Main changes in the revised version and detailed answers to the reviewers

We thank the reviewers for their comments and suggestions, which helped us to improve our paper. The main changes in the revised version concern the following issues:

- We added some sentences in order to: i) better explain how PET was considered, ii) present the KGE results of the calibration and validation of the hydrological model, as requested by two reviewers, iii) add pragmatic results on the interpretation of the hydrographs, as requested by one reviewer, which clarifies the role of the bias correction methods in improving forecast quality.
- We shortened considerably the paper, as suggested by two reviewers and by the Editor, by cutting several words and shortening several sentences, which we thought could be deleted or rephrased without affecting the message the paper conveys. We also deleted an entire section on the corrective factors, which we believe was not essential for the comprehension of the main findings of the paper.

All other minor changes and answers to the reviewers are detailed below.

Reviewer 1

Reviewer's comment (RC): While the Introduction is well balanced and gives useful insight on previous work on the topic and also references supporting the envisaged methodology, I found that the final paragraphs should possibly include more information on the novelty of the present manuscript. Also in the methodological section some more referencing is needed. See minor comments for this.

Authors' reply (AR): We changed the following in the Introduction to better emphasize the novelty of our study (end of line 29, beginning of line 30, page 3): "Despite these recent works, and to the knowledge of the authors, no previous study has compared bias correction methods and their impact on streamflow forecasting in a systematic way, with a focus on understanding how the main attributes of forecast performance are impacted by bias correction.

This paper aims to provide insights into the way bias correcting seasonal precipitation forecasts can contribute to the skill of seasonal streamflow predictions, notably in terms of overall performance, reliability, sharpness and skilful lead time. It investigates the potential of bias corrected ECMWF System 4 forecasts to improve streamflow forecasts at extended lead times over 16 catchments in France. An in-depth comparison of eight variants of linear scaling and distribution mapping methods applied over the 1981-2010 period is presented. Section 2 presents…"

RC: 4-4-15: We learn here about the meteorological forcing. It is clear to me how you use precipitation, but as a forecast and as SAFRAN product. Concerning Potential Evapotranspiration (PET), only SAFRAN is declared. I'd like you to declare which PET is used in retrospective forecasts forced by the ECMWF products. If it is from ECMWF, you should state why you are not post-processing it. If you use SAFRAN, you should be able to assess how much uncertainty are you neglecting by using the best observed estimates of PET instead of using a forecasted value (which you need to do as soon as you will deploy the system in real-time). In our experience, for basins not affected by snow-melt, the post-processing of relative-humidity data (an important proxy the evaporation demand by the atmosphere) helps improving the estimation of hydrological droughts (Jörg-Hess et al., 2015).

AR: The potential evapotranspiration (PET) used to force the hydrological model is, in fact, the mean interannual PET. 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 for each year from 1958 to 2010. PET for each year is calculated using SAFRAN. Regardless of the precipitation scenario fed to the model (historical precipitations or System 4), the PET scenario used as input to the model is always the same: the series of mean interannual PET

corresponding to the forecast period. With this setup, we can focus on the changes in skill that can solely be attributed to the bias correction of precipitations, which is in the aim of our study. Adding the uncertainty of temperature forecasts in the analysis would in fact require a different framework. For instance, we would need to set up multi-variable bias corrections to take into account the dependencies between precipitation and temperature, or we would need to consider the impact of observed trends in time series of observed temperatures in some regions in France prior to post-processing and ESP forecasting. This is beyond the scope of this study, although interesting for further investigations and specific operational setups.
In the revised version, we clarified the way PET is considered by adding this sentence in Section 2.1: "The interannual potential evapotranspiration was then computed in each catchment, i.e. for a given day of the year, we computed the average potential evapotranspiration for this day over all available years (1958 to 2010)", and the following in Section 2.2: "Here, the series of interannual potential evapotranspiration corresponding to the forecast period was systematically used as input to the hydrological model. With this setup, we aimed to isolate the influence of precipitation forecast inputs on the quality of streamflow forecasts."

RC: 4-25: I just reviewed another paper on seasonal forecasting where authors did not show any score concerning their calibration/validation and I amended it. Same here. I am happy with a table as supplementary material.
AR: The table below summarizes the scores obtained in calibration and validation of the GR6J model. In the revised version, we added the following sentence in Section 2.2: "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 0.95 to 1.02."

Catchment	Calibration KGERQ	Validation KGERQ	Validation C2MQ	Validation Bias
1	0.93	0.92	0.75	0.99
2	0.93	0.92	0.65	0.97
3	0.94	0.94	0.64	0.95
4	0.94	0.94	0.72	0.98
5	0.94	0.94	0.69	1.00
6	0.95	0.95	0.77	1.02
7	0.95	0.95	0.79	0.97
8	0.97	0.97	0.87	0.98
9	0.97	0.97	0.84	1.01
10	0.89	0.88	0.58	1.00
11	0.95	0.95	0.81	0.96
12	0.95	0.95	0.82	0.96
13	0.93	0.93	0.86	0.95
14	0.96	0.96	0.88	0.97
15	0.97	0.97	0.84	0.98
16	0.95	0.94	0.81	0.96

RC: 24-18: I like this evaluation very much, just, I miss some quantification supporting the description based on visual inspection you are giving. Be pragmatic.
AR: Thank you very much for this suggestion. The MAE and coverage probability provide a good quantification to support the description. The values of MAE, coverage probability 90% (COV 5-95) and 50% (COV 25-75) obtained by each forecasting system over the displayed period (April 2004 to April 2007) are shown below. They show that the ensembles based on past precipitations and past streamflow (HistQ and ESP) are more accurate over the chosen period (lower MAE values), but that EDMD-m performs better in terms of coverage probability. We included this quantitative analysis in the interpretation of the hydrographs, as suggested by the reviewer.

	HistQ	ESP	LS-m	EDMD-m
MAE (m^3/s)	3.81	4.06	4.26	4.26
COV 90 % (5-95)	97 %	92 %	85 %	89 %
COV 50 % (25-75)	66 %	60 %	46 %	51 %

RC: 25-3: The discussion section is here quickly merged with the conclusions. The only link to current literature one is expecting here merely consists in a enumeration of possible post-processing of the forecasts with currently available methods. Here some more effort has to be shown to make also this section a valuable part of the manuscript.

AR: The reviewer is right that this section reflects more our conclusions. A posteriori, we think that sections 6.3 and 6.4 reflect the discussions, putting the results into a broader perspective. We thus renamed the last section "Conclusions".

RC: 26-2: You address here the issue of implementation in operational systems. Again, declare how you deal the PET, and then re-evaluate the potential for real-time operations.

AR: We added some sentences to better explain how PET was considered (see reply above) and we deleted the sentence on operational issues to avoid introducing a discussion that is not the focus of this paper.

RC: 2-11: I guess here you should give one or two references for the statistical models, too. Eg. Some approaches relating winter snowpack to summer-flows (e.g.: Godsey et al., 2014; Jenicek et al., 2016).

AR: Thank you for pointing out to these interesting references. We added Jenicek et al. (2016) (mentioning references therein) as suggested.

RC: 5-4: Please support the "one-year-leave-out cross-validation method" with a reference.

AR: We added the following reference and changed the "one-year-leave-out" denomination to "leave-one-year-out" for consistency:

Arlot, S. and C. Alain. A survey of cross-validation procedures for model selection. Statist. Surv. 4 (2010), 40-79. doi:10.1214/09-SS054. http://projecteuclid.org/euclid.ssu/1268143839.

RC: 6 -2: Please support "Precipitation and streamflow forecasts are evaluated with deterministic and probabilistic scores commonly used in ensemble forecasting" with a reference, e.g. Brown et al EVS paper.

AR: In the process of reducing the length of the paper, this sentence was removed. Yet, references supporting the evaluation criteria can be found in Section 3.3.

RC: 8-15: Nice idea to use the ensemble of past-streamflow observations as a reference. If you would "sort-out" some past years by means on analogues techniques you might get a very challenging set of members for your ensemble forecast. Have you tried this?

AR: This is precisely the topic of another paper that we have just submitted to this special issue. It is available here: http://www.hydrol-earth-syst-sci-discuss.net/hess-2016-285/

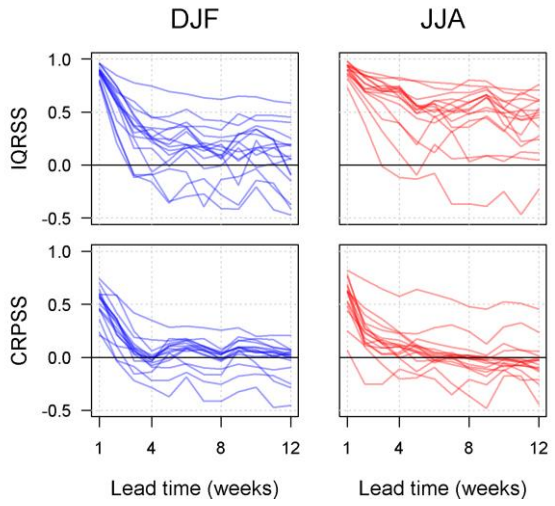
RC: 8-22: Another interesting feature here. This definition of gain is very elucidative. Can you maybe elaborate on pro and contra of this kind of "gain" definition with respect to scores based on cost-loss considerations?. Why choosing such a large gap of day between the classes? Have you tried to make a 30-day moving window? Or a 15-days moving window?

AR: Thank you for the comment. We chose to evaluate the gain in terms of anticipation in response time, rather than in terms of relative economic value (REV), for instance, since cost-loss considerations would need an evaluation of (or additional assumptions on) mitigation costs, avoidable losses, as well as unavoidable losses for each studied catchment. Here, we may assume that increasing the anticipation response time could increase time for preparedness, which would decrease costs and losses related to missed events or actions taken with no or little anticipation to a critical situation. The cost-loss approach would need to be applied considering this evolution of forecasting with time since in a seasonal forecasting system one has several forecasts or months ahead to detect a potential critical situation and

act accordingly. Actions and consequences would need to be stratified according to the time available for action in order to have this aspect reflected in an evaluation score.
The gap was chosen to help represent the improvements due to bias correction at a monthly time scale of reference. It seemed to us that a month ahead could be a good minimum of time necessary to adapt any mitigation actions once a critical situation is forecasted by a seasonal forecasting system. As shown in Figures 8 and 9, this choice seems to be appropriate to a joint representation in a plot, while differentiating situations for a useful analysis.
We did not try to use windows larger than 7 days. The objective of the rolling mean was to smooth the skill curves and remove the high frequency variations of the skill at the daily time step. Seven days appeared to be enough to smooth the curves, while keeping the moving mean as a good estimate of the gain in lead time.

RC: 9 – 2 & 9 -19: Both in Figure 2 & 4 CRPSS is showing increasing skill at weeks 5 and 9. We are also used to "struggle" in interpreting such cycles. Do you have some ideas on your particular case here?
AR: We have also spent some time trying to interpret these cycles. Despite a closer look at the data and the scores, under different angles, we could not see any systematic reasons for these cycles. We think it may be related to several correlated aspects, such as the type of forecasting model/system, the forcings, the behaviour of the catchment, etc.

RC: 11 – Figure 4: How would look like this figure if you use the "ensemble based on past streamflow" as a reference?
AR: The following figure represents the IQRSS and the CRPSS of the streamflow forecasts without bias correction, when the ensemble based on past streamflow is used as reference. It can be compared to Figure 4 to see that the skill is higher with this reference, and to Figure 13 to see that bias correction has also increased the skill of forecasts with regard to past streamflow.



RC: 13 – Figure 6: Right margin is cropped. Additionally, the "too wet"=red is not really intuitive.
AR: We exchanged the blue and red colours in the scale so that blue corresponds to overestimation and red corresponds to underestimation. We also increased the right margin.

RC: 13 – 2: "the 2-month" or the "month-2"? If you mean the one for the second month of the forecast I would find more adequate to use "month-2".
AR: We agree with the reviewer and changed the occurrences of "the 2-month" to "month-2" in the revised version.

RC: 17 – Figure 8 (and later 9): I like such Figures because the train my brain. Tell me if I am reading it wrong: If a look at a certain score in a certain season than for a particular bias correction method a percentage of the basins is showing improvement in lead time. Of this percentage a distinct portion shows improvement of let's say 60 to 90 days. So largest improvement is in the PIT-Skill in summer and Winter for the EDMD methods.
Right?
AR: Thank you. Your reading of the figure is absolutely correct.

RC: 22 – 8: This would be the only heading with a question mark. Maybe replace this with a sentence
AR: In fact, the question mark was a typo and we removed it.

RC: 23 – Figure 15: is there any special reason (beside readability) for having different scales in the three panels?
AR: No, there is no special reason, apart to zoom in on the case of the CRPSS. Following this comment, we changed the figures and used the same scales in the axes.

Reviewer 2

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 (see our reply above). We added the following sentence in Section 2.2: "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 0.95 to 1.02."

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: We cut several words and sentences along the text and removed an entire sub-section concerning the corrective factors (Section "Comparison of bias correction factors for LS and EDMD methods"), which reduced the length of the paper.

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 added a reference to the review paper of Yuan et al. (2015).

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 used. 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 the 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: 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.

In the revised version, we clarified the way PET is considered by adding one sentence in Section 2.1 and one sentence in Section 2.2 (see also our reply to Reviewer 1 above).

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 also 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.

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. In order to clarify it in the revised version, we added the following sentence: "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: This was done and only one sentence appears in the revised version: "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.

In the revised version, we clarified this point. We changed the first sentence of Section 3.3 which was too general to the following: "The quality of the forecasts was evaluated as a function of lead time and for the winter (December-January-February), the spring (March-April-May), the summer (June-July-August) and the autumn (September-October-November) seasons".

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.

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

AR: We replaced "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: The manuscript was shortened by removing sentences, repetitions and by removing a whole sub-section. Concerning combining figures with and without bias correction from EDMD-m, this would in fact disturb the logic of the paper since, in between these figures, we propose an in-depth analysis of the impact of the bias corrections per month and for each tested method. We then preferred to keep the organization as it was first proposed.

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.

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.
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.
In the process of reducing the length of the paper, this paragraph was removed in the revised version.

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, but 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.
In the revised version, we followed the reviewer's suggestion and removed this section.

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 light on 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 replaced 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: We changed the sentence to: "Overall, after bias correction, streamflow forecasts are sharper than ESP in most catchments and for most lead times".

