Response to Reviewer#1

The authors want to thank Reviewer#1 for the valuable comments on our manuscript. We provide below our answers to the comments.

Reviewer 1

As an outsider to the professional academic world, I feel that I cannot speak with unquestionable credibility to the novelty or scientific soundness of this manuscript – I am simply not familiar enough with the wealth of recent research into seasonal hydrologic forecasting. However, I can supply my overall impression of this work, which may be useful given my background in operational hydrologic forecasting.

Reviewer's comment (RC): The authors reference several studies that utilized approaches similar to the one undertaken here – conditioning historical observation-based ensembles to improve forecasts generated from these ensembles. Thus, the fundamental direction of the current study is not overly original. However, the manner in which the conditioning was applied – using GCM- and climatology-derived precipitation indices to select the most relevant historical ensembles – does appear to be a novel approach.

Authors' reply (AR): Our study is clearly not the only one in the topic. As we have mentioned on Page 3, lines 2-3, the conditioning or weighting of past observations based on climate signals is a recent topic of research, and several studies are emerging to investigate the best approaches to be used in order to extend the skill of seasonal hydrometeorological predictions. In this regard, our study aims to contribute to this research area and the approach and application that we propose in the paper is original. Studies that use GCM-derived precipitation indices as conditioning indices are rare and applications of conditioning approaches over mid-latitudes (in our case, France), where the reliability of seasonal weather predictions is, in general, low, are important contributions to the community. In addition, in this study:

- this specific conditioning method was applied simultaneously to a set of sixteen catchments, which, to our knowledge, had not been done previously,
- we distinguish between performance associated with sharpness and reliability,
- the latest GCM forecasts in date for Europe (System 4 from ECMWF) were used. This is of importance since seasonal forecasting in meteorological centres has been improving in the past decade, and to our knowledge, no other study tried this conditioning approach as extensively as in our study with these latest forecasts available.

We propose to add a sentence at the beginning of Section 1.3 (Scope of the study) to highlight these points and the originality of our paper in the light of the existing literature.

RC: The potential utility of this approach is presented well in Figure 2, where the precipitation indices generated from the GCM hindcasts (ECMWF Sys4) are compared against those generated from the historical observations. As the authors state, the Sys4 indices perform at least as well as the base indices overall (CRPSS), even outperform at one month lead time, but are consistently sharper (IQRSS). Further, the Sys4 indices have good reliability overall (Figure 3). The reliability of the indices falters when looking at only drier than normal or only wetter than normal conditions, but this seems to be unavoidable with any forecasting approach.
Despite the prefaced potential of using the Sys4 precipitation indices to condition, or subset, historical ensembles, this study's results offer just marginal practical insight:

1) Subsetting the ensembles based on the precipitation indices improve the HistQ performance more than it does the ESP performance. This result is not very useful, however, since the HistQ approach is rudimentary (and likely rarely used), and the primary benefit of the conditioning is seen during short lead times (which is simply the effect of blending from the last streamflow observation).

AR: The reviewer has clearly understood the aims of the analysis illustrated in Figure 2 and we acknowledge the positive comment provided. We believe this first step in analysing the performance of the precipitation-based conditioning indices is essential prior to analysing the conditioned outputs in terms of streamflow, given the non-linearity in the transformation of precipitation into streamflow in a hydrological model. We also included HistQ in our study because this is a "poorman's approach" that can serve as a naïve benchmark, where no hydrological model but only a long streamflow time series of records is available. One of the objectives of our study was to see whether this rudimentary approach could be turned into a valuable one provided that precipitation anomalies are available. While we agree that the improvement in the first days/weeks can be due to the assimilation of the last streamflow observation, the effect of this data assimilation technique usually decreases with lead time. The improvement observed in sharpness, for instance, is then mostly due to the conditioning. This improvement is one of the aims of several operational seasonal hydrological prediction systems: obtain sharper predictions, while maintaining reliability.

RC: 2) For ESP, SPI-conditioning appears to outperform SUM-conditioning, but this statement is qualitative at best and neither set of conditioned ensembles provides any notable improvements over the base ensembles. Compared to the base ESP ensembles, the sharpness of the ESP_SPI3 ensembles was improved by up to 10% but the reliability was degraded by up to 40% (Figure 7).

