In particular, the study focusses on the effect of bias-correcting precipitation in different ways. Main conclusions are that linear scaling and EDMD bias-correction with monthly calibration windows perform better than other methods. In general, bias-correction was found to improve the skills, but the result varies for the different skill scores. Often, a trade-off between decreasing sharpeness and increasing reliability was found when applying bias-correction. When comparing the differences of the bias-correction effect on precipitation and streamflow, it was found that in cases when sharpeness and overall performance increased by bias-correction, it increased more strongly for streamflow than for precipitation. The opposite was found to be true for reliability.

General comments

The study addresses a relevant scientific question and the methods used are sound. A few results need some further clarification and/or discussion and I have listed the open issues in the detailed comments. Additionally to the detailed comments, I would like to add two general comments.

Reviewer's comment (RC):

- The study uses mainly modelled streamflow as a reference. Nevertheless, I miss some indication of the hydrological model performance in the 16 basins. This is particularly relevant since also the observed streamflow is used as a reference forecast in one part of the manuscript, and this analysis would critically depend on systematic biases of the hydrological model.

Authors' reply (AR): Reviewer 1 also recommended indicating the performance of the hydrological model. Please refer also to our answer to his review. We proposed to add the following sentence on line 25: "…applied to root-squared flows. We obtained an average KGE of 0.95 in calibration and 0.94 in validation over the sixteen catchments. The bias obtained in simulation ranges from -0.02 to 0.05." Both KGE values and bias values show good performance of the model. We provide the table of all values hereafter:

Catchment	Calibration KGERQ	Validation KGERQ	Validation C2MQ	Validation 1-Bias
1	0.93	0.92	0.75	0.01
2	0.93	0.92	0.65	0.03
3	0.94	0.94	0.64	0.05
4	0.94	0.94	0.72	0.02
5	0.94	0.94	0.69	0.00
6	0.95	0.95	0.77	-0.02
7	0.95	0.95	0.79	0.03
8	0.97	0.97	0.87	0.02
9	0.97	0.97	0.84	-0.01
10	0.89	0.88	0.58	0.00
11	0.95	0.95	0.81	0.04
12	0.95	0.95	0.82	0.04
13	0.93	0.93	0.86	0.05
14	0.96	0.96	0.88	0.03
15	0.97	0.97	0.84	0.02
16	0.95	0.94	0.81	0.04

RC: • The manuscript covers a large body of results and is therefore lengthy. I think that it could be streamlined without losing too much information.
AR: The third reviewer and the editor also recommended shortening the article. This will be our priority when producing the revised version.

RC: Over all, I suggest acceptance of the manuscript after my comments have been taken into account. I'm looking forward to the revised manuscript.
AR: We thank the reviewer for this positive appreciation of our paper.

RC: Page 3, line 1: Some reference needed to support the statement that linear scaling and distribution mapping are widely used methods in seasonal forecasting.
AR: We will explicitly add specific references showing the application of these bias corrections methods in seasonal forecasting.

RC: Page 4, line 13: Which parametrization was used to derive potential evapotranspiration?
AR: The calculation of the evapotranspiration was done prior to this study and is embedded in the database we use in the Catchment hydrology team at IRSTEA. It follows the Oudin formula, which can be found in Equation (2) of the reference Oudin et al. (2005). K_1 is set to 100 and K_2 to 5, as shown in Equation (3) of this same paper. This reference is cited in our paper.

RC: Page 4, line 23: What is meant by interannual potential evapotranspiration? I would have understood the manuscript in such a way that potential evapotranspiration is derived from raw, i.e. non-bias-corrected, forecasts, but in this case, the term interannual potential evapotranspiration does not make sense. I probably misunderstood something and would like that the authors clarify the manuscript in that respective.
AR: Reviewer 1 also requested that the PET used in the paper should be better explained. For a given day of the year, the estimated PET on this day is assumed to be the mean of all PET computed for this day of the year, in all available years. Here, the mean interannual PET is the average of the PET calculated from observed temperatures for each year from 1958 to 2010. This will be clarified in the revised version (please refer also to our answer to his review).