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 the revised version, in order to clarify this issue, the sentence was changed to: "Streamflow forecasts generated from…up to twelve weeks in some catchments. This was expected because ensembles based on hydrological modelling benefit from knowledge of initial hydrologic conditions."

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). 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.

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

AR: We added 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.

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: Thank you. Reviewer 1 also appreciated this figure and suggested some improvements, which we implemented in the revised version. We therefore added 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: Thank you for pointing this out. The caption was corrected in all occurrences of this problem.

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 for the dotted lines was added in the captions of the four figures.

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. Figure 6 was modified in the revised version to take into account this reviewer's comment.

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 added a legend for the colours in the figure.

RC: Page 19, line 10: precipitation instead of precipitations
AR: This was corrected in the revised version.

Reviewer 3

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): We have cut several sentences along the text and removed an entire analyses concerning the corrective factors (Section Comparison of bias correction factors for LS and EDMD methods), which reduced the length of the paper.

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 based on the PIT diagram and the IQR. In order to clarify this issue, we added the following sentence to Section 3.3.1: "In this study, we considered that for two reliable systems, the sharpest one is the best (Gneiting et al., 2007)."

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. 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), 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. We would like to avoid adding a figure that is already presented in another easy-to-access paper that we are referencing (Laio and Tamea, 2007), especially since we are asked to shorten the paper. However, to make the PIT interpretation clearer, we modified some sentences in the description of this score in Section 3.3.1, and we added a more explicit title to the x axes of the PIT diagrams in our figures. The terms describing and explaining the shapes of the PIT diagram do not refer to "points" anymore (we have linked our points with lines for a better visualization of the results of the 16 catchments in a unique PIT diagram, and we think that deleting references to "points" will make the PIT diagram easier to understand).

[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: The scores are described in Section 3.3.1, and details on their characteristics can be found in the references provided. The way they are impacted by bias correction is illustrated throughout the results and discussed in Section 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. 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 studied have a dominant pluvial regime and are not strongly influenced by snow. In Section 2.2, 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 slightly changed the description of the model in the revised version (Section 2.2), which we hope will make this part of the paper clearer.

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. References on the bias correction methods are already provided in the Introduction. We also added a reference to Yuan et al. (2015), which gives an extensive review of methods and a list of additional references.

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: These points were corrected 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. We rephrased it in the revised version to make it clearer.

RC: page 3, line 18: add "heavily" before "influenced"

AR: This was 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). The following text was added in Section 2.1: "The interannual potential

evapotranspiration was then computed in each catchment, i.e. for a given day of the year, we computed the average potential evapotranspiration for this day over all available years (1958 to 2010)."

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. We added the following sentence in Section 2.2: "This setup is also consistent with the fact that our catchment set is dominated by a pluvial 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. 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 in the paper). 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 were 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 1958 to 2010. Observed streamflow data were available for different time periods, ranging from 36 years to 52 years depending on the catchment, and up to 2010. This was specified in Section 3.3.2 of the revised version.

RC: page 8, line 23: "caused" instead of "brought"
AR: This was 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. In the revised version, we added "first" before "lead time beyond which" 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: The reviewer is right. In the process of shortening the paper, this sentence was removed 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"
AR: This was corrected in the new version.

RC: page 14, line 30: add "cumulative" before "probability"
AR: This section was deleted in the revised version in order to shorten the length of the paper.

RC: page 17, figure 8: "Fraction of catchments" instead of "Number of catchments"
AR: This was 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: The reviewer is right. We changed the sentence to "In some catchments, the values of IQR are lower, but bias corrected forecasts remain sharper than the reference (i.e., skill scores are mostly greater than zero)" to clarify this point.

RC: page 19, line 3: replace "in regards to" by "with respect to"
AR: This was corrected 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: Combining figures with and without bias correction from EDMD-m, would in fact disturb the logic of the paper since, in between these figures, we propose an in-depth analysis of the impact of the bias corrections on forecast quality for each month and each tested method. We thus preferred to keep the organization as it was originally proposed.

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 added a legend for the four seasons represented in the figure.

Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts

Louise Crochemore[1], Maria-Helena Ramos[1], and Florian Pappenberger[2,3]

[1]Irstea, Hydrosystems and Bioprocesses Research Unit, 1 rue Pierre Gilles de Gennes, F- 92 761, Antony, France.
[2]ECMWF, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK.
[3]School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS, UK.

Correspondence to: Louise Crochemore (louise.crochemore@irstea.fr)

Abstract. Meteorological centres make sustained efforts to provide seasonal forecasts that are increasingly skilful, which has the potential to benefit streamflow forecasting. Seasonal streamflow forecasts can help to take anticipatory measures for a range of applications, such as water supply or hydropower reservoir operation and drought risk management. This study assesses the skill of seasonal precipitation and streamflow forecasts in France to provide insights into the way bias correcting precipitation forecasts can improve the skill of streamflow forecasts at extended lead times. We apply eight variants of bias correction approaches to the precipitation forecasts prior to generating the streamflow forecasts. The approaches are based on the linear scaling and the distribution mapping methods. A daily hydrological model is applied at the catchment scale to transform precipitation into streamflow. We then evaluate the skill of raw (without bias correction) and bias corrected precipitation and streamflow ensemble forecasts in sixteen catchments in France. The skill of the ensemble forecasts is assessed in reliability, sharpness, accuracy, and overall performance. A reference prediction system, based on historical observed precipitation and catchment initial conditions at the time of forecast (i.e., ESP method), is used as benchmark in the computation of the skill. The results show that, in most catchments, raw seasonal precipitation and streamflow forecasts are often more skilful than the conventional ESP method in terms of sharpness. However, they are not significantly better in terms of reliability. Forecast skill is generally improved when applying bias correction. Two bias correction methods show the best performance for the studied catchments, each method being more successful in improving specific attributes of the forecasts: the simple linear scaling of monthly values ~~contribute~~ contributes mainly to increasing forecast sharpness and accuracy, while the empirical distribution mapping of daily values is successful in improving forecast reliability.

1 Introduction

Numerous activities with economic, environmental and political stakes benefit from knowing and anticipating future streamflow conditions at different lead times. ~~While flood forecasting requires forecasts up to several hours or days ahead, other areas such as water supply reservoir operations or drought risk management need forecasts for the months or season ahead. Regardless of the considered lead time, streamflow forecasting~~ Streamflow forecasting systems are frequently ~~updated~~ developed to take the latest useful information content into account (e.g. last observed discharges, soil moisture or snow cover) and ~~developed~~ to make use of numerical weather model outputs to extend the range of skilful predictions.

Seasonal forecasts have shown to perfectly fall within a context of proactive risk management, for example, for drought management (e.g. Wilhite et al., 2000; Dutra et al., 2014; Mwangi et al., 2014; Wetterhall et al., 2015). Extended-range forecasting systems can be valuable ~~tools~~ to help decision-makers in planning long-term strategies for water storage (Crochemore et al., 2016) and to support adaptation to climate change (Winsemius et al., 2014). Nevertheless, several users still remain doubtful whether seasonal forecasts can be trustworthy or skilful enough to enhance decision-making ~~in an operational context~~ (Rayner et al., 2005). Lemos et al. (2002) list the performance of seasonal forecasts, the misuse of seasonal forecasts by end-users and the lack of consideration of end-users' needs in the development of products as major obstacles to the widespread use of seasonal forecasting in North-East Brazil. It is therefore crucial to assess the potential of available seasonal forecasting products and communicate on the assets and shortcomings of the different approaches ~~that can benefit~~ for the water sector (Hartmann et al., 2002).

Seasonal forecasting methods in hydrology can be broadly divided into two categories: statistical methods, which use a statistical relationship between a predictor and a predictand (e.g. Jenicek et al., 2016, and references therein), and dynamical methods, which use seasonal meteorological forecasts as input to a hydrological model. More recently, mixed approaches have been investigated ~~in the attempt~~ to take advantage of initial land surface conditions, seasonal predictions of atmospheric variables and the predictability information contained in large-scale climate features (see Robertson et al., 2013; Yuan et al., 2015, and references therein). Ensemble Streamflow Prediction (ESP; Day, 1985) is a dynamical method that is widely used to forecast low flows and reservoir inflows at long lead times (Faber and Stedinger, 2001; Nicolle et al., 2014; Demirel et al., 2015). It consists in using historical weather data as input to a hydrological model whose states were initialized for the time of the forecast. The ESP method is also used along with the Reverse-ESP method to determine the relative impacts of meteorological forcings and hydrological initial conditions on the skill of streamflow predictions (Wood and Lettenmaier, 2008; Shukla et al., 2013; Yossef et al., 2013). An alternative dynamical method consists in using seasonal forecasts from regional climate models (RCMs) (Wood et al., 2005). This approach yields better results when seasonal predictability is enhanced by meteorological forcings~~rather than by initial conditions~~. Climate model outputs may also be more suitable to capture the specific climate conditions at the time of the forecast, whereas ESP-based methods will be limited to the range of past observations and challenged by climate non-stationarity.

The use of climate model outputs in hydrology has however some methodological implications. ~~Outputs~~ For instance, outputs are produced for ~~grid scales that are usually too coarse for streamflow forecasting at the catchment scale. This~~ coarse grid scales, which can lead to errors in capturing forecast uncertainty and ~~introduce significant~~ induce biases. Post-processing (including bias correction ~~techniques and~~ downscaling ~~procedures~~) is usually a necessary first step prior to using climate model outputs to model streamflow. A range of methods has been proposed in the literature~~and the best method usually depends on~~ , with performance varying depending on the modelling chain ~~being investigated~~ and the studied area ~~, with levels of performance that may vary with the forecast horizon or the targeted application~~ (Christensen et al., 2008; Gudmundsson et al., 2012).

~~Bias correction is usually an integral part of post-processing techniques applied to forecasting systems.~~ Weather forecasting has performed bias correction of numerical model outputs through model output statistics (MOS) for decades. In hydrologic ensemble prediction systems, post-processing has become more and more popular in the last decade, particularly for medium-

range ensemble forecasting (e.g. Weerts et al., 2011; Zalachori et al., 2012; Verkade et al., 2013; Madadgar et al., 2014; Roulin and Vannitsem, 2015). In seasonal forecasting, two popular bias correction methods are linear scaling and distribution mapping (Yuan et al., 2015). Linear scaling corrects the mean of the forecasts based on the difference between observed and forecast means, whereas distribution mapping matches the statistical distribution of forecasts to the distribution of observations. These approaches ~~, which can also be applied to improve the performance of ESP forecasts (Wood and Schaake, 2008),~~ focus on increasing forecast skill and reliability, by reducing errors in the forecast mean and improving forecast spread.

Studies comparing different bias correction methods in seasonal hydrological forecasting are still rare in the literature. However, we can find studies reviewing and comparing methods to bias correct RCM outputs and quantify climate change impacts, although their efficiency in this context is still a topic of discussion (Ehret et al., 2012; Muerth et al., 2013; Teutschbein and Seibert, 2013). Teutschbein and Seibert (2012) compared six methods, among which linear scaling and parametric distribution mapping, to bias correct RCM simulations of precipitation and temperature in Sweden. The authors recommended using the distribution mapping method for current climate conditions. They also highlighted the need to assume that bias correction procedures are stationary to correct future climate projections and evaluate changes in flow regimes. In Norway, Gudmundsson et al. (2012) proposed a comparison of eleven methods to bias correct RCM precipitation~~. The methods derived from distribution transformations (e.g. distribution~~ , including distribution mapping based on fitted theoretical ~~distributions), parametric transformations such as linear scaling, and nonparametric transformations such as distribution mapping based on empirical distributions~~or empirical distributions and linear scaling. Their study highlighted the differences between the bias corrections and the necessity to test methods prior to their application. The authors recommended using nonparametric methods since these methods were the most effective to reduce the bias and did not require any approximations of the empirical distributions.

The European Centre for Medium-range Weather Forecasts (ECMWF) produces seasonal forecasts from GCM simulations (Molteni et al., 2011). Weisheimer and Palmer (2014) evaluated the reliability of the precipitation forecasts issued by ECMWF System 4 on a scale ranging from "dangerous" to "perfect". Over the world, ~~precipitation~~ forecasts often fell within the "marginally useful" category. In France, they were ranked as "marginally useful" during wet winters and summers, "not useful" in dry winters, and "dangerous" in dry summers. Kim et al. (2012) also evaluated the skill of System 4 precipitation and temperature forecasts at the global scale. Despite good overall performances, they identified systematic biases, e.g. a warm bias in the North Atlantic. Several studies have proposed to bias correct ECMWF System 4 ~~precipitation~~ forecasts in different contexts. Di Giuseppe et al. (2013) applied a spatially-based precipitation bias correction to improve malaria forecasts. Trambauer et al. (2015) applied a linear scaling method to forecast hydrological droughts in Southern Africa. In the same context, Wetterhall et al. (2015) applied a quantile mapping method to daily precipitation values, and showed that bias correction was able to improve the skill of ~~the system to forecasts~~dry spell forecasts.

Despite these recent works, and to the knowledge of the authors, no previous study has compared bias correction methods and their impact on streamflow forecasting in a systematic way, with a focus on understanding how the main attributes of forecast performance are impacted by bias correction. This paper aims to ~~further investigate the~~ provide insights into the way bias correcting seasonal precipitation forecasts can contribute to the skill of seasonal streamflow predictions, notably in terms of

overall performance, reliability, sharpness and skilful lead time. It investigates the potential of bias corrected ECMWF System 4 forecasts to improve streamflow forecasts at extended lead times ~~. By comparing several~~ over 16 catchments in France. An in-depth comparison of eight variants of linear scaling and distribution mapping methods ~~, the study provides insights into the way bias correcting seasonal precipitation forecasts can contribute to the skill of seasonal streamflow predictions. Forecasts are evaluated~~ applied over the 1981-2010 period ~~in 16 catchments in France~~is presented. Section 2 presents the catchment set, the forecast and observed data, as well as the hydrological model used. Section 3 presents the bias correction methods investigated, as well as the calibration and evaluation frameworks adopted. Results are presented in Sections 4 to 6 for the quality of the raw (uncorrected) and the bias corrected forecasts. In Section 7, conclusions and limitations are discussed.

