# Response to Reviewer#2

The authors want to thank Reviewer#2 for the valuable comments, which will help us to enhance our paper. We provide below our answers to the comments.

### Reviewer 2

This study proposes an approach to improve short- and long-range (10-90 days) streamflow forecasts by conditioning resampled historical observations based on ECMWF System 4 forecasts. The conditioning is applied on both precipitation and streamflow records. Results are compared with historical resampled streamflow and ensemble streamflow prediction (ESP) as reference forecasts. Overall, the paper is well written and provides good assessments of different model performances. Nevertheless, I am concerned with the proposed method to improve streamflow forecasts (selection of resampled data based on GCM forecasts) as well as the results (week performance of the proposed method). Therefore, I think the paper is not ready for publication and requires major revision.

Authors' reply (AR): We thank the reviewer for the evaluation. Our aim was also to demonstrate through an extensive analysis the limitations and assets of the different conditioning approaches, notably when looking at the main attributes of forecast quality that are often searched by developers and users of forecasting systems (i.e., overall performance as measured by the CRPS, reliability and sharpness). We think that our paper provides useful insights to how hydrological seasonal forecasts can benefit from conditioning information. Our study also shows that the analysis of the usefulness of a forecasting system should not be restricted to evaluating some scores of forecast quality. It should also be extended to show how better forecasts impact the forecasting of the main variables of interest for a specific user and its decision-making context (in our paper, low-flow forecasting). In this regard, we think that, even if weak performance of seasonal forecasts is often observed in mid-latitudes (as is the case of our study catchments in France), progress can be obtained by reporting on experiments that focus on trying to understand where benefits can be expected. We think that the reviewers' comments received on this paper will greatly help us to improve our paper for its potential future publication.

---

**Reviewer's comment (RC):** Major comments:
1) The manuscript states that (P4, L9) the aim of this study is "to generate forecasts that benefit from the reliability of climatology-based ensembles and the sharpness of System 4 precipitation forecasts." First the proposed method does not seem to benefit from the sharpness of System 4, rather the reason for increased precision (sharpness) in the conditioned forecasts is due to the reduced ensemble size which is independent of the System 4's degree of uncertainty. Second, the results (e.g. Figures 4-5) show that except for some marginal improvements in forecasts for short lead times (Figure 4 upper row), the proposed method degrade the performance of the reference methods (CRPSS and PITSS are negative). In several instances in the manuscript (such as P9, L17) the authors discuss the improvements to the sharpness of the forecasts using their conditioning approach while reliability and performance have declined compared to the reference methods which undermines the sharpness improvements. The authors state that "...the PIT diagrams at 45 days show that this decrease does not affect the overall reliability of the conditioned ensembles" This again shows that the proposed method has not been able to improve upon the conventional approaches.

Authors' reply (AR): Following also the comments of Reviewer #1 (see also our replies to Reviewer #1), we understand that the aims of our study need to be clarified. Our general aim is stated on lines 6-7, Page 4: "(…) to investigate how selecting historical data based on forecast precipitation indices contributes to the skill of seasonal streamflow forecasts". The aim stated on line 9-10, Page 4 ("The aim is to generate forecasts that benefit from the reliability of climatology-based ensembles and the sharpness of System 4 precipitation forecasts") refers to the aim behind selecting the conditioning approaches to investigate how these can improve seasonal hydrological prediction. We agree that this is not clear as stated in the paper and we will clarify it in the revised version.

It is also interesting to note that one of the results illustrated in the paper is the discussions one can have around the importance of having forecasts of improved reliability and sharpness. When the reviewer states that "reliability and performance have declined compared to the reference methods which undermines the sharpness improvements", we believe that this a point of view and an interesting topic of discussion: Why degrading reliability undermines improvements in sharpness? Is it overall true or does it depend on the hydrological application? Can a user be so interested in improving sharpness that he accepts the cost of losing a bit of reliability? We do not mean that ensembles do not need to be reliable, on the contrary, we believe this is a quality that we should preserve when bringing improvements to a probabilistic or ensemble-based forecasting system. But sometimes a compromise between reliability and sharpness needs to be reached, and this is part of the results we show here (see also our other paper Crochemore et al., 2016, recently published). We illustrate how different approaches have different limitations, but also different assets. In our opinion, this is an important contribution, notably to better meet operational expectations.

---

**RC:** 2) The proposed method selects forecast ensemble members based on their closeness to some statistics (P8, L17). The procedure to choose the number of ensemble members to keep, however, is not explained. Is the number of selected runs subjectively chosen? If so a sensitivity analysis needs to be conducted.