RC: Page 5, lines 3-4: Just a comment, nothing to change: leave-one-year-out might result in the validation years not being really independent, as interannual serial correlation might be quite high. Maybe it would be interesting to test larger block sizes in future studies.
AR: Definitely. This point was also recently raised in a HEPEX bog post by colleagues from CSIRO (http://hepex.irstea.fr/how-good-is-my-forecasting-method-some-thoughts-on-forecast-evaluation-using-cross-validation-based-on-australian-experiences/). We think that a more-than-one-year-leave-out procedure could potentially fit better for one of our catchments, which has a high base-flow index. We believe that its impact on the other catchments would be lower, given the length of our calibration periods. Also, the impact is expected to be lower when calibrating the hydrological model than when implementing the bias corrections. In any case, it would certainly be interesting to test it in a future study, where more catchments could also be included and focus could be put on this aspect.

RC: Page 6, lines 21-25: In the case of EDM-m and GDM-m, only 29 data points are used to derive a cumulative distribution function for the reference data. This is a rather low number of data points, potentially leading to estimated cumulative distributions that are non-robust. Maybe, and this is of course rather speculative without analyzing the data, this could be a reason for the worse bias validation of EDM-m and GDM-m in Fig. 6.
AR: This is an interesting point and could, as suggested, partly explain the poorer performance of EDM-m and GDM-m. Nevertheless, it is also worth noting that it is difficult to have much more years available for the calibration of these correction methods, since the meteorological reforecast archive

needs to be homogeneous (i.e., based on the same model) over the period. The fact that bias correction methods require long time series of forecasts is a well-known limitation in the field. The point raised by the reviewer is worth mentioning in the revised version and we will include a comment on it.

RC: Page 6, lines 21-25: I'm not aware of a study that applied gamma distribution fitting for monthly precipitation data. Could you please cite a study to support the method GDM-m? I'm a bit worried that the gamma distribution might not be a good choice for monthly mean precipitation values.

AR: The choice of a cumulative distribution function for precipitation (or streamflow) data is always a challenging one. The Gamma distribution is often assumed to be suitable and fitted to precipitation sums. Some examples of the gamma distribution fitted to monthly precipitations are:

Zekai S. and A. G. Eljadid (1999) Rainfall distribution function for Libya and rainfall prediction, Hydrological Sciences Journal, 44:5, 665-680, DOI:10.1080/02626669909492266,

Husak, G. J., Michaelsen, J. and Funk, C. (2007), Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. Int. J. Climatol., 27: 935–944. doi:10.1002/joc.1441

The gamma distribution is also often used when computing the SPI. Examples are:

Lavaysse, C., Vogt, J., and Pappenberger, F.: Early warning of drought in Europe using the monthly ensemble system from ECMWF, Hydrol. Earth Syst. Sci., 19, 3273-3286, doi:10.5194/hess-19-3273-2015, 2015.

X. Lana, A. Burgueño, M. D. Martínez and C. Serra: A review of statistical analyses on monthly and daily rainfall in Catalonia. 2009. Tethys (Journal of Weather & Climate of the Western Mediterranean), 6, 15–29, 2009, doi:10.3369/tethys.2009.6.02.

In the preliminary steps of our study, we visually compared several distributions to fit to monthly precipitations in the selected catchments. The gamma distribution showed the best fit to the empirical distributions.

RC: Page 6, lines 25: It is unclearly written how exactly the EDM and GDM correction is applied to daily values. I assume it is done as such that the monthly values are corrected following the quantile mapping procedure. After that, a correction factor is estimated between the corrected and the uncorrected monthly mean value and this correction factor is applied to all daily values. The text on line 25 is though misleading as the actual correction in a quantile-mapping framework is the mapping of the uncorrected values to the cumulative probability space, from which a corrected value is derived following an inverse mapping based on the reference data. As the mapping is calibrated for monthly values, it cannot be used for daily values directly. Please clarify the text.