2 Data and hydrological model

2.1 Seasonal forecasts and observed data

~~This study is based on daily~~ Daily seasonal precipitation forecasts come from ECMWF System 4 ~~. System 4 provides a 51-member forecast ensemble~~ , which provides ensemble forecasts for the next seven months at a TL255 (about 0.7°) spatial resolution~~(Molteni et al., 2011). ECMWF retrospectively produced forecasts~~ , for the period running from 1981 to 2010 ~~These~~ (Molteni et al., 2011). Forecasts are composed of 51 ensemble members for February, May, August and November, and 15 members for the other months. ~~For the purpose of~~ In this study, ~~the 1981-2010 forecasts were aggregated at the catchment scale (i.e.,~~ areal precipitations were computed for each catchment~~). Only~~ , and only the first 90 days of the forecast horizon were considered.

~~Observed precipitation data~~ Daily observed precipitations used for the calibration and evaluation of the bias correction methods come from the 8x8 km grid resolution SAFRAN reanalysis of Météo-France (Quintana-Seguí et al., 2008; Vidal et al., 2010). ~~Daily values are available at an 8x8 km grid resolution covering France.~~ They were also aggregated at the catchment scale. Mean areal potential evapotranspiration was computed for each catchment based on daily observed temperatures from the SAFRAN reanalysis (Oudin et al., 2005). The interannual potential evapotranspiration was then computed in each catchment, i.e. for a given day of the year, we computed the average potential evapotranspiration for this day over all available years (1958 to 2010). Daily streamflow data at the outlet of each catchment come from the French national archive (*Banque Hydro*).

2.2 Studied catchments and hydrological model

The catchment set was selected from the database in Nicolle et al. (2014). It comprises 16 catchments ~~spread over France~~ in France (Fig. 1) with a dominant pluvial regime. Catchments show an average solid fraction of precipitation below 10% and are thus not heavily influenced by snow. Their main characteristics are shown in Table ~~1, and their location in Fig.~~ 1.

We applied the conceptual, reservoir-based GR6J hydrological model (Pushpalatha et al., 2011) at the daily time step. The model ~~is composed of~~ has three reservoirs (one for the production function and two for the routing function), and one unit hydrograph to account for flow delays. The model inputs are daily precipitation and potential evapotranspiration at the

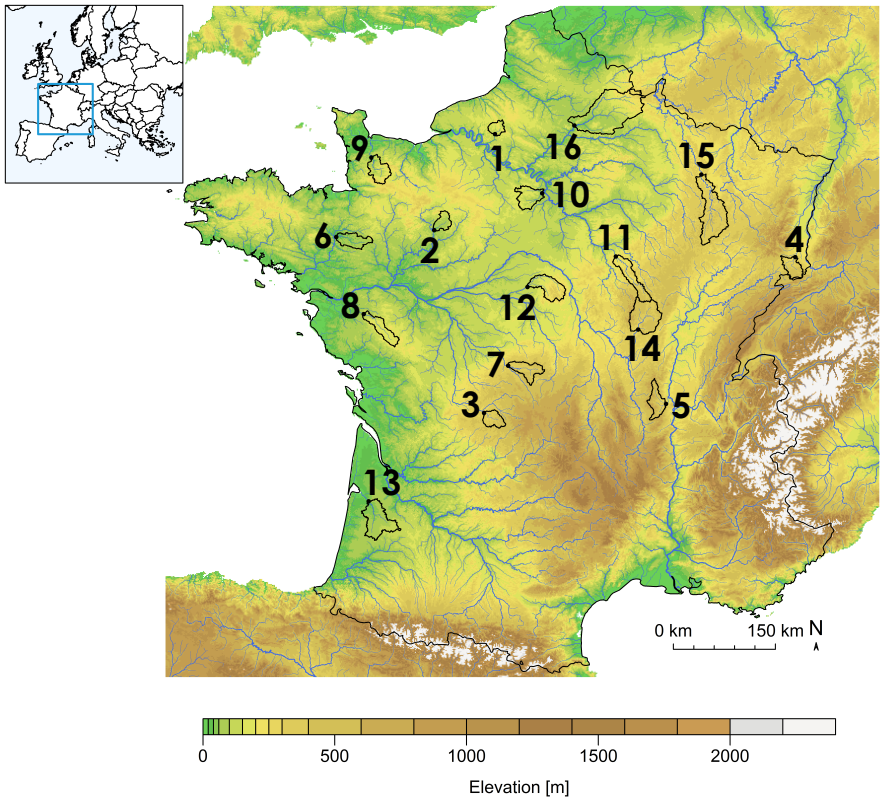


Figure 1. Location of the ~~16~~ studied catchments in France, identified by their numbers (see Table 1~~for details~~).

catchment scale. The model output is the daily streamflow at the catchment outlet. ~~Interannual potential evapotranspiration was used to focus solely on the~~ Here, the series of interannual potential evapotranspiration corresponding to the forecast period was systematically used as input to the hydrological model. With this setup, we aimed to isolate the influence of precipitation forecast inputs on the quality of streamflow forecasts. This setup is also consistent with the fact that our catchment set is dominated by a pluvial regime. The model was calibrated in each catchment with the Kling-Gupta Efficiency (Gupta et al., 2009) 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 0.95 to 1.02. When the model is applied to forecast streamflow, the model states are initialized by running the model in simulation mode for the year preceding the forecast date. The last observed streamflow ~~at the time of forecast~~ is then used to update the levels of the routing reservoirs before issuing the forecast~~s~~.

3 Methods

3.1 Overview of the calibration-~~evaluation~~ approach

~~Bias correction methods were calibrated and evaluated in each catchment over the 1981-2010 period. The one-year-leave-out~~ The leave-one-year-out cross-validation method (Arlot and Celisse, 2010) was applied to calibrate ~~and evaluate the methods~~ the bias correction methods in each catchment over independent periods within the 1981-2010 period. Given a target application year~~within the study period~~, all available years but the target year are used in the calibration process. Results of the calibration are then applied to the target ~~application~~ year and bias corrected forecasts are evaluated against observations.

In the calibration step, we considered two approaches~~. The simplest calibration uses~~ : (1) all days of the years within the calibration dataset ~~. An alternative approach consists in calibrating~~ are used, (2) the bias correction methods are calibrated for each calendar month. Additionally, since we are dealing with forecasts issued up to 90 days ahead, and since forecast performance varies with lead time, calibration also takes the lead time into account. ~~In this study, lead~~ Lead times were grouped from 1 to 30 days, 31 to 60 days and 61 to 90 days ahead. The calibrated bias correction factors are then applied to the daily values of the ensemble precipitation forecasts in the target ~~application~~ year. The hydrological model is forced by ~~precipitation forecasts and streamflow ensemble forecasts are obtained. The modelling chain is applied to~~ raw and bias corrected precipitation forecasts~~.~~ ~~Precipitation and streamflow forecasts are evaluated with deterministic and probabilistic scores commonly used in ensemble forecasting~~, which results in streamflow ensemble forecasts.

3.2 Bias correction methods

We applied the linear scaling (LS) and the distribution mapping (DM) methods to the raw System 4 precipitation forecasts. The DM method was applied ~~following three variants:~~ considering the empirical distribution of monthly values (EDM), a fitted gamma distribution of monthly values (GDM), and the empirical distribution of daily values (EDMD). Each method was applied on a monthly (-m) or a yearly (-y) basis (Table 2).

3.2.1 Linear scaling of precipitations

~~The LS method~~ LS consists in correcting the monthly mean values of the forecasts to match the monthly mean values of the observations. A scaling factor (or bias) is calculated considering the ratio between the observed and the forecast (ensemble mean) values. A scaling factor higher (lower) than 1 indicates that the mean ensemble forecast underpredicts (overpredicts) the mean observed value. A value of 1 indicates no bias in the forecasts. The scaling factor obtained through calibration is then applied as a multiplicative factor to correct raw daily precipitation forecasts.

3.2.2 Distribution mapping of precipitations

~~The DM method~~ DM consists in correcting the precipitation forecasts so that their statistical distribution matches that of the observations. There are several ways to match forecast and observed distributions or quantiles, and existing techniques

mainly differ on how the ~~forecast and observed~~ cumulative distribution functions (CDF) are considered. In some techniques, a parametric distribution is fitted to the ~~forecast and observed~~ datasets, while in others the empirical distributions and linear interpolations between data points or estimated quantiles are considered. ~~In any case, observed and forecast CDFs must be determined from long data series.~~

In this study, the calibration of the DM method was first carried out considering empirical (EDM) and gamma-fitted (GDM) distributions of observed and forecast (ensemble mean) precipitation values averaged monthly. A third variant considered directly the empirical distribution of the daily values of the ensemble members (EDMD). These variants are listed in Table 2. After calibration, bias correction is applied to the daily precipitation forecasts of each ~~application~~ target period. In the case of EDM and GDM, ~~all daily values are~~ the monthly values are first corrected based on the ~~correction suited to their monthly average~~distribution mapping procedure. Then, for a given month, the ratio of the corrected monthly value and the non-corrected one is used to correct all daily values within this month. In the case of EDMD, each daily precipitation value of each forecast member is corrected individually.

3.3 Evaluation framework

~~For each catchment, daily forecasts are issued once every month, up to 90 days ahead, during the 1981-2010 period.~~ The quality of the forecasts was evaluated ~~at the weekly time step (i.e., daily forecasts and observations are averaged over the week). Scores were computed~~ as a function of lead time and for the winter (December-January-February), the spring (March-April-May), the summer (June-July-August) and the autumn (September-October-November) seasons. Four criteria were used to assess reliability, sharpness, accuracy and overall performance of the ~~ensemble~~ forecasts (Gneiting et al., 2007; Eslamian, 2015; Musy et al., 2015).

3.3.1 Evaluation criteria

Reliability is a forecast attribute that refers to the statistical consistency between observed frequencies and forecast probabilities. In this study, it ~~is~~ was evaluated with the Probability Integral Transform (PIT) diagram (Gneiting et al., 2007; Laio and Tamea, 2007). The PIT diagram is the cumulative distribution of the ~~positions of the observation within the cumulative forecast distribution . A reliable forecasthas a PIT diagram superposed~~ PIT values, which are defined by the values of the predictive distribution function at the observations, computed at each time step. In the case of a reliable forecast, the observations uniformly fall within the predictive distribution and the PIT diagram coincides with the 1:1 diagonal. If the PIT diagram ~~shows a curve~~ is systematically above (below) the diagonal, the observed values are too frequently located in the lower (upper) parts of the forecast distribution, suggesting a systematic bias of the forecasts towards overprediction (underprediction). If the ~~points in the diagram are too concentrated in the vicinity of the end points (0 and 1), forecasts are too narrow and observations fall more frequently than expected on~~ PIT diagram tends to resemble a horizontal line, observations fall too frequently in the tails of the forecast distribution, indicating that forecasts are too narrow. On the contrary, ~~too many points concentrated~~ if the PIT diagram is closer to a vertical line, too many observations fall in the midrange ~~indicate a forecast distributionthat is~~ of the forecast distribution, indicating that forecasts are too wide. We also represented the Kolmogorov significance bands at +0.1

and -0.1 from the bisector, which ensure a 5% significance. In order to numerically compare results among catchments, we also computed the area between the curve of the PIT diagram and the 1:1 diagonal, as proposed by Renard et al. (2010). The smaller this areais, the more reliable the ensemble.

Sharpness is a property of the forecasts only. It refers to the concentration of the predictive distribution and indicates how spread the members of an ensemble forecast are. In this study, sharpness was evaluated with the 90% interquantile range (IQR; Gneiting et al., 2007)(IQR), i.e. the difference between the 95^{th} th and the 5^{th} th percentiles of the forecast distribution. The final IQR score is the average of the interquantile range at each time step of the evaluation period. The narrower the IQRis, the sharper the ensemble. In this study, we considered that, given two reliable systems, the sharpest one is the best (Gneiting et al., 2007).

The accuracy of the forecasts is assessed with the mean absolute error (MAE). The MAE computes the average (over the evaluation period) of the absolute difference between the forecast ensemble mean and the observed value. Smaller MAE values correspond to more accurate forecasts.

LastLastly, the Continuous Rank Ranked Probability Score (CRPS) evaluates the overall performance of the forecasts. It is defined as the integral of the squared distance between the cumulative distribution of the forecast members and a step function for the observation (Hersbach, 2000). The CRPS score is the average of this integral computed at each time step of the evaluation period. The lower the CRPSis, the better the overall performance of the forecasts.

3.3.2 Skill scores

Forecast skill is evaluated by comparing the performance of a given forecast system with the performance of a reference forecast. The skill score is computed for a given lead time i.

$$SkillScore_i = 1 - \frac{Score_i^{Syst}}{Score_i^{Ref}} \quad (1)$$

When the skill score is superior (inferior) to zero, the forecast system is more (less) skilful than the reference. When the skill score it is equal to zero, the system and the reference have equivalent skill.

The skill scores were computed for the probabilistic scorespresented in the previous section (. They are noted PITSS, IQRSS and CRPSShereafter). The reference used to evaluateprecipitation forecasts is based on past observations and is representative of the catchment climatology: for a given day and year, it is the ensemble of precipitation values observed on that same Gregorian day in other years of the observation period . The reference used to evaluate streamflow forecasts (1958 to 2010). Two reference streamflow forecasts are used. The first is the Ensemble Streamflow Prediction (ESP), which corresponds to the streamflow ensemble obtained when the reference precipitation ensemble is used as input to the hydrological model. Pappenberger et al. (2015) highlight the importance of the reference chosen to compute skill scores and list a number of options for streamflow forecasting. The ESP is a commonly used method in seasonal forecasting. It allows applying the same hydrological modelling setup to both the precipitation forecasts and the reference precipitation ensemble. Therefore, differences in performance are mainly due to differences between the precipitation inputs to the model. One would expect that precipitation and streamflow forecasts perform better than precipitation climatology or ESP, at least in the first lead times. At

~~longer lead times, natural variability should end up being a sound forecast. In our study, we also used an ensemble~~ The second reference is based on past streamflow observations (on the same day as the given forecast day, in a 36- to 52-year period running up to 2010) to evaluate performance. This ~~allows to use as reference an ensemble that~~ reference ensemble does not use any precipitation forecasts or hydrological model.

Finally, several studies have shown that the ensemble size induces a bias when computing skill scores with ensembles of different sizes. This bias usually leads to an underestimation of the skill of the forecast system when the system has fewer members than the reference. Ferro et al. (2008) provide a synthesis of previous studies on the influence of ensemble size on probability scores and propose a correction factor to remove the bias in the computation of CRPS skill scores. This correction was applied to compute the CRPSS in this study. Since the ensemble size of System 4 precipitation forecasts varies with the month, we used the ensemble size averaged over one year.