AR: For a given forecast period, the conditioning statistic is calculated for each member of the System 4 forecast. We thus have an ensemble of forecast statistics of the same size as the System 4 ensemble for the forecast period. For each member of this ensemble of forecast statistics, the closest historical scenario is identified and used as ensemble member (i.e. as a local temporal realization for that forecast statistic). We will clarify this in Section 2.4.2.

---

**RC:** 3) The method conditions the resampled precipitation and streamflow data to GCM forecasts. However, GCM forecasts are uncertain particularly at seasonal scales. That might explain why the overall results do not show improvements compared with conventional ESP. In particular, authors need to discuss how the method will perform in regions with high topographical variations (considering that the low-resolution GCMs cannot capture the regional hydroclimatic variations). Related to this please discuss why you compare the proposed conditioning approach (based on SYS4) with results of SYS4?

AR: The idea behind this conditioning is that, even though GCM forecasts are uncertain at seasonal scales, coarse precipitation statistics (such as the SPI or monthly sums) may be easier to predict than precipitation time series. The performance of System 4 in predicting these coarse statistics is presented in Figures 2 and 3. Based on these results, we could expect the conditioning to improve sharpness.
The idea behind the comparison with Sys4 was to evaluate how the conditioned ensembles resemble the forecasts directly derived from System 4 time series in terms of reliability and sharpness. Another idea was to check the added value of conditioning compared to using Sys4 alone. We propose to add a sentence at the beginning of Section 3.2.2 or 3.2.3 to clarify why we make the comparison.

---

**RC:** 4) Please clarify which are the statistics (section 2.4.2) calculated for each ECMWF ensemble member separately or for the average of the 51 ensemble runs?

AR: The statistics were calculated for each member so as to obtain an ensemble of statistics (see also our reply above). We will clarify this in the revised version.

---

**RC:** 5) P8, L25: "when directly selecting scenarios from past streamflow observations, the last observed streamflow is added as a conditioning criterion in the computation of the Euclidian distance." This is problematic as the last observed (previous year's(?)) streamflow is not a good indicator of the next year's streamflow in particular with regard to high and low flows which are driven by several hydroclimatic factors that do not necessarily repeat at consecutive years.

AR: In fact, the hydrological model is run at the daily time step and "the last observed streamflow" refers to the observed streamflow on the day of issuing the forecast (Section 2.2). We will make sure that this is clear in the revised version.

---

**RC:** 6) Resampled precipitation is considered to drive the hydrologic model, however, the mean interannual potential evapotranspiration is used instead of the resampled one. Considering that PET might have a substantial role in low flow forecasts, I recommend using the resampled PET as well.

AR: We used the mean multi-annual PET instead of the resampled one when conditioning ESP in order to compare it with System 4 streamflow forecasts. Indeed, System 4 streamflow forecasts are also produced by forcing the model with the mean multi-annual potential evapotranspiration.
As a matter of fact, we had first produced the results in Figures 5, 6 and 7 for the resampled PET (PET for the years resampled based on precipitation). The results we obtained were very close to those presented here.

---

**RC:** 7) P12, L12: "The rankings are based on the visual evaluation of Figure 5." Visual evaluation is not an appropriate ranking approach.

AR: For a more quantitative ranking, we will consider ranking the methods by using averaged skill scores in the revised version.

---

**RC:** 8) Results of section 3.4 are based on only one drought event for one catchment and cannot provide sufficient evidence for the overall performance of the methods.

AR: We agree; the aim of Section 3.4 is purely illustrative. We do not aim at providing a statistical assessment of overall performance, especially as the illustration refers to rare events in hydrologic risk assessment. We will clarify this and pay attention not to drawn any general conclusions on the statistical performance of the systems from the analysis of the figure.

---

**RC:** 9) P6, section 2.3.1 Please elaborate further on the differences between CRPS and PIT and how they should be interpreted when they show inconsistent results (e.g. Fig 4).

AR: The CRPS is the sum of several terms, one representing reliability and one being influenced by sharpness (Hersbach, 2000). Therefore, the CRPS can be stable even though reliability is deteriorated, provided that sharpness, for instance, is improved. We will add a few words in Section 2.3.1 in the revised version to clarify this.

---

**RC:** 10) Multi-model averaging methods (such as simple mean, Bayesian Model Averaging (BMA) etc.) (Duan et al. 2005, Najafi et al. 2015, Raftery et al. 2005) have shown to improve short and long term hydrologic forecasts. I would suggest discussing the application of these approaches to merge the ensemble of forecasts obtained from different methods in this study.

AR: This can be an interesting topic for further studies. We will consider it in the discussion/perspectives presented at the end of the paper in the revised version.

---

**RC:** Specific comments:

- Abstract "…forecasts based on GCM outputs can offer sharper ensembles… :": does "sharper" refer to more precise? Related to this please define "sharpness" and "reliability" before using these terms, in the Introduction.