AR: The reviewer has understood correctly how the EDM and GDM methods are applied. To clarify the text, we propose to add on line 24, page 6: "In the case of EDM and GDM, the monthly values are first corrected based on the distribution mapping procedure. Then, for a given month, the ratio of the corrected monthly value and the non-corrected ones is used to correct all daily values within this month."

RC: Page 8, lines 23-24: The first sentence in this paragraph is redundant. Consider removing it.

AR: Following the reviewer's suggestion, we propose to merge the sentences on lines 23-24, page 8 into a single sentence: "To investigate the gain in performance brought by bias correction methods, we use the raw (uncorrected) forecasts as reference in the computation of the skill scores."

RC: Page 8, lines 27-28: According to section 3.3, all data was first converted to weekly means, thus a seven-day moving average cannot be derived. Please clarify the contradictions.

AR: When computing skill scores with reference to the ESP or historical streamflow, we computed scores based on weekly-averaged precipitation or streamflow. But when we computed the skill scores with reference to the raw System 4 forecasts (to calculate the UFL), scores were computed for daily values. In this case, the moving average allowed us to remove the high frequency variations in the skill

scores while looking at the impact of bias corrections on daily forecast values. We will clarify this in the revised version.

RC: Page 9, line 12: Why is the value +0.1 and -0.1 for the deviations from the diagonal chosen?

AR: Laio and Tamea (2007) propose to calculate the position of these "tolerance" lines to correspond to a significance test: *« The Kolmogorov bands are two straight lines, parallel to the bisector and at a distance q(α)/sqrt(n) from it, where q(α) is a coefficient, dependent upon the significance level of the test α (e.g., q(α = 0.05) = 1.358, see D'Agostino and Stephens, 1986). The test is passed when the curves remain inside these confidence bands. »*. In our case, the Kolmogorov significance bands should be approximately at 0.15 to correspond to a 5 % significance test. The 0.1 bands we use are thus a good conservative choice to test deviations from the diagonal.

Laio F. and Tamea S.: Verificatin tools for probabilistic forecasts of continuous hydrological variables. HESS, 11, 1267-1277, doi: 10.5194/hess-11-1267-2007,2007.

RC: Page 9, line 19: Unclear use of the word "translate".

AR: We propose to replace "translate" by "indicate" in the revised version.

RC: Sections 4-6: The presentation of the results could be improved and shortened. When I read the manuscript, I would have liked to have the comparison of the raw and bias corrected forecasts closer together and I suggest combining the discussion of the raw and the EMDD corrected forecasts. It would be much easier for the reader to follow the discussion if, for e.g., figure 2 and 10 are to be combined into one figure. Similarly for all other seasonal skill score figures in sections 4 and 6.

AR: Reviewer 3 also proposed to have the figures of Sections 4 and 6 closer together for an easier visual comparison of the performances before and after bias correction. We will consider these suggestions in the revised version of the paper.

RC: Page 11, line 8-9: I thought that the reference forecast is the streamflow simulated using the reference precipitation. Thus, any model deficiencies regarding low flows should not affect the skill score as also the reference forecast would suffer from those deficiencies. Also, similarly as for the low flows, the PIT diagram reports difficulties to forecast the high flow. What could be the reason for this issue? The explanations give in the manuscript so far are not fully convincing.

AR: The reference forecast in the computation of the skill scores uses the hydrological model (in all figures but Figure 13), and model deficiencies cannot be detected based on the graphs of skill scores, as well noted by the reviewer. Here, however, the explanations proposed on lines 8-9 refer to the PIT diagrams of Figure 5 (which are not expressed as skill scores) and, more specifically, to the lack of reliability observed in the summer season (JJA). The tendency to have observations below the forecast range is obtained with both the streamflow simulated with System 4 precipitations (in red) and the streamflow simulated with the reference precipitation (in grey). This is why we make the assumption that this lack of reliability is due to the hydrological model rather than the precipitation forcings. We also should note that it is hard to distinguish between high and low flows based on the PIT diagram solely. In summer, we have observed an under-dispersion of forecasts, but also a strong tendency to have observations falling below the forecast range. From the hydrographs, we also observed that a large part of the observations falling in the lowest forecast range in summer can be associated with low flows. The PIT diagram thus needs to be analysed together with the hydrographs to better separate the effects of under-dispersion on high or low flows. We will clarify this in the revised version of the paper.