3.3.3 Gain in lead time from bias correcting seasonal forecasts

~~Skill scores can be computed to indicate~~ To investigate the gain in performance brought by bias correction methods~~. To that effect~~, we use the raw (uncorrected) forecasts as reference in the computation of the skill scores. An indicator of forecast performance can be derived~~from the evolution of these skill scores~~: the lead time up to which bias corrected ~~seasonal~~ forecasts have more skill than raw forecasts. Nicolle et al. (2014) defined an indicator named UFL (Useful Forecasting Lead time) as ~~"the~~ the first "lead time beyond which model performance is not at least 20% better than benchmark performance". Here, we considered the lead time beyond which the seven-day moving average of the skill score becomes negative. UFL values were then grouped in four categories: (1) None: no improvement over the forecast reference, (2) <30: gain up to 30 days, (3) <60: gain greater than 30 days and up to 60 days and (4) >60: gain greater than 60 days.

4 Quality of the raw seasonal forecasts

4.1 Performance of raw precipitation forecasts

Figure 2 presents the evolution of IQRSS and CRPSS with lead time, for winter (DJF) and summer (JJA). Each line corresponds to a catchment. Skill in sharpness and overall performance is very similar in winter and in summer (as well as in spring and autumn, not shown). Precipitation forecasts are overall sharper than historical precipitations in the large majority of catchments and up to long lead times. Some exceptions appear for lead times longer than three weeks, and especially in winter (wetter season in the majority of catchments). In terms of overall performance, precipitation forecasts clearly have skill up to two to three weeks ahead for 7-day averaged areal precipitation. At longer lead times, they are equivalent or perform slightly worse than historical precipitations.

Figure 3 shows the PIT diagrams for lead times of 30 and 90 days, for winter and summer. Grey lines represent the reliability of historical precipitations and coloured lines represent the reliability of System 4 precipitation forecasts in each catchment. Dotted lines represent ~~deviations of +0.1 and -0.1 from the bisector~~the Kolmogorov significance bands to ensure a

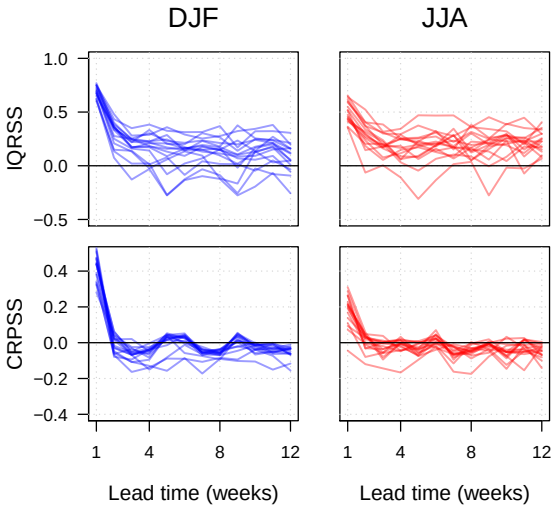


Figure 2. Skill of raw weekly precipitation forecasts as a function of the lead time for all catchments and ~~all~~ for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is historical precipitations. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

5% significance test. The two seasons yield very similar results (also observed in spring and autumn, not shown). In all catchments and for both lead times, historical precipitations are reliable, as expected. Seasonal precipitation forecasts also show some reliability, but tend to overpredict precipitations in both seasons and at both lead times. The concentration of points in the zero end points in most of the curves ~~of the System 4 forecasts~~ shows that low values of the observations are too often falling in the lower tail of the forecast distribution. This effect tends to decrease with increasing lead time. This is an indication that forecasts are too narrow and overpredict the lowest observations. It can also ~~translate~~ indicate a difficulty of the system to forecast null precipitation.

4.2 Performance of raw streamflow forecasts

Streamflow forecasts are generated by using raw ~~seasonal~~ precipitation forecasts as input to the hydrological model. Forecast skill is evaluated using the ESP method as reference (Fig. 4). Differences in forecast skill between the winter and summer seasons are more noticeable when evaluating streamflow forecasts rather than precipitation forecasts. Streamflow forecasts generated from raw precipitation forecasts are sharper than ESP up to twelve weeks ahead in most catchments (IQRSS above zero in Fig. 4). Approximately, only four catchments stand out in both seasons with lower skill than ESP (six in spring and one in autumn, not shown). However, even in these catchments, sharpness can be improved using seasonal precipitation forecasts for lead times up to three weeks in winter (as well as in spring and autumn, not shown). Concerning overall performance (CRPSS in Fig. 4), skill can be observed for lead times up to four weeks in some catchments. ~~In winter, as well as in spring and autumn (not shown), this is observed in the majority of catchments, while in summer, this concerns only a couple of~~

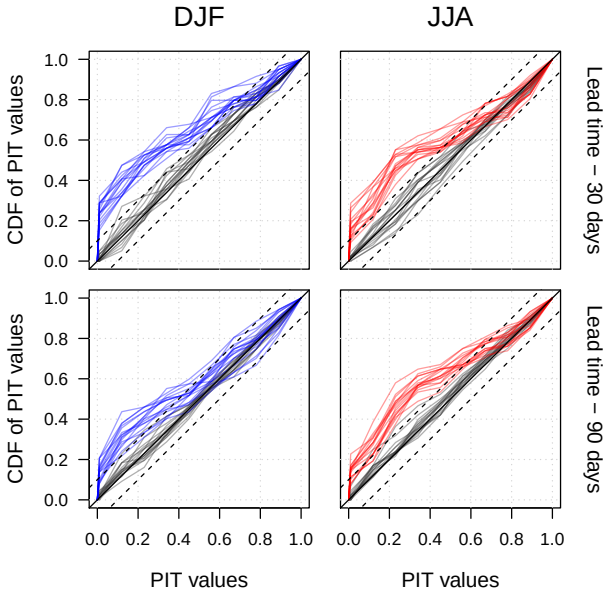


Figure 3. PIT diagram of raw precipitation forecasts (coloured lines) and historical precipitations (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

~~catchments.~~ At longer lead times, ESP and ~~streamflow forecasts generated from raw precipitation forecasts~~ raw streamflow forecasts are equivalent in most catchments for the winter season. In summer, as well as in spring and autumn (not shown), the difference in skill at longer lead times is more pronounced and most catchments have a ~~clearly~~ negative skill in terms of overall ~~forecast~~ performance.

PIT diagrams are shown for each catchment, for the winter and summer seasons, and for lead times of 30 and 90 days (Fig. 5). In winter and spring (not shown), ESP ~~forecasts and seasonal streamflow forecasts generated from raw precipitation forecasts~~ and raw streamflow forecasts show good reliability, although the curves above the diagonal indicate that forecasts are slightly overpredicting streamflow. Streamflow forecasts for the autumn season (not shown) also show good reliability, but with a tendency to underpredict streamflow. In summer (Fig. 5, right), streamflow forecasts from both ~~, ESP forecasts and forecasts generated from raw seasonal precipitation~~ ESP and raw forecasts, show problems in forecast reliability. PIT curves clearly indicate a concentration of points at the end points of the diagram and, consequently, narrow ensemble forecasts. In most catchments, 20% to 60% of observed values fall in the lowest interval of the forecast distribution or below it~~, i. e., outside the forecast range.~~ . Although reliability is slightly improved with lead time, streamflow ~~ensemble~~ forecasts remain under-dispersive at 90 days of lead time. This could be the result of at least two factors acting alone or jointly: a difficulty of the hydrological model to reach the lowest streamflow values in the simulations of the recession periods, and the influence of not considering uncertainties in the hydrological initial conditions at the time of forecasting.

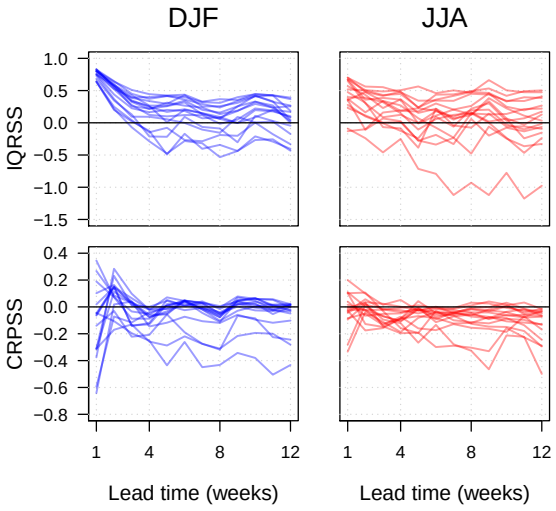


Figure 4. Skill of weekly streamflow forecasts from raw precipitation forecasts as a function of the lead time for all catchments and ~~all~~ for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is Ensemble Streamflow Prediction. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

4.3 Summary of the quality of raw seasonal forecasts

Skill in the overall performance of System 4 raw precipitation forecasts, at the catchment scale and over a reference forecast based on past observed precipitations, was observed up to two to three weeks in the studied catchments. When looking at streamflow forecasts generated from ~~the input of raw seasonal forecaststo a hydrological model~~raw precipitation forecasts, skill over the traditional ESP method was observed up to four weeks, but only in few catchments. The asset of System 4 raw precipitation forecasts and related streamflow forecasts over historical precipitations and ESP, respectively, resides mainly in their sharpness. However, the evaluation of forecast quality shows also that forecasts are often too narrow and suffer from underprediction or overprediction. Improving forecast reliability, while maintaining forecast sharpness is clearly a challenge. ~~In the following section, we investigate the presence of biases in System 4 precipitation forecasts and the impact of bias correction on seasonal precipitation and streamflow forecasts.~~

5 Bias correction of seasonal precipitation forecasts

5.1 Overview of the effectiveness of the bias correction methods

Forecast bias, i.e. the ratio between the mean observation and the average forecast ensemble mean, was computed for each catchment over the 1981-2010 period. The bias was computed for each calendar month, but also considering the whole year. Figure 6 shows the biases expressed as deviations from 1 (i.e., $1-Bias$), before and after applying the bias correction methods.

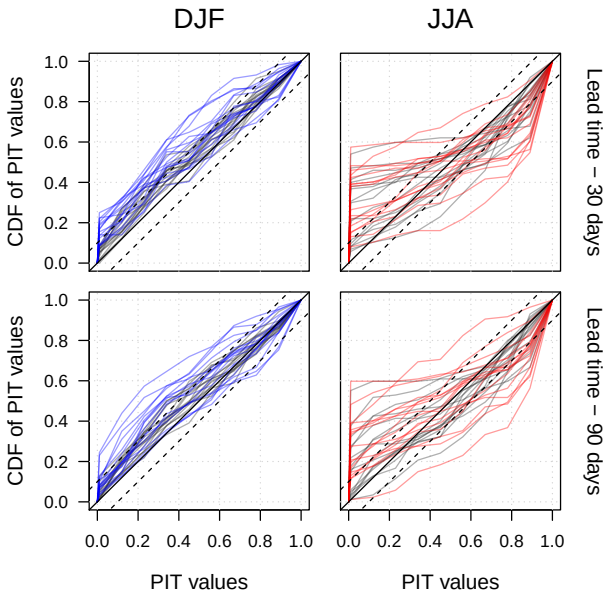


Figure 5. PIT diagram of streamflow forecasts from raw precipitation forecasts (coloured lines) and Ensemble Streamflow Prediction (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

It illustrates the results obtained in four catchments at the ~~2-month~~ month-2 lead time (i.e., ~~considering the~~ forecasts issued for day 31 to day 60~~in the forecast range~~). The effectiveness of each bias correction method can be ~~easily seen from the coloured charts~~observed: unbiased forecasts have a deviation equal to 0 (white~~colour~~); positive deviations (red~~colour~~) and negative deviations (blue~~colour~~) indicate overprediction and underprediction, respectively. A deviation equal to 0.75 (-3) can be interpreted as the mean forecast being four times larger (smaller) than the mean observation. Overall, when computing the deviations for all monthly lead times of the forecast range, we observed that the biases vary more with the calendar month of the forecast horizon than with lead time. For this reason, we only show the ~~2-month~~ month-2 lead time.

In general, seasonal forecasts tend to overpredict precipitations over the year in most catchments. Overprediction tends to occur near the end of the winter (rainy) season and throughout the spring season. Conversely, precipitations tend to be underpredicted from the end of the summer (dry) season and until the beginning, and sometimes throughout, the autumn season. The four selected catchments illustrate the variety of conditions we encountered in the bias correction analysis. In catchment 2, precipitations could be considered unbiased when carrying the analysis over the year. However, this result hides monthly underpredicting and overpredicting biases which compensate over the year. In this catchment, forecasts tend to overpredict from February to June and underpredict from July to October. The yearly result may also be a reflection of the lack of important biases in the months of December and January, which are, climatologically, the rainiest months~~in this catchment~~. This type of variation in bias was also observed in catchments 6, 11, 12 and 13. In catchment 4, precipitation forecasts are strongly over-

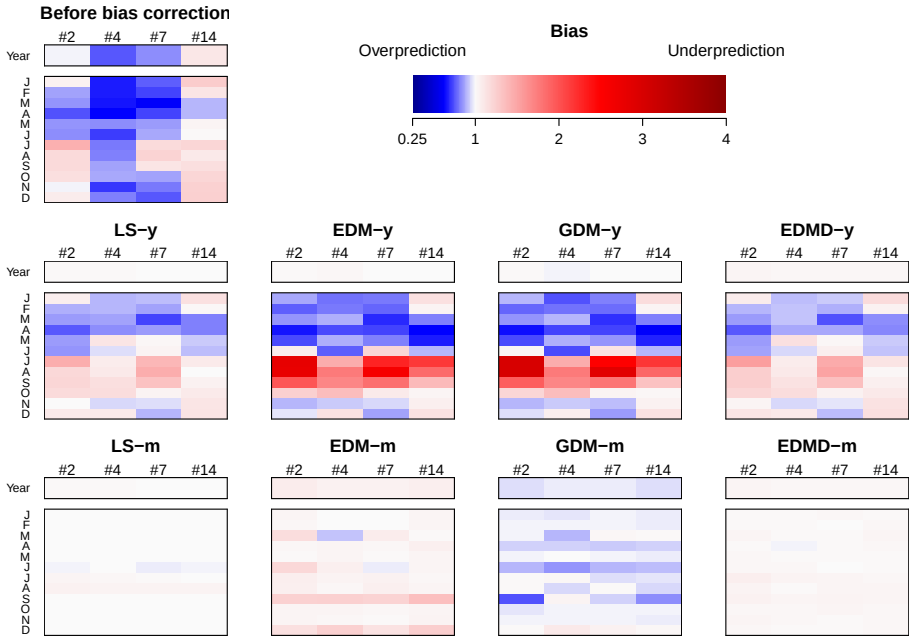


Figure 6. ~~Deviation of the~~ Bias in precipitation ~~bias from 1,~~ for catchments 2, 4, 7 and 14, over the 1981-2010 period. The ~~deviation~~ bias is shown for the whole year (top line) and for each calendar month. The bias is only shown for lead times between 31 and 60 days. Blue-shaded areas ~~(negative values)~~ represent a tendency of ~~underpredicting~~ overpredicting precipitations and red-shaded areas ~~(positive values)~~ represent a tendency of ~~overpredicting~~ underpredicting precipitations. The top left graph represents the bias of raw precipitation forecasts, and each of the other graphs represents the bias after applying one of the bias correction methods.

