Response to Reviewer#1

The authors want to thank Dr. Zappa for his valuable review which will help to enhance the document significantly. We provide below our answers.

Reviewer 1

The authors are presenting a generally well-conceived study focussed on possible improvement of seasonal streamflow forecasts by applying bias correction to the forcing precipitation forecasts. I found the manuscript well written, with appealing and adequate artwork and well discussed results. I found that the whole manuscript well supports the results the authors are enumerating. The scores the chose are adequate and the combination of scores allows drawing conclusions that are not blended by the scores that show maximal improvement with respect to the defined references.
It is not clear which PET forcing is used in forecasts mode. If the authors use the observation-based reanalysis of PET in combination with the precipitation forecasts, then the appeal of this study would be quite reduced, in my "operational-minded" opinion.
Furthermore, the discussion and conclusions section is not adequately considering previous studies.

Authors' reply (AR): We thank the reviewer for his comments and we provide a detailed reply to the use of PET and to the final session of the manuscript below.

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): Thank you for pointing this out. We propose to change 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. We will clarify the way PET is considered and our reasons for doing so in the revised version.

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 following table summarizes some scores on the calibration and the validation of the GR6J model. Since other reviewers and the editor recommended decreasing the length of the paper, we propose to summarize this information in a sentence on line 25 in the revised version: "…applied to root-squared flows. We obtained an average KGE of 0.95 in calibration and 0.94 in validation over the sixteen catchments. The bias obtained in simulation ranges from -0.02 to 0.05."

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

RC: 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. We propose to include in the plots of Figure 16 a quantification of the performance of the systems over the period presented (April 2004 to April 2007). Notably, the coverage probability provides a good quantification to support the description. Here below, we summarize the values of MAE, coverage probability 90% (COV 5-95) and 50% (COV 25-75) obtained by each forecasting system over the displayed period. From these values, we observe that the ensembles based on past precipitations and past streamflow (HistQ and ESP) are more accurate over the chosen period (lower MAE values).

We also observe that the coverage probability of EDMD-m is the best over the chosen period. We propose to add these values to the plots of Figure 16 and to include a sentence on it in the interpretation of the hydrographs.

	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: We will revise this section. For instance, we think we can discuss our results in the light of those of Gudmundsson et al. (2012) and Teutschbein and Seibert (2012). However, we will not be able to add too much text, since the length of the paper was an issue for the other two reviewers and the editor.

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: See previous reply for PET. We will consider deleting this sentence or clarifying the limitations of the framework we adopted for real-time operations.

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, we will consider adding these two references in the revised version.

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

AR: We propose to add the following reference:
Arlot, Sylvain; Celisse, 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: We will add these references:

Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environmental Modelling & Software 25, 854 – 872. doi:http://dx.doi.org/10.1016/j.envsoft.2010.01.009

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G. and Mason, S. (2008), Forecast verification: current status and future directions. Met. Apps, 15: 3–18. doi: 10.1002/met.52

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practicioner's Guide in Atmospheric Science. John Wiley.

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 a second paper that is in preparation, and that should be submitted to this special issue (adding it here would result in a very long and unreadable paper). Past years were selected based on precipitation indices derived from seasonal forecasts. The resulting ensemble based on past-streamflow observations was compared with the ensemble of all past streamflow observations, the ESP and the streamflow forecasts obtained from precipitation forecasts and EDMD-m. We do not want to spoil the conclusions of this second paper here, so we invite the reviewer to check this special issue for our next submission.

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.

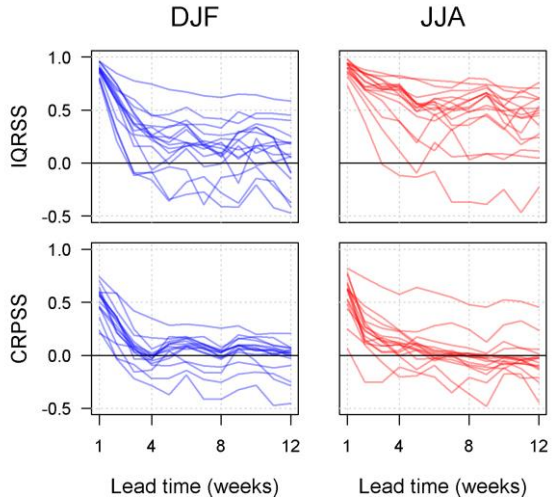
We did not try to use windows larger than 7 days. The objective of the rolling mean was to smooth the skill curves and remove the high frequency variations of the skill at the daily time step. Seven days appeared to be enough to smooth the curves, while keeping the moving mean as a good estimate of the 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. This remark is interesting and we would be glad to have more insights from other researchers on this too.