RC: Page 14, lines 19-23: The reasoning is unclear to me, probably due to an unclear explanation how the bias-correction works. If it is done in the way I described in the comment regarding EDM and GDM, I don't think that the reasoning is

correct. Everything stated for the monthly correction would also apply for the daily correction. Also on the daily time scale, the rank structure (see comment below) of the forecast is not the same as for the reference data. In both cases (monthly and daily correction), the distribution mapping should be able to correct differing rank structures and remove biases in the monthly mean effectively. In fact, I would have expected the daily correction to perform worse than monthly correction when evaluated on the monthly scale since it is not targeted to the monthly scale but the daily scale. I rather think it has to do with a higher sensitivity of monthly corrections to overfitting as evaluated within the cross-validation framework. Admittedly, distribution mapping can lead to unforeseen effects and it might very well be that I'm wrong. If the authors are convinced that their reasoning is correct, I would like them to describe in the reply a case where the distribution mapping fails in more detail, for e.g. by showing how the reference and forecast distribution look like and how the mapping fails to come up with a correct monthly mean value.

Page 14, lines 19 and 21: Usage of the term "time structure" seems to be misleading. I understand this term in a way that it refers to the temporal sequence of values, i.e. that the day n in the reference corresponds to day n in the forecasts. However, distribution mapping does not have this requirement. It is rather the rank structure as I would call it: Rank n in the reference has to correspond to the rank n in the forecasts. Please correct the terminology or explain in more detail what "time structure" means.

AR: We believe that compensation effects (linked to data aggregation) may occur when evaluating monthly values with bias corrected daily values. Since daily corrections are more numerous than monthly corrections, this can result in more flexibility and daily correction performing better than monthly correction when evaluated at the monthly scale. This is more or less similar to monthly correction performing well when evaluated at the yearly scale. However, as mentioned by the reviewer, this may also be linked to a "higher sensitivity of monthly corrections to overfitting". Further studies would be necessary to conclude more firmly on this issue. We will revise the text in order to be more careful on the comment we made.

Concerning "time structures", we agree that the terminology may not be very clear. We meant that, for the DM methods to be efficient, the uncorrected and corrected values with the same rank (in their respective cumulative distributions) should also occur at the same time (e.g. as observed in the hyetograph). Therefore, the "time evolution" of values should be consistent and DM methods will be more efficient if they are applied on forecast hyetograph that are not too discrepant. We believe that this can go unnoticed in performance evaluation if daily values are corrected and aggregated at the monthly scale for evaluation. However, it will be harder to cover up if we apply and evaluate the corrections at the same time scale. We will be more precise about this in the revised version in order to clarify our idea.

RC: Section 5.2: In my opinion, this section does not give new information which is not already present in figure 6 (time varying bias-correction factors can be inferred from the panel "Before bias correction") and I suggest removing it for the sake of shortening the result section. The only new aspect is that the correction factors for EDMD vary more than for LS, but this comparison is not valid in my point of view as one should not compare a mean correction factor with a correction factor for a quantile level. I'm pretty sure that if you would calculate the correction factor for the mean in the case of EDMD, it would be very similar to the LS factor.

AR: The idea behind this plot was not to compare LS and EDMD average correction factors (we fully agree with the reviewer that this comparison is not valid). In fact, we wanted to give an additional element to understand the different features behind the LS and the EDMD methods. The main conclusions from this figure are that (1) EDMD can correct the frequency of null precipitations, whereas LS cannot, and (2) correction coefficients do not vary much from one application year to the other (especially with LS) and, therefore, in operational contexts, one can choose a more parsimonious calibration of the bias correction method applied. We will consider the ways to shorten this section in the shortened revised version of the paper.