predicting observations in all calendar months and thus over the year. This catchment stands out because in no other catchment do we observe a similarly strong and systematic bias. This catchment is the one located at the ~~most eastern~~ easternmost part of France. Its main river (l'Ill) is a tributary of the Rhine river. It has its sources in the Jura mountains and receives several tributaries from the Vosges mountains. In catchment 7, precipitations are overpredicted over the year, with the strongest positive deviations concentrated during the rainy season, basically from November to April. The same behaviour is found in catchments 5, 10 and 15. Interestingly, catchments with a clear overprediction, i.e. catchments following the patterns depicted in Fig. 6 for catchments 4 and 7, correspond to the catchments in which System 4 raw precipitation and streamflow forecasts showed low skill in sharpness and/or overall performance. ~~Last~~Lastly, catchment 14 is representative of catchments 1, 3, 8, 9 and 16 in the database. Forecasts slightly underpredict precipitations over the year, with a tendency to underpredict precipitations in all seasons but the spring season, ~~whose~~ when precipitations are slightly overpredicted.

Figure 6 also presents the remaining biases after the application of the eight bias correction methods~~to the raw precipitation forecasts. We present the results over the whole year and for each month. The same four selected catchments illustrate the results for the 2-month lead time.~~ . All correction methods are effective to correct biases of precipitation forecasts over the year. However, ~~this is not observed in the bias correction for each calendar month. Results~~ results for the methods calibrated

on a yearly basis (LS-y, EDM-y, GDM-y, EDMD-y) show that the absence of bias over the year is mainly achieved through an effect of compensation between over and underprediction among the calendar months. Particularly EDM-y and GDM-y methods show a strong pattern of monthly biases, even after bias correction, towards overprediction of precipitations in winter and spring, and underprediction in summer and autumn.

By constructionWhen looking at monthly biases, monthly calibrated methods perform much better when looking at monthly biasesby construction. LS-m and EDMD-m are particularly effective in all catchments. Forecasts corrected with EDM-m tend to slightly underpredict precipitations, while forecasts corrected with GDM-m tend to overpredict precipitations. This may be an effect of the application of distribution mapping based on monthly values. Distribution mapping requires that the time structure of forecast and observed precipitation are coherent, so that upper forecast values are shifted towards upper observed values and conversely. However, raw monthly forecast means from System 4 do not always reproduce the time structure of monthly observations and often fail to reach extreme monthly values. Therefore, correction factors obtained with a distribution mapping based on monthly values show poorer performance, and the method can wrongly increase or decrease daily precipitation values.

Comparison of bias correction factors for LS and EDMD methods

The LS and EDMD methods showed more effectiveness in reducing bias in the precipitation forecasts. In order to better understand how the two methods compare, we plotted in Fig. 7 their correction factors for catchment 7 over the 1981-2010 period for the 2-month lead time. Black lines represent correction factors from LS. Each day, one correction factor is applied to all members of the ensemble forecast at the 2-month lead time. Grey-shaded areas represent the range of correction factors applied with EDMD, and darker grey lines represent the median correction factor. For EDMD, each precipitation value has a specific correction factor depending on its probability of occurrence. Therefore, for a given day and lead time, the number of correction factors is equal to the number of ensemble members.

LS-y provides relatively constant bias correction factors over the study period. Since, on average, precipitations in catchment 7 are overpredicted by System 4 forecasts, this correction factor is smaller than 1. The bias correction factors are obtained with the one-year-leave-out calibration framework. It is interesting to note that removing one year within the 30 years of the calibration period has little impact over the calibrated correction factors, even for an extreme dry year such as 1989 in this catchment. With EDMD-y, correction factors vary for each day of the study period. These factors remain smaller or close to 1. Their median values are smaller than the LS-y correction factors and the maximum values are slightly greater than the LS factors. When calibrated monthly, correction factors obtained with LS-m depict a variation, ranging from 0.6 to 1.2. They present a recurring pattern over the year, which follows what was shown in Fig. 6, i.e., that precipitations in catchment 7 are, on average, overpredicted during the winter and spring seasons, leading to correction factors smaller than 1, and underpredicted from July to September, leading to bias correction factors greater than 1. This pattern in the factors indicates that the LS method might be further simplified to provide correction factors that would solely vary with the calendar month, regardless of the year, or in the case of LS-y, be constant over the target period. Correction factors computed with EDMD-m present a similar pattern to the one observed with LS-m, but their range is more variable, with values between 0 and 1.4. This method is

particularly interesting because, as opposed to LS, it also corrects the frequency of precipitation days, given the null values of some correction factors.

Bias correction factors applied to each day of the record period with the LS and EDMD methods. Correction factors are only shown in the case of catchment 7 and for the second month lead of the precipitation forecasts. The top graph presents correction factors obtained with LS and EDMD calibrated over the whole year, and the bottom graph presents correction factors obtained with LS and EDMD calibrated monthly.

5.2 Impact of bias correction on the useful forecasting lead time

The four criteria used to evaluate reliability, accuracy, sharpness and overall performance were applied to the precipitation forecasts bias corrected with each of the eight bias correction methods. They were also applied to the seasonal streamflow forecasts generated from inputting the different bias corrected precipitation forecasts to the hydrological model. Skill scores were computed with the raw seasonal precipitation Skill scores of bias corrected precipitations and related streamflow forecasts were computed using raw forecasts as referenceforecast for precipitation, and with the (raw) streamflow forecasts generated from raw precipitation forecasts as reference forecast for streamflow. For each variable (precipitation and streamflow), each evaluation criterion, each bias correction method, each catchment and each catchment and season, we obtained the corresponding UFL (Useful Forecasting Lead time) . We then and evaluated the proportion of catchments falling in each UFL group (as defined in Section 3.3.3). Results are shown in Fig. 8 7 and Fig. 98, for precipitation and streamflow forecasts, respectively.

In Fig. 87, the two bias correction methods that stand out regarding overall performance (CRPS), in all seasons, are LS and EDMD. This is in accordance with our previous results on the efficiency of each method to correct biases. When looking more closely at improvements in the PIT criterion, as measured by the UFL, EDMD clearly stands out from the other methods. The proportion of catchments with skill improvement over raw precipitation forecasts is almost always 100%, and skill is often extended up to 60 days and more. The other methods are quite equivalent to each other, although LS performs slightly better, with greater improvements in larger proportions of catchments, especially in winter and spring, for reliability (PIT), accuracy (MAE) and overall performance (CRPS). In terms of sharpness (IQR), the best performing method varies with the season. Precipitation forecasts in spring (MAM) are sharper when corrected with methods calibrated monthly, while forecasts in summer and autumn are sharper with methods calibrated yearly. To effectively address the tendency to overestimate spring precipitations, the multiplicative correction factor of a monthly calibrated bias correction for the spring season will be smaller than 1, and much smaller than the correction factor obtained with a yearly calibrated correction. Therefore, the spring interquartile range will be further reduced by the method calibrated monthly than by the method calibrated yearly. This reasoning only applies to LS, EDM and GDM since EDMD corrects each ensemble member independently.

Figure 9 shows the results for the streamflow forecasts. 8 shows that LS and EDMD methods are able to extend the lead time of bias corrected predictions streamflow forecasts further than other methods, and for a higher proportion of catchments in the large majority of seasons and criteria. Again, EDMD methods yield the best improvements in reliability. LS yields results slightly better than EDMD in sharpness and accuracy. EDM and GDM clearly have lower performance, except in some cases in sharpness and for spring and summer.

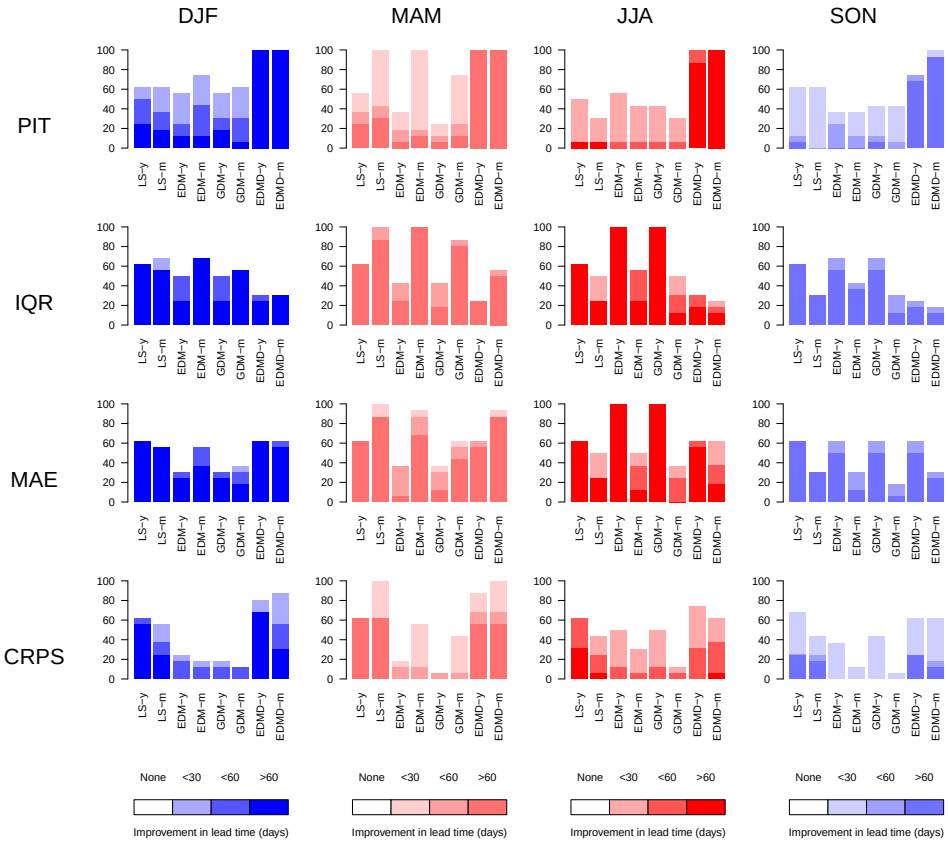


Figure 7. ~~Number~~ Fraction of catchments (%) in each UFL value category, i.e. ~~number~~ fraction of catchments in which bias corrections increase the lead time up to which seasonal precipitation forecasts have skill ~~in regards~~ with respect to raw seasonal precipitation forecasts. Each row corresponds to an evaluation criterion and each column corresponds to a season. Colour shades indicate the UFL category, i.e. the lead time up to which precipitation forecasts are improved.

5.2 Summary of the comparison of bias correction methods

In general, LS and EDMD bias correction methods show good performance for precipitation and streamflow forecasts, although in a distinct way. While EDMD clearly improves forecast reliability, LS shows better performance in improving sharpness ~~. In terms of streamflow forecasts, LS and EDMD are the methods that offer the best performance. Again, EDMD may be preferred if focus is placed on forecast reliability, while LS may be preferred if sharpness and accuracyare the criteria one is looking to improve~~and accuracy. Since streamflow forecasts generated from raw System 4 precipitation forecasts are already, in most of the studied catchments, sharper than the ESP reference, but lack reliability (as shown in Fig. 4 and Fig. 5), it seems appropriate to give priority to a correction method that improves reliability, while providing good overall performance. Therefore, in the following, we will only consider the monthly calibrated version of EDMD (EDMD-m) to further investigate the skill of bias

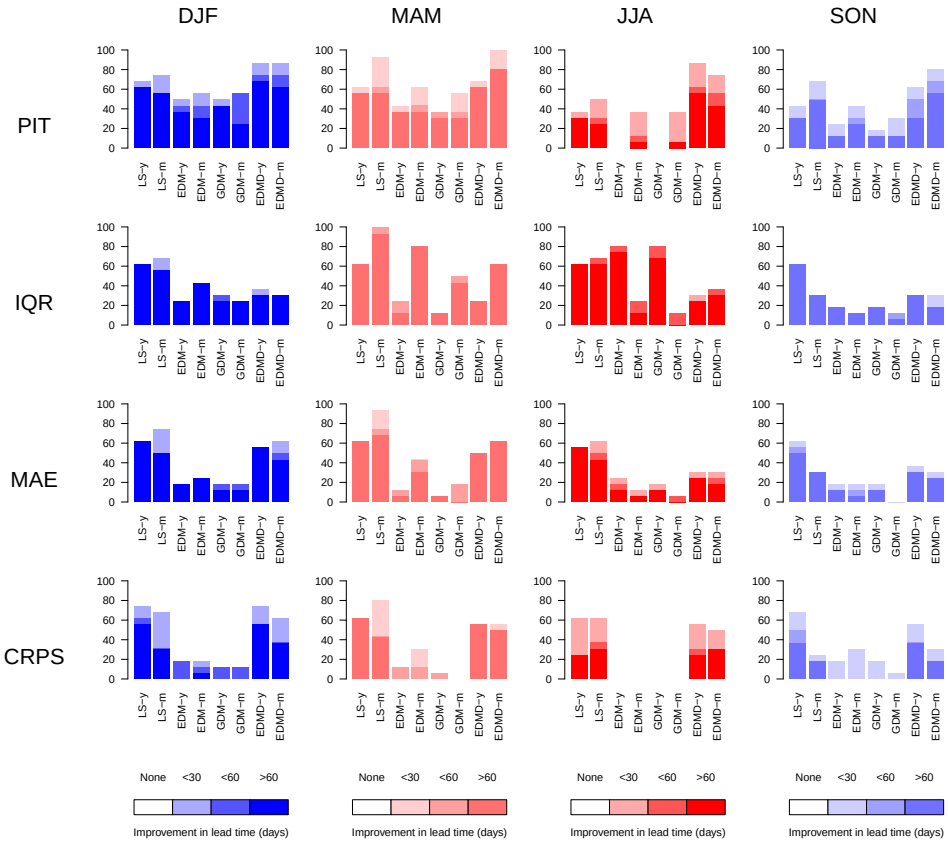


Figure 8. ~~Number~~ Fraction of catchments (%) in each UFL value category, i.e. ~~number~~ fraction of catchments in which bias corrections increase the lead time up to which seasonal streamflow forecasts have skill ~~in regards~~ with respect to seasonal streamflow forecasts generated from raw seasonal precipitation forecasts. Each row corresponds to an evaluation criterion and each column corresponds to a season. Colour shades indicate the UFL category, i.e. the lead time up to which streamflow forecasts are improved.

corrected seasonal forecasts~~in the 16 selected French catchments~~. The monthly version is chosen to ensure that monthly biases are removed and that the correction will perform relatively equally in all seasons, while avoiding the "mis-estimation" of forecast skill (Hamill and Juras, 2006).