AR: Sharper refers to the range of possible future scenarios. It is a property of the ensembles and do not depend on the observations (as is the case of accuracy). We will add short definitions to the concepts of sharpness and reliability in the revised version.

---

**RC**: - L15: ECMWF System 4: Please expand the full name.

AR: We will expand the full name in the revised version.

---

**RC**: - Abstract: "The four conditioned precipitation scenarios were used as input to the GR6J hydrological model to obtain eight conditioned streamflow forecast scenarios": The statement is vague as to how four precipitation scenarios result in eight streamflow scenarios?

AR: Indeed, we will rewrite this sentence to clarify the methodology.

---

**RC**: - P2, L19: ESP is one of the streamflow forecast methods which need to be discussed here. Also please note that in ESP all historical meteorological forcings can be resampled to run the hydrological model (not just precipitation as stated in LP2, L27)

AR: We see ESP as a hybrid approach: it is statistical in terms of precipitations (climatology) and dynamical when it comes to streamflow (referring to the use of a hydrological model). Therefore, we propose to add a sentence after "More importantly, some studies have shown that the two approaches can complement and benefit from each …" to better introduce the ESP as a type of combination of the two approaches. We will also change the definition of ESP to refer to meteorological forcing to a hydrological model.

---

**RC**: - P4, L3 Statement is not clear "although the ensemble conditioned from historical streamflows, which was the…"

AR: We propose to change this to "They found that the GCM-conditioned ensemble outperformed the ESP method. Nevertheless, the ensemble conditioned from historical streamflows was the most reliable. In addition, decisions based on that ensemble completely eliminated flood damage and generated more energy than decisions based on the other two ensembles."

---

**RC**: - P4, L12-15: Please move to the results section.

AR: We agree and will consider moving the text in the revised version.

---

**RC**: - P4, L17: Please define "discrimination"

AR: The discrimination of a system is its capacity to detect an event defined by a threshold. We will add a definition in Section 2.3.1, when presenting the ROC score.

---

**RC**: - P5, L3: Please explain how many grid cells lie within each catchment in average. How was the aggregation performed? Please also indicate the forecast starting date.

AR: Each catchment is covered by one to four grid cells. The aggregation method is a simple weighted mean of precipitations from different grid cells, based on the area of the catchment covered by each cell. Forecasts are issued for the 1st of each month. We will clarify this in the revised version.

---

**RC:** - P5, L23: What do you mean by "systematically"?

AR: We meant that ESP_SPI3 is the only forecasting system that belongs to the best Friedman category for the three lead times (category c at two weeks, and category b at five and twelve weeks). We will remove "systematically" and state explicitly that we refer to "all studied lead times".

**RC:** - P5, L31-33: What is the range of KGE values? Please show the equations for KGE and 1-bias and include their ranges.

AR: We will add the range of KGE values. We will also explain the way the bias was computed. However, we would prefer to avoid adding the equations for these two criteria since they are only mentioned once and a reference article is already provided for the KGE.

**RC:** - P6, L9: Please change "The CRPS averages over the evaluation period the area between the cumulative forecast distribution…" to "The CRPS averages the area between the cumulative forecast distribution… over the evaluation period." Similarly, for L12.

AR: We will correct this.

**RC:** - P7, L3: What is the "reference"? Is it HisQ? Please define.

AR: We will add the information.

**RC:** - I suggest bringing section 2.4 before section 2.3.

AR: We will consider changing the position of these two sections in the revised version.

**RC:** - Figure 2: What is the difference between SUM1-3 and SUM3

AR: SUM3 is the sum of precipitations over the 3-month forecast horizon. SUM1-1 corresponds to the sum of precipitations over the first month of the forecast horizon, SUM1-2 the second month and so on. We will clarify this in the revised version.

**RC:** - P9, L1 "The reference forecast used to compute the skill scores is historical precipitations (i.e. climatology)": Do you mean hydrologic model simulation driven by historical precipitation?

AR: The reference here is historical precipitations. The analysis refers to precipitations only and not to hydrological model simulations. We evaluate precipitation indices derived from GCM-outputs and compare them to the precipitation indices derived from all historical years of precipitation. In other words, we compare the performance of the precipitation inputs used to obtain System 4 streamflow forecasts, to the performance of the precipitation inputs used to obtain ESP.

**RC:** - P9, L3 "SPI forecasts issued from System 4 are reliable overall and in standard precipitation conditions" please provide a reference

AR: This sentence is based on the analysis of Figure 3, which we will explicitly cite in the revised version.

## References

Crochemore, L., M.-H. Ramos, F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, 20: 3601-3618

Hersbach, Hans. “Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems.” *Weather and Forecasting* 15, no. 5 (2000): 559–70.