RC: Section 5.3 and figures 8 and 9: I very much like this analysis. I'm not sure though if I really understand the analysis completely. MAE is partly related to the

bias analysis in figure 6, i.e. if biases in figure 6 are substantial, then MAE should be even larger since MAE does not allow for a compensation of errors. EDM-y and GDM-y have large biases throughout the year in figure 6, and in some cases and particularly in summer, the bias is even larger than in the uncorrected data. However, in figure 8 the two methods stick out for MAE and IQR in summer lead to skill improvements in all catchments up to a lead time >60 days. To me, this seems to be contradicting. Could you please explain this particularity?

AR: Thank you. Fig. 8 reflects, somehow, the ability to bring skill to the (corrected) forecasts in terms of lead time and expressed as a percentage of catchments where improvements (comparatively to the raw forecasts used as reference) were seen. Even if EDM-y and GDM-y methods result in forecasts that still present some strong biases (as seen in, and commented from, Fig. 6), these may result in MAE values smaller than MAE values computed from the raw forecast. This is enough to characterize a relative gain in skill and, if this is observed over all lead times, the UFL will be >60 days and count in the percentage represented in Fig. 8. It seems contradictory at first sight, as well observed by the reviewer, but can, computationally, happen (e.g. due to the different aggregations: MAE is computed with daily values over a season, the bias is computed with monthly values over a month, or when biases change from under to over prediction or vice versa after correction). Overall, it is interesting to note that forecast skill is definitely hard to evaluate as there are many facets that one can look at. We tried to explore this in this paper and shed lights into the different aspects that can better inform forecast users.

RC: Page 19, line 1: If I read the figure 10 correctly, there are negative skill score values and therefore, the statement that the skill scores are always larger than zero does not hold.

AR: You are right. We propose to replace the sentence with: “Nevertheless, bias corrected forecasts remain sharper than the reference (i.e., skill scores are mostly greater than zero).”

RC: Page 20, line 3: If I read the figure 12 correctly, there are negative skill score values and therefore, the forecast performs sometimes worse than ESP, which is the opposite of what is stated on this line.

AR: By “overall” we meant to indicate that. We propose to clarify it by changing to: “Overall, after bias correction, streamflow forecasts are sharper than ESP in all catchments and seasons (only the winter and summer seasons are shown but results are similar for the spring and autumn seasons)”.

RC: Page 20, lines 12-13: It is not clear to me why this is expected. I would expect that comparison to streamflow climatology is a harder check and therefore the skillfull lead time should be smaller than in the comparison to the baseline reference run since also the hydrological model bias deteriorates the skill. I surely misunderstand something but I think it would be good to add a bit more explanation in the manuscript.

AR: It is usually expected that ensembles based on streamflow climatology have less skill than ensembles based on hydrological modelling, at least in the first lead times, because ensembles based on hydrological modelling benefit from knowledge of initial hydrologic conditions. For instance, here, the states of the GR6J model are first initialized by running the model with observed inputs for a year prior to the forecast date. Therefore, ensembles based on streamflow climatology are supposed to be less skilful for forecast lead times that are impacted by initial hydrologic conditions. In France, studies have shown that these lead times extend to a month, on average. We will make sure it is clearer in the revised version.

RC: Page 22, line 4: As for the uncorrected forecast discussion, I do not understand why it is the hydrological model that causes the problems with low-flow overestimation. The reference data is also output of the same hydrological model driven by the reference precipitation data. I would therefore rather think that it is some characteristics in the input data which the bias-correction cannot correct for that causes the problem (for e.g. dry-spell lengths). If the authors still think their statement holds, I would like to have a bit more explanations why this can be the case.

AR: The main discussion here is about reliability, which can clearly still be improved for streamflows. Comments on the model performance are linked to the analysis of the simulated and observed hydrographs, which complement the PIT analysis. The lack of reliability in streamflow forecasts may come from the input data, but not solely (as shown in Fig. 3, which analyses the reliability of the precipitation forcing; please also refer to the answer given above referring to Fig. 5, **RC: Page 11, line 8-9**). A lack of spread in hydrologic initial conditions may also play a role in the reliability of streamflow forecasts. That is why we referred to the needs of accounting for other sources of uncertainty, with, for instance, additional post-processing (lines 4-6). This may not be clearly stated and we will clarify it in the revised version.