6 Skill scores of bias corrected seasonal forecasts

6.1 Performance of bias corrected precipitation forecasts

Figure ~~10~~ 9 (for sharpness and overall performance) and Fig. ~~11~~ 10 (for reliability) present the skill of seasonal precipitation forecasts bias corrected with EDMD-m. Skill scores are computed with historical precipitation as the reference. In order to

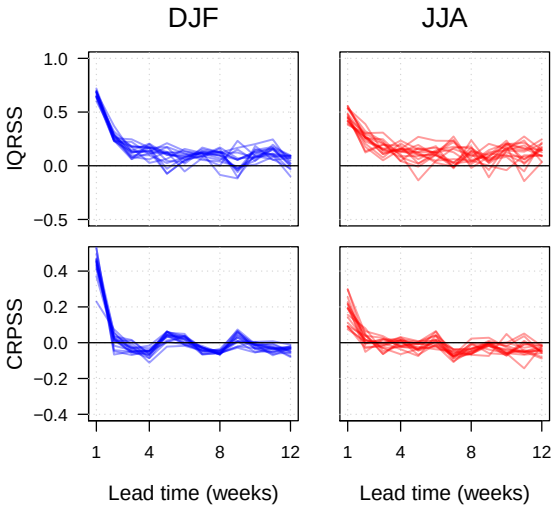


Figure 9. Skill of weekly precipitation forecasts corrected with EDMD-m as a function of the lead time for all catchments and ~~all~~ for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is historical precipitations. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

better evaluate the impact of bias correction on forecast skill, the y-axes in Fig. ~~10~~ 9 are the same as in Fig. 2. The comparison of these two figures shows that bias correcting the raw ~~System 4~~ forecasts reduces the differences in skill between catchments. After bias correction, catchments present very similar evolutions of the skill with the lead time. ~~We can also infer that, after bias correction, in~~ In some catchments, the values of IQR ~~and CRPS are lowerthan before bias correction. Nevertheless,~~ are lower, but bias corrected forecasts remain sharper than the reference (i.e., skill scores are ~~always~~ mostly greater than zero). In the catchments where the raw forecasts performed worse than historical precipitations (i.e., skill scores lower than zero in Fig. 2), bias corrected forecasts become sharper and gain skill~~in regards to the reference~~. Forecast skill in overall performance (CRPSS) is observed up to two to three weeks ahead~~, after which forecasts attain skill equal to that of the reference forecast~~. Skill is improved in catchments that performed worse than the reference prior to bias correction (i.e., skill scores lower than zero in Fig. 2). Figure ~~10~~ 9 illustrates these findings for winter (DJF) and summer (JJA), but results are similar for spring and autumn (not shown).

Figure ~~11~~ 10 shows that the most remarkable improvement in performance due to bias correction is achieved in reliability. While precipitation forecasts had a tendency to overpredict prior to bias correction, bias corrected precipitations are reliable in all catchments. Figure ~~11~~ 10 shows the results for winter and summer, and for lead times of 30 and 90 days, but conclusions are similar in the other seasons and lead times (not shown). Even though a slight tendency to overpredict ~~precipitations~~ precipitation remains in winter for short lead times, the improvements are noticeable. The EDMD-m bias correction was able to address the concentration of points in the zero end point observed in Fig. 3 for the raw forecasts.

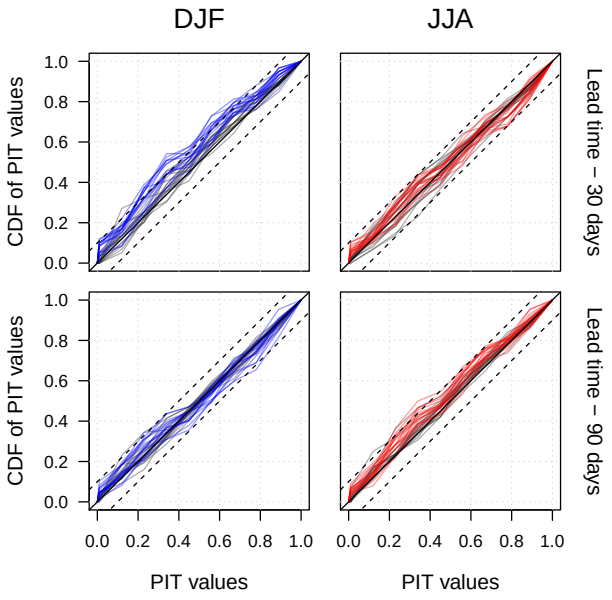


Figure 10. PIT diagram of precipitation forecasts corrected with EDMD-m (coloured lines) and historical precipitations (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

6.2 Performance of bias corrected streamflow forecasts

The quality of the streamflow forecasts generated from the precipitation forecasts corrected with EDMD-m is investigated in Fig. ~~12~~ 11 and Fig. ~~13~~ 12 (IQRSS and CRPSS) and in Fig. ~~14~~ 13 (PIT diagrams). These figures can be compared to Fig. 4 and Fig. 5 ~~which were obtained from the analysis of streamflow forecastsgenerated from raw precipitation forecasts~~for raw streamflow forecasts. As seen with precipitation forecasts, bias correction also reduces the differences in streamflow forecast skill between catchments and seasons (Fig. ~~12~~11). Again, this translates into a loss in skill in catchments with the sharpest ensemble forecasts before bias correction, but also in a gain in skill in catchments where raw streamflow forecasts had negative skill. 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)~~most catchments and for most lead times. In terms of overall performance (CRPSS), the skill of streamflow forecasts was largely improved, especially in catchments that had very low skill prior to bias correction (i.e., CRPSS values well below zero in Fig. 4). In winter, autumn and spring, skill over the ESP reference is observed up to four weeks ahead in several catchments (even up to five weeks ahead in spring and autumn), while in summer, it is observed up to two to three weeks. At longer lead times, streamflow forecasts show an overall performance equivalent or slightly lower than the performance of the ESP method. Some studies use past streamflow observations (referred to as streamflow climatology) as the reference forecast to assess the skill of streamflow forecasts (e.g. Trambauer et al., 2015; Wetterhall et al., 2015). Figure ~~13~~ 12 shows the skill in overall performance

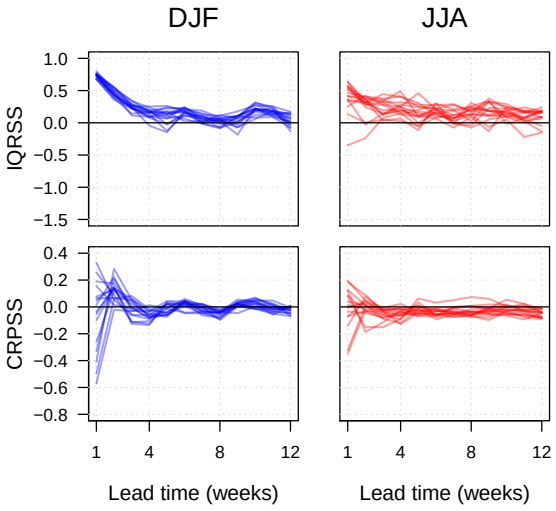


Figure 11. Skill of streamflow forecasts obtained from precipitation forecasts corrected with EDMD-m as a function of the lead time for all catchments and ~~all~~ for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and the CRPS (bottom) and the reference is Ensemble Streamflow Prediction. Each column corresponds to a target season. Each line represents the skill score in a catchment for forecast horizons within the target season.

and sharpness when streamflow climatology is used as reference~~to calculate the skill of EDMD-m bias corrected forecasts. As expected, streamflow~~ . Streamflow forecasts generated from bias corrected precipitation forecasts are sharper and present better overall performance than streamflow climatology, even for lead times of up to twelve weeks in some catchments. This was expected because ensembles based on hydrological modelling benefit from knowledge of initial hydrologic conditions. In one catchment (catchment 1), skill scores are systematically higher than the scores of the other catchments. In this catchment, streamflow climatology is very wide, with interannual variability of the same order of magnitude as interseasonal variability.

The PIT diagrams in Fig. ~~14~~ 13 show that the reliability of streamflow forecasts is also improved after bias correcting precipitation forecasts. In winter (DJF) and spring (not shown), streamflow forecasts are now reliable and equivalent to ESP, although forecasts still show a slight tendency to overpredict streamflows. In autumn (not shown), streamflow forecasts are also reliable in most catchments, but with a tendency to underpredict streamflows. Summer (JJA) streamflow forecasts are also more reliable ~~than they were prior to~~ after bias correction, but they still depict poor reliability and show that there is room for improvements. As shown by other studies in ensemble forecasting (Zalachori et al., 2012; Verkade et al., 2013; Roulin and Vannitsem, 2015), a simple bias correction of meteorological inputs is obviously not enough to achieve streamflow forecast reliability. In our case, the difficulties of the hydrological model in reaching lower streamflow values remain. This highlights the need for taking into account other sources of hydrological modelling uncertainties and including additional post-processing, targeting directly streamflow forecasts.

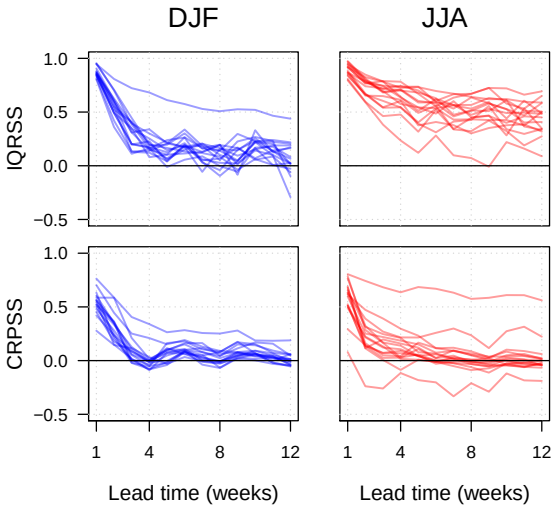


Figure 12. Skill of EDMD-m debiased streamflow forecasts as a function of the lead time for all catchments and ~~all~~ for the winter (DJF) and summer (JJA) seasons. The skill is computed based on the IQR (top) and based on the CRPS (right) and the reference is historical streamflow. Each column corresponds to the target season of forecast lead times. Each plotted line represents the performance of a catchment.

6.3 How improvements in precipitation forecasts propagate to streamflow forecasts~~?~~

We have seen that the use of reliable precipitation forecasts as input to a hydrological model does not automatically generate reliable streamflow forecasts. In order to further understand how improvements in precipitation forecasts propagate to streamflow forecasts, we compared the skill scores of EDMD-m bias corrected precipitation forecasts with the skill scores of the streamflow forecasts generated from these bias corrected precipitations. We focused the analysis on the four catchments previously selected as representative of the database, i.e. catchments 2, 4, 7 and 14.

Figure ~~15 presents the results for the~~ 14 presents the CRPSS, IQRSS and the PITSS (PIT area) in these four catchments, when raw forecasts are used as reference. The reference forecast for the computation of the skill scores of the bias corrected forecasts is the raw forecast. The skill thus represents a measure of the improvement due to bias correction. Skill scores were averaged over lead times of 10 ~~days~~ to 90 days.

In overall performance (CRPSS), bias correcting precipitation forecasts either led to a gain in skill in both precipitation and streamflow forecasts, as in catchments 4 and 7 and in some seasons in catchment 2, or to a skill equivalent to the skill prior to bias correction, as in catchment 14. Since catchments 4 and 7 were the ones with the most biased forecasts (cf. Fig. 6), there was more room for improvement in these catchments. Catchment 14 had the smallest bias of the four catchments. Bias correction had thus little impact on precipitation forecasts, and therefore also on streamflow forecasts. Interestingly, the improvement achieved in streamflow is always superior to the improvement achieved in precipitation, or equivalent when there was no gain in skill. It seems therefore that a small improvement in the overall performance of precipitation inputs (as measured by the CRPS) can translate in a greater improvement in streamflow forecasts.

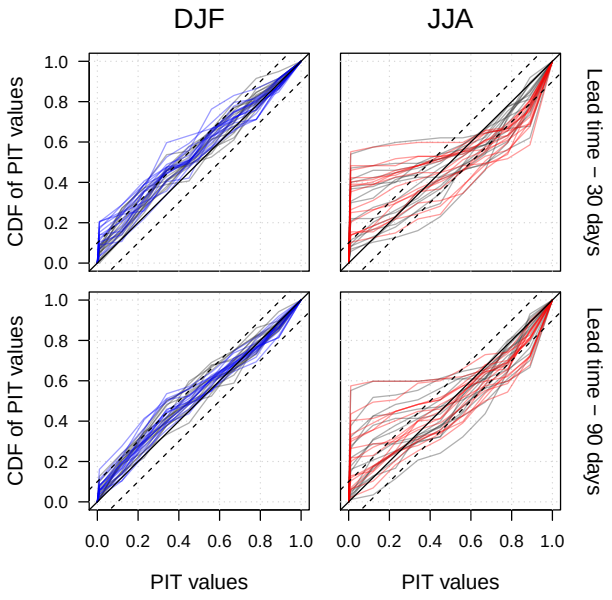


Figure 13. PIT diagram of streamflow forecasts obtained from precipitation forecasts bias corrected with EDMD-m (coloured lines) and Ensemble Streamflow Prediction (grey lines) for lead times of 30 days (top) and 90 days (bottom). Each column corresponds to a target season. Each line represents the PIT diagram in a catchment for forecast horizons within the target season. Dotted lines represent the Kolmogorov significance bands for a 5% significance test.