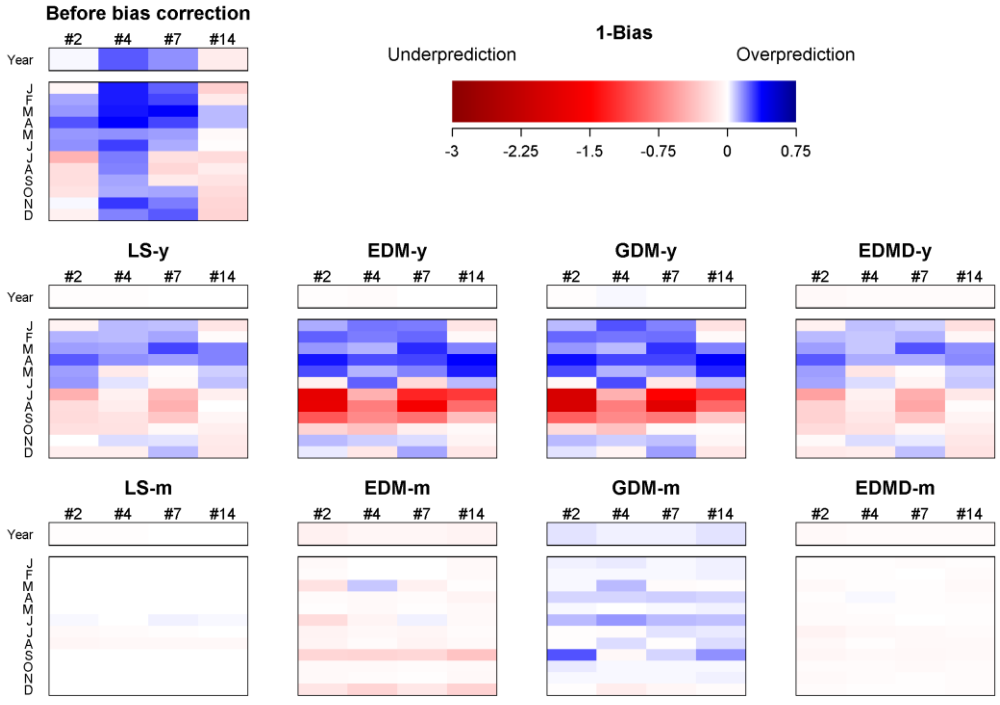
RC: 11 – Figure 4: How would look like this figure if you use the "ensemble based on past streamflow" as a reference?

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: The reviewer is right. Exchanging the blue and red colours in the scale, and increasing the right margin lead to the following figure. We propose to replace the original figure with this one:



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, the occurrences in 12 – 9, 13 – 2, 14 – 10, 14 – 27, 14 – 28 of "the 2-month" will be changed 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?

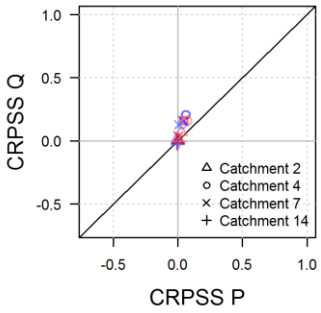
AR: Thank you. Your reading of the figure is absolutely correct. We will revise our description of the results to make sure it is clear to the readers.

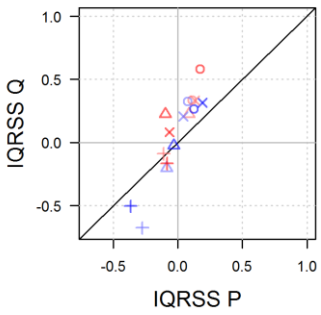
RC: 22 – 8: This would be the only heading with a question mark. Maybe replace this with a sentence

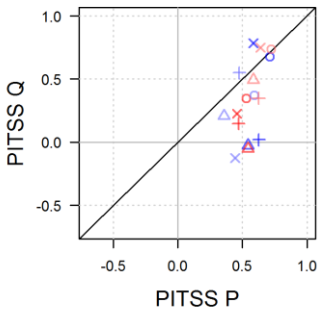
AR: The question mark was a typo and we will remove it.

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. Keeping the same scales gives the figure shown here below. With the same scales, we do not see clearly the impacts on CRPS, but we can better see the relative improvements between the different forecast attributes. For instance, we can see that the impact of bias correction is more seen in sharpness and reliability, rather than in overall performance. We will consider which could be the best figure to show when preparing the revised version of the paper.







Final considerations:
I find this manuscript is a very solid communication for the growing community dealing with seasonal forecasting in hydrology. It uses a strong set of data and robust statistics and comes to valuable conclusions. I indicated some weakness that let me recommend to the editors to ask for moderate revisions for this manuscript.
Best regards
Massimiliano Zappa
Birmensdorf, 23. March 2016

AR: Thank you again for your comments that greatly contribute to improving our paper.