RC: Page 27, lines 13-15: References needed

AR: We propose to add the following references:

Hamlet, A. F. and Lettenmaier, D. P.: Columbia River Streamflow Forecasting Based on ENSO and PDO Climate Signals, J. Water Resour. Plan. Manag., 125(6), 333–341, doi:10.1061/(ASCE)0733-9496(1999)125:6(333), 1999.

van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J. and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resour. Res., 49(5), 2729–2746, doi:10.1002/wrcr.20251, 2013.

Werner, K., Brandon, D., Clark, M. and Gangopadhyay, S.: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts., J. Hydrometeorol., 5(6), 1076–1090, 2004.

In addition, the last paragraph will include the following reference:

Ionita, M., Boroneant, C., Chelcea, S., 2015. Seasonal modes of dryness and wetness variability over Europe and their connections with large scale atmospheric circulation and global sea surface temperature. Climate Dynamics 45, 2803–2829. doi:10.1007/s00382-015-2508-2

RC: Section 6.3: Just a comment: I very much like this analysis.

AR: Thank you!

RC: Section 6.4: Although I like the illustrative character of this section, it stands a bit loose within the rest of the manuscript. I suggest to either motivate the section better or, for the sake of brevity, to remove it. In my opinion, the main statements of this sections have already been made, i.e. increased sharpeness after bias-correction compared to ESP.

AR: Reviewer 1 also appreciated this figure, so we prefer to keep it. He proposed some improvements that we think will better motivate the section: adding a quantification of what is shown in this figure. Notably, we show that the coverage probability of the streamflow forecasts is improved after bias correction compared to ESP (see our answers to Reviewer 1 for details). Additionally, studies have shown the need to combine statistical evaluations with visual evaluations. Even though this is hard to achieve in probabilistic forecasting, we wanted to propose a visual appreciation of the ensembles to have a better overview of how bias corrections affect streamflow forecasts.

RC: Figure 2: "… and all seasons." The figure only shows two seasons, please correct the caption.

AR: Correct: "… all seasons" will be replaced by "and the winter (DJF) and summer (JJA) seasons".

RC: Figures 3, 5, 11, 14: The dashed lines should be explained in the figure as well, and not just in the text describing figure 3.

AR: The explanation will be added in the captions.

RC: Figure, 6: Although certainly correct, I do not see a reason why to transform the simple relative bias into 1-bias. I understand that this transformation turns the bias into a skill score. However, in my opinion, the interpretation is not following the one for skill scores anyway. The perfect bias-correction would not

yield 1 but 0. I suggest plotting the relative bias without transformation. The scale would be much easier interpretable as it directly refers to a percentage over- or underestimation.

AR: We used this transformation so that "no bias" corresponded to the null value, over-prediction corresponded to positive values and under-prediction corresponded to negative values. This representation of the scale seemed more intuitive, but the reviewer is right that the interpretation in terms of percentage is easier without this transformation. We will test and consider the reviewer suggestion when preparing the revised version.

RC: Figures 8 and 9: Why are there different color scales for the different seasons?

AR: The four colours are supposed to help the reader identify the four seasons throughout the article. These colours include the blue and red colours used throughout the paper: blue for winter, lighter blue for autumn, red for summer and lighter red for spring. In these figures, the four colour scales in the legend are needed to clarify the colour shades related to the percentage of catchments in each category (e.g. to avoid light blue (autumn) being mistaken for a shade of bright blue (winter)).

RC: Figure 15: What are the colours standing for? There is probably also an error in the caption where it reads "shown for all seasons".

AR: The colours represent the four seasons as mentioned in the reply above. We will add an explanation in the figure.

RC: Page 19, line 10: precipitation instead of precipitations

AR: This will be modified in the revised version.