If we look at the skill in sharpness (IQRSS) and in reliability (PITSS)~~of the ensemble forecasts~~, we observe different behaviours. In sharpness, a loss in skill was observed in catchments 2 and 14, while a gain was observed in catchments 4 and 7. When a gain was achieved, the gain is superior in streamflow forecasts than in precipitation forecasts. ~~If we look at~~ In reliability, skill was always improved by bias correcting the precipitation forecasts, with skill scores always superior to 0.3. The gain in streamflow is mainly positive, but not always, as in the case of precipitation forecasts. Although the majority of skill scores are superior to 0.1, some values are below ~~the zeroskill score line~~zero. The gain in reliability from the application of bias correction to ~~raw~~ precipitation forecasts is, in general, superior in precipitation forecasts than ~~it is~~ in streamflow forecasts.

Based on our results, we can say that in catchments with small biases, here represented by catchments 2 and 14, overall performance was mainly stable from precipitation to streamflow forecasts. However, in these catchments, a gain in reliability was generally associated with a loss in sharpness. In catchments with greater biases, here represented by catchments 4 and 7, overall performance, sharpness and reliability were improved for both precipitation and streamflow forecasts by simply bias correcting the precipitation forecasts.

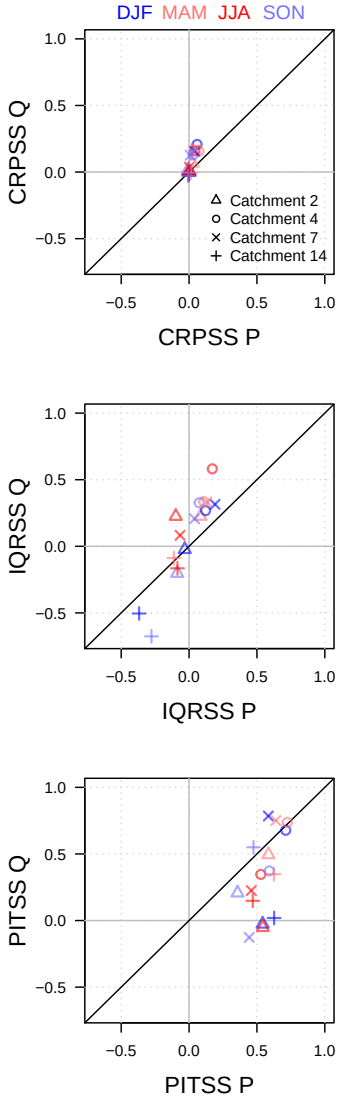


Figure 14. Skill scores of streamflow forecasts after correction with EDMD-m against skill scores of precipitation forecasts after correction with EDMD-m. The skill score of forecasts corrected with EDMD-m is computed ~~in regards to~~ using raw forecasts as reference. It is then averaged over lead times 10 to 90 days to obtain a single value. Results are shown for all four seasons in four selected catchments (Catchments 2, 4, 7 and 14). Skill scores were obtained based on the CRPS (top), the IQR (middle) and the PIT diagram area (bottom). The 1:1 diagonal corresponds to an equivalent performance increase in precipitation and streamflow.

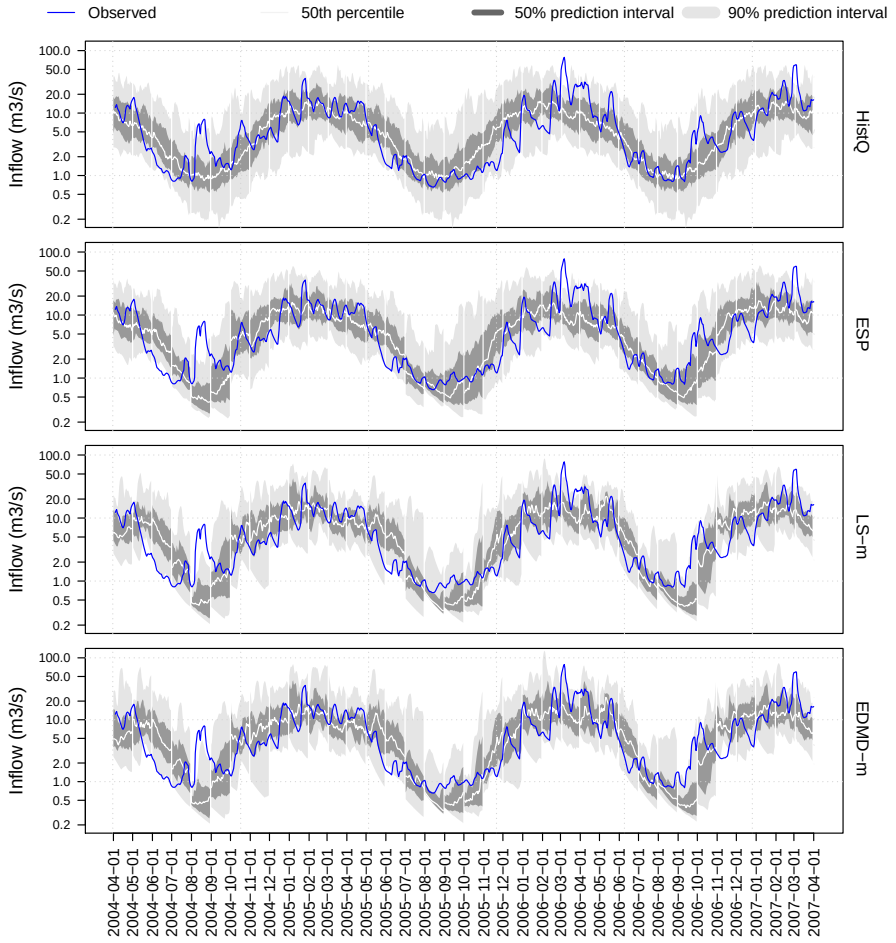


Figure 15. Hydrographs obtained with historical streamflow, ESP, seasonal forecasts corrected with LS-m and seasonal forecasts corrected with EDMD-m in catchment 7 from 1 April 2004 until 1 April 2007. The vertical axis is logarithmic. The blue line represents the observed streamflow. The grey shaded areas present the forecasts issued in the previous month, i.e. 31 to 60 days prior to the observations.

6.4 Example of forecast hydrographs in a selected catchment

Figure ~~16~~ 15 presents the hydrographs of the forecasts obtained from historical streamflow (HistQ), ESP, and seasonal forecasts bias corrected with LS-m and EDMD-m, from April 2004 to April 2007 in catchment 7. We show forecasts for lead times from 31 days to 60 days~~, i. e., forecasts issued in the previous month.~~ . Ensemble forecasts are represented by the median forecasts and two prediction intervals: the ~~25%-75% interval containing~~ 50% ~~of the ensemble members (~~interval (between the 25th and 75th percentiles; dark grey zone), and the ~~5%-95% interval with~~ 90% ~~of the ensemble members (~~interval (between the 5th and 95th percentiles; light grey zone). Observed streamflow is also shown. In this catchment, seasonal forecasts had a strong bias and bias correction methods performed well.

The hydrograph for historical streamflow (HistQ plot) represents the interannual variability in ~~streamflow in~~ the catchment, except that the forecast year is excluded for cross-validation. ~~It relies on past observations of streamflow and does not include seasonal meteorological forecasts. We can see that~~ Visually, the observations fall within the forecast ranges in most cases. The actual coverage of the 90% and 50% prediction intervals is 97% and 66% respectively, which indicates ~~, as expected with climatology, good forecast reliability. However, the forecast lacks sharpnessduring low-flow periods~~forecast overdispersion and poor sharpness. Accuracy of the median forecast (50^{th} percentile) is, in general, good ~~, although~~ with a mean absolute error (MAE) of 3.8 m^3/s, although, visually, we observe that too high and low peak flows are not well reproduced.

The forecasts obtained with the ESP method use past observations of precipitation as input to the hydrological model rather than seasonal meteorological forecasts. They show visible improvements in sharpness~~during low-flow periods, while reliability seems preserved~~, notably during low-flow periods. The 90% and 50% prediction intervals actually cover 92% and 60% of the observations, respectively. Accuracy of the median forecasts seems equal or lower than observed with ~~historical streamflow~~HistQ, which is consistent with an MAE of 4.1 m^3/s.

The hydrographs representing the streamflow forecasts obtained from bias corrected System 4 precipitation ~~seasonal~~ forecasts show forecasts that are sometimes even sharper than ESP forecasts, as seen, for instance, for the rising limb in 2005. ~~Overall, the observed streamflow falls within the forecast ranges.~~ In some situations, as in the peak event in August 2004, prediction intervals of bias corrected ~~seasonal~~ forecasts, particularly in the EDMD-m case, are closer to observations than ESP forecasts. In general, visual differences in quality between seasonal streamflow forecasts obtained from precipitation forecasts corrected with LS-m and EDMD-m are hardly noticeable. ~~Although EDMD-m forecasts seem to present slightly larger prediction intervals, which could result in better reliability but lower sharpness comparatively to LS-m,~~ For instance, the accuracy of their median forecasts is ~~practically identical . The visual inspection of these graphs for all catchments indicates similar results. Although our analyses and evaluation criteria have indicated the~~ identical with an MAE of 4.3 m^3/s. However, the 90% and 50% prediction intervals of EDMD-m ~~as the preferred method for the studied catchments, LS-m also yields good improvements in precipitation and streamflow forecasts. Since this method is easier to implement, it can be an alternative to the application of EDMD-m in operational forecasting systems~~forecasts actually cover, respectively, 89% and 51% of the observations, which indicates better performance in terms of reliability comparatively to LS-m, for which the actual coverage of these prediction intervals is 85% and 46%, respectively.

7 Conclusions

We assessed the quality of ECMWF System 4 precipitation forecasts for seasonal streamflow forecasting in 16 catchments in France. We evaluated areal precipitation forecasts over the catchments and streamflow forecasts generated from inputting precipitation forecasts to a lumped hydrological model. Results show that, in most catchments, raw (uncorrected) System 4 precipitation forecasts are sharper than precipitation climatology (i.e., ensemble forecasts built from past observed precipitations) in all seasons. However, raw precipitation forecasts show poor reliability and a tendency to overpredict precipitations. Likewise, streamflow forecasts generated from raw System 4 precipitations are sharper, but far less reliable than forecasts based

on the ESP approach (i.e., ensemble forecasts obtained from running the hydrological model with current initial conditions and past observed precipitations). Yet, in overall performance, raw precipitation forecasts yield improvements up to two weeks in all catchments over precipitation climatology, and streamflow forecasts yield improvements up to three to four weeks over ESP in some catchments. In general, improving forecast reliability, while maintaining (or not diminishing too much) forecast sharpness, was clearly a challenge for bias correction methods.

An in-depth analysis of the biases of System 4 seasonal precipitation forecasts showed strong monthly biases sometimes hidden at the scale of the year, depending on the catchment. Bias correction methods calibrated over the whole year were therefore less efficient when evaluating forecasts over calendar months. In the majority of catchments, the empirical distribution mapping of daily values (EDMD) or the simple linear scaling method (LS) applied to raw ~~System 4~~ precipitation forecasts showed more effectiveness in correcting the yearly but also the monthly biases. These methods also gave the highest increase in overall performance for streamflow forecasting. Empirical distribution mapping of daily values calibrated for each calendar month (EDMD-m) was particularly efficient to increase reliability of precipitation and streamflow forecasts, while linear scaling (LS-m) led to higher improvements in sharpness and accuracy.

The EDMD-m bias correction method was further investigated to better understand its impact on the skill of bias corrected seasonal forecasts~~in the studied catchments~~. Overall, the application of bias correction reduced the differences in forecast performance between seasons and catchments for precipitation and streamflow forecasts. Also, bias correction ensured that precipitation and streamflow forecasts were at least equivalent in performance to the historical precipitations and ~~streamflow forecasts based on historical precipitations~~the ESP method, respectively, up to three months ahead. In catchments with greater biases, overall performance, sharpness and reliability were improved for both precipitation and streamflow forecasts by simply bias correcting the precipitation forecasts. Overall performance was mainly stable in catchments with small biases. However, in these catchments, a gain in reliability was generally associated with a loss in sharpness. The evaluation of forecasts after bias correction, for the purposes of operational applications on water and risk management, may therefore involve a trade-off between sharpness and reliability. Furthermore, while precipitation forecast reliability is improved with bias correction, the evaluation of streamflow forecast reliability shows that there is still room for improvement. Notably, bias correction of precipitation inputs was not enough to achieve good reliability in summer streamflow forecasts. This highlighted the need for adding a step of streamflow post-processing to the forecasting system.

This study compared eight simple bias correction methods to correct precipitation seasonal forecasts and investigated how ~~one of them impacts~~ they impact the skill of streamflow forecasts. The catchments studied were not influenced by snowmelt flows and thus only precipitation was considered in the bias correction procedures. In other contexts, it may be interesting to also include bias correction of temperature forecasts, with appropriate methods to consider space-time interdependencies of the meteorological variables. The explicit consideration of temperature forecasts could also benefit the skill of low flow forecasts in summer, when evapotranspiration can play a crucial role.

Several other approaches for post-processing and bias correction exist, for instance, based on MOS techniques, space-time disaggregation schemes or Bayesian Model Averaging (Gneiting et al., 2005; Raftery et al., 2005; Liu et al., 2013; Hemri et al.,

2014). These could be investigated to contribute to the comprehensive comparison of options for bias correcting precipitation and temperature forecasts prior to seasonal streamflow forecasting.

~~Last~~Lastly, other forecasting methods selecting historical precipitations based on climate indicators have been investigated in the literature for seasonal hydrological forecasting in regions where strong correlations have been observed, e.g. in the United States or in Australia (Hamlet and Lettenmaier, 1999; Werner et al., 2004; van Dijk et al., 2013). In France, weak correlations have often shown that climate indicators may not be adapted to forecast precipitations at the seasonal scale. However, the use of indicators derived from seasonal forecasts could potentially improve the selection of past precipitation scenarios, which might enhance the skill of ESP methods to forecast streamflow.

Acknowledgements. This work was partly funded by the Interreg IVB NWE programme of the European Union, project DROP (Benefit of governance in DROught adaptation). The first author acknowledges Dr. Christopher A. T. Ferro for his insights on probabilistic scores.

References

Arlot, S. and Celisse, A.: A survey of cross-validation procedures for model selection, Statist. Surv., pp. 40–79, doi:10.1214/09-SS054, http://projecteuclid.org/euclid.ssu/1268143839, 2010.

Christensen, J. H., Boberg, F., Christensen, O. B., and Lucas-Picher, P.: On the need for bias correction of regional climate change projections of temperature and precipitation, Geophysical Research Letters, 35, L20 709, doi:10.1029/2008GL035694, http://dx.doi.org/10.1029/2008GL035694, 2008.

Crochemore, L., Ramos, M.-H., Pappenberger, F., van Andel, S. J., and Wood, A. W.: An experiment on risk-based decision-making in water management using monthly probabilistic forecasts, Bulletin of the American Meteorological Society, p. (in press), doi:10.1175/BAMS-D-14-00270.1, http://dx.doi.org/10.1175/BAMS-D-14-00270.1, 2016.

Day, G.: Extended Streamflow Forecasting Using NWSRFS, Journal of Water Resources Planning and Management, 111, 157–170, 1985.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, Hydrology and Earth System Sciences, 19, 275–291, doi:10.5194/hess-19-275-2015, http://www.hydrol-earth-syst-sci.net/19/275/2015/, 2015.

Di Giuseppe, F., Molteni, F., and Tompkins, A. M.: A rainfall calibration methodology for impacts modelling based on spatial mapping, Quarterly Journal of the Royal Meteorological Society, 139, 1389–1401, doi:10.1002/qj.2019, http://dx.doi.org/10.1002/qj.2019, 2013.

Dutra, E., Wetterhall, F., Di Giuseppe, F., Naumann, G., Barbosa, P., Vogt, J., Pozzi, W., and Pappenberger, F.: Global meteorological drought – Part 1: Probabilistic monitoring, Hydrology and Earth System Sciences, 18, 2657–2667, doi:10.5194/hess-18-2657-2014, http://www.hydrol-earth-syst-sci.net/18/2657/2014/, 2014.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions "Should we apply bias correction to global and regional climate model data?", Hydrology and Earth System Sciences, 16, 3391–3404, doi:10.5194/hess-16-3391-2012, http://www.hydrol-earth-syst-sci.net/16/3391/2012/, 2012.

Eslamian, S.: Handbook of Engineering Hydrology: Modeling, Climate Change, and Variability, Handbook of Engineering Hydrology, CRC Press, https://books.google.fr/books?id=-8LMBQAAQBAJ, 2015.

Faber, B. A. and Stedinger, J. R.: Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, Journal of Hydrology, 249, 113–133, doi:10.1016/S0022-1694(01)00419-X, 2001.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorological Applications, 15, 19–24, doi:10.1002/met.45, http://dx.doi.org/10.1002/met.45, 2008.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098–1118, doi:10.1175/MWR2904.1, http://dx.doi.org/10.1175/MWR2904.1, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 243–268, http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2007.00587.x/full, 2007.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations - a comparison of methods, Hydrology and Earth System Sciences, 16, 3383–3390, doi:10.5194/hess-16-3383-2012, http://www.hydrol-earth-syst-sci.net/16/3383/2012/, 2012.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, 2009.

Hamill, T. M. and Juras, J.: Measuring forecast skill: is it real skill or is it the varying climatology?, Quarterly Journal of the Royal Meteorological Society, 132, 2905–2923, doi:10.1256/qj.06.25, http://dx.doi.org/10.1256/qj.06.25, 2006.

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

Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R.: Confidence Builders: Evaluating Seasonal Climate Forecasts from User Perspectives, Bulletin of the American Meteorological Society, 83, 683–698, doi:10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2, 2002.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T.: Trends in the predictive performance of raw ensemble weather forecasts, Geophysical Research Letters, 41, 9197–9205, doi:10.1002/2014GL062472, http://dx.doi.org/10.1002/2014GL062472, 2014GL062472, 2014.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559–570, 2000.

Jenicek, M., Seibert, J., Zappa, M., Staudinger, M., and Jonas, T.: Importance of maximum snow accumulation for summer low flows in humid catchments, Hydrology and Earth System Sciences, 20, 859–874, http://www.hydrol-earth-syst-sci.net/20/859/2016/, 2016.

Kim, H.-M., Webster, P. J., and Curry, J. A.: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter, Climate Dynamics, 39, 2957–2973, doi:10.1007/s00382-012-1364-6, http://link.springer.com/10.1007/s00382-012-1364-6, 2012.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrology and Earth System Sciences, 11, 1267–1277, doi:10.5194/hess-11-1267-2007, http://www.hydrol-earth-syst-sci.net/11/1267/2007/, 2007.

Lemos, M., Finan, T., Fox, R., Nelson, D., and Tucker, J.: The Use of Seasonal Climate Forecasting in Policymaking: Lessons from Northeast Brazil, Climatic Change, 55, 479–507, doi:10.1023/A:1020785826029, 2002.

Liu, Y., Duan, Q., Zhao, L., Ye, A., Tao, Y., Miao, C., Mu, X., and Schaake, J. C.: Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai river basin, Hydrological Processes, 27, 57–74, doi:10.1002/hyp.9496, http://dx.doi.org/10.1002/hyp.9496, 2013.

Madadgar, S., Moradkhani, H., and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrological Processes, 28, 104–122, doi:10.1002/hyp.9562, http://dx.doi.org/10.1002/hyp.9562, 2014.

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., Magnusson, L., Mogensen, K., Palmer, T., and Vitart, F.: The new ECMWF seasonal forecast system (System 4), ECMWF Tech. Memo., 656, 49 pp., availableat:http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/601-700/tm656.pdf, 2011.

Muerth, M. J., Gauvin St-Denis, B., Ricard, S., Velázquez, J. A., Schmid, J., Minville, M., Caya, D., Chaumont, D., Ludwig, R., and Turcotte, R.: On the need for bias correction in regional climate scenarios to assess climate change impacts on river runoff, Hydrology and Earth System Sciences, 17, 1189–1204, doi:10.5194/hess-17-1189-2013, http://www.hydrol-earth-syst-sci.net/17/1189/2013/, 2013.

Musy, A., Hingray, B., and Picouet, C.: Hydrology: A Science for Engineers, CRC Press, https://books.google.fr/books?id=AHTSBQAAQBAJ, 2015.

Mwangi, E., Wetterhall, F., Dutra, E., Di Giuseppe, F., and Pappenberger, F.: Forecasting droughts in East Africa, Hydrology and Earth System Sciences, 18, 611–620, doi:10.5194/hess-18-611-2014, http://www.hydrol-earth-syst-sci.net/18/611/2014/, 2014.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J. M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrol. Earth Syst. Sci., 18, 2829–2857, doi:10.5194/hessd-10-13979-2013, 2014.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall-runoff model ? Part 2 — Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, Journal of Hydrology, 303, 290–306, 2005.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, Journal of Hydrology, 522, 697 – 713, doi:http://dx.doi.org/10.1016/j.jhydrol.2015.01.024, http://www.sciencedirect.com/science/article/pii/S0022169415000414, 2015.

Pushpalatha, R., Perrin, C., Mathevet, T., and Andreassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, Journal of Hydrology, 411, 66–76, doi:10.1016/j.jhydrol.2011.09.034, 2011.

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, Journal of Applied Meteorology and Climatology, 47, 92–107, doi:10.1175/2007JAMC1636.1, http://dx.doi.org/10.1175/2007JAMC1636.1, 2008.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Monthly Weather Review, 133, 1155–1174, doi:10.1175/MWR2906.1, http://dx.doi.org/10.1175/MWR2906.1, 2005.

Rayner, S., Lach, D., and Ingram, H.: Weather Forecasts are for Wimps: Why Water Resource Managers Do Not Use Climate Forecasts, Climatic Change, 69, 197–227, doi:10.1007/s10584-005-3148-z, 2005.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, W05 521, doi:10.1029/2009WR008328, http://doi.wiley.com/10.1029/2009WR008328, 2010.

Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrology and Earth System Sciences, 17, 579–593, doi:10.5194/hess-17-579-2013, http://www.hydrol-earth-syst-sci.net/17/579/2013/, 2013.

Roulin, E. and Vannitsem, S.: Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors, Hydrological Processes, 29, 1434–1449, doi:10.1002/hyp.10259, http://dx.doi.org/10.1002/hyp.10259, 2015.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrology and Earth System Sciences, 17, 2781–2796, doi:10.5194/hess-17-2781-2013, http://www.hydrol-earth-syst-sci.net/17/2781/2013/, 2013.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of Hydrology, 456-457, 12–29, doi:10.1016/j.jhydrol.2012.05.052, http://linkinghub.elsevier.com/retrieve/pii/S0022169412004556, 2012.

Teutschbein, C. and Seibert, J.: Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions?, Hydrology and Earth System Sciences, 17, 5061–5077, doi:10.5194/hess-17-5061-2013, http://www.hydrol-earth-syst-sci.net/17/5061/2013/, 2013.

Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E., and Uhlenbrook, S.: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa, Hydrology and Earth System Sciences, 19, 1695–1711, doi:10.5194/hess-19-1695-2015, http://www.hydrol-earth-syst-sci.net/19/1695/2015/, 2015.

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 Resources Research, 49, 2729–2746, doi:10.1002/wrcr.20251, http://dx.doi.org/10.1002/wrcr.20251, 2013.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, Journal of Hydrology, 501, 73–91, http://www.sciencedirect.com/science/article/pii/S0022169413005660, 2013.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, doi:10.1002/joc.2003, http://dx.doi.org/10.1002/joc.2003, 2010.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255–265, doi:10.5194/hess-15-255-2011, http://www.hydrol-earth-syst-sci.net/15/255/2011/, 2011.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, Journal of The Royal Society Interface, 11, 20131 162, doi:10.1098/rsif.2013.1162, http://rsif.royalsocietypublishing.org/content/11/96/20131162.abstract, 2014.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S.: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts., Journal of Hydrometeorology, 5, 1076–1090, http://dx.doi.org/10.1175/JHM-381.1, 2004.

Wetterhall, F., Winsemius, H. C., Dutra, E., Werner, M., and Pappenberger, E.: Seasonal predictions of agro-meteorological drought indicators for the Limpopo basin, Hydrology and Earth System Sciences, 19, 2577–2586, doi:10.5194/hess-19-2577-2015, http://www.hydrol-earth-syst-sci.net/19/2577/2015/, 2015.

Wilhite, D. A., Hayes, M. J., Knutson, C., and Smith, K. H.: Planning for drought: Moving from crisis to risk management, JAWRA Journal of the American Water Resources Association, 36, 697–710, doi:10.1111/j.1752-1688.2000.tb04299.x, http://dx.doi.org/10.1111/j.1752-1688.2000.tb04299.x, 2000.

Winsemius, H. C., Dutra, E., Engelbrecht, F. A., Archer Van Garderen, E., Wetterhall, F., Pappenberger, F., and Werner, M. G. F.: The potential value of seasonal forecasts in a changing climate in southern Africa, Hydrology and Earth System Sciences, 18, 1525–1538, doi:10.5194/hess-18-1525-2014, http://www.hydrol-earth-syst-sci.net/18/1525/2014/, 2014.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophysical Research Letters, 35, L14 401, doi:10.1029/2008GL034648, http://doi.wiley.com/10.1029/2008GL034648, 2008.

Wood, A. W. and Schaake, J. C.: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, Journal of Hydrometeorology, 9, 132–148, doi:10.1175/2007JHM862.1, http://dx.doi.org/10.1175/2007JHM862.1, 2008.

Wood, A. W., Kumar, A., and Lettenmaier, D. P.: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States, Journal of Geophysical Research: Atmospheres, 110, D04 105, doi:10.1029/2004JD004508, http://dx.doi.org/10.1029/2004JD004508, 2005.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resources Research, 49, 4687–4699, doi:10.1002/wrcr.20350, http://doi.wiley.com/10.1002/wrcr.20350, 2013.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, Wiley Interdisciplinary Reviews: Water, pp. 523–536, doi:10.1002/wat2.1088, http://dx.doi.org/10.1002/wat2.1088, 2015.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Advances in Science and Research, 8, 135–141, doi:10.5194/asr-8-135-2012, http://www.adv-sci-res.net/8/135/2012/, 2012.

Table 1. Number, name, surface, and mean annual precipitation, potential evapotranspiration and streamflow for the studied catchments.

#	River	Gauging station	Surface (km^2)	Mean annual precipitation (mm/yr)	Mean annual evapotranspiration (mm/yr)	Mean annual flow (mm/yr)
1	Andelle	Vascoeuil	377	952	628	332
2	Orne Saosnoise	Montbizot [Moulin Neuf Cidrerie]	501	735	696	163
3	Briance	Condat-sur-Vienne [Chambon Veyrinas]	605	1100	706	427
4	Ill	Didenheim	668	956	664	309
5	Azergues	Lozanne	798	931	689	296
6	Seiche	Bruz [Carcé]	809	732	696	181
7	Petite Creuse	Fresselines [Puy Rageaud]	853	899	680	316
8	Sèvre Nantaise	Tiffauges [la Moulinette]	872	898	712	331
9	Vire	Saint-Lô [Moulin des Rondelles]	882	958	629	448
10	Orge	Morsang-sur-Orge	934	658	680	131
11	Serein	Chablis	1119	842	675	220
12	Sauldres	Salbris [Valaudran]	1220	803	684	240
13	Eyre	Salle	1678	1025	785	323
14	Arroux	Etang-sur-Arroux [Pont du Tacot]	1792	981	655	390
15	Meuse	Saint-Mihiel	2543	948	639	372
16	Oise	Sempigny	4320	805	639	250

Table 2. Bias corrections applied: corresponding abbreviations, method used for calibration and description.

Abbreviation	Calibration based on	Description
LS-y LS-m	the whole year calendar months	Linear scaling of monthly values
EDM-y EDM-m	the whole year calendar months	Empirical distribution mapping of monthly values
GDM-y GDM-m	the whole year calendar months	Gamma distribution mapping of monthly values
EDMD-y EDMD-m	the whole year calendar months	Empirical distribution mapping of daily values