AR: The comparison between SPI-conditioning and SUM-conditioning over ESP is illustrated in Figure 5. From this figure, we can see that conditioning on SPI (third and fourth columns) provides better scores over (or at least do not degrade the score of) the reference (base ESP ensembles) than conditioning with SUM (first and second columns). For instance, based on the IQRSS results, we can see that the SUM-conditioning may decrease sharpness in some cases, whereas the SPI-conditioning guarantees to maintain or even increase sharpness. Based on the PIT diagram analysis, the SUM-conditioning causes an overprediction of observations, whereas the SPI-conditioning clearly limits this effect. Figure 7, mentioned by the reviewer, provides the means for a more quantitative analysis. However, it is restricted to the results for SPI-conditioning, since this one was already qualitatively better than the SUM-conditioning from the analysis in Figure 5. Figure 7 shows that we can lose in reliability (PIT area) in some catchments when comparing ESP-SPI to the base ESP ensembles (mainly at longer lead times), but that, in general, we gain in sharpness. The loss in reliability does not necessarily mean that the ensemble becomes "unreliable". As illustrated in Figure 5, ESP-SPI ensembles are still not far from the diagonal of perfect reliability of the PIT diagram. Here again this relates to one of our objectives: how can we obtain sharper predictions while still having reliable ensembles.

RC: 3) The conditioning improved the performance of HistQ ensembles in forecasting low flow events and variables, but the conditioned ensembles were still less skilful than the Sys4 and ESP/ESP_SPI3 ensembles.

AR: The reviewer is right. If we look at Figure 6, HistQ_SPI3 appears to be less skilful than Sys4, ESP and ESP_SPI3 in terms of overall performance. Nevertheless, this ensemble has characteristics of interest for low-flow forecasting. In Figure 9, this ensemble systematically is in the best category for deficit duration, probably because it represents recessions better than the model does. This is quite an advantage of this ensemble. Again, we were interested in studying this ensemble HistQ because it can be a benchmark and a simple approach if a hydrological model is not used. We believe that investigating the possible ways of improving the HistQ approach is useful and, notably, provides insights to how the model (and its performance) influences (for better or worse) the quality of streamflow predictions.

RC: 4) The authors state that the ESP_SPI3 approach "systematically appears to be one of the best options to forecast deficit volumes." However, this conclusion is very subjective, as it is not authoritatively substantiated by the results presented in Figures 9 and 10.

AR: Figure 9 shows the results for the deficit duration, and not for the deficit in volumes. In Figure 10, which represents the reliability of the forecasting systems in terms of deficit volumes, ESP_SPI3 is the only forecasting system that belongs to the best Friedman category (non-parametric statistical test of significance) for the three lead times shown (category c at two weeks, and category b at five and twelve weeks) (these results are representative of the other lead times that are not shown). This is why we stated that it is the best method. We propose to replace "systematically" by "for all three lead times" to make the sentence clearer.

RC: Although several pages of this manuscript are spent discussing the results in great detail, and the authors walk through the discussion in a relatively clean, scientific manner, much of this discussion is centered around tangential topics. For example, the comparisons between the conditioned ESP/HistQ ensembles to the Sys4 ensembles seem irrelevant given that the conditioning did little to improve, and actually degraded in some cases, the skill compared to the base ensembles. Thus, comparing the conditioned ensembles to the Sys4 ensembles is equivalent to comparing the base ensembles to Sys4, which of course is unnecessary. The results should be restricted to and presented with the stated goal of the study in mind – improving the skill of historical observation-based ensemble forecasting systems.

AR: We will consider shortening the discussion to better highlight the key results of the paper.

RC: Unfortunately, because there is little to report on the utility of applying this conditioning method to seasonal streamflow and low flow forecasting, the authors may need to redesign and/or include other experiments before resubmitting this paper. One suggestion, actually offered by the authors, is to examine the utility of using SPEI to condition the ensembles. Although the SPI is likely sufficient to appropriately subset historical precipitation ensembles, it may not be sufficient from a streamflow perspective. It seems likely that the relative magnitude of an individual SPI value may not always be translated into a similar relative magnitude flow or volume value if ET is a major hydrologic control in the watershed of interest (i.e. late season streamflows can be very different following extended dry but mild vs extended dry but hot conditions). Thus, conditioning the ensembles with both precipitation- and temperature-driven indices may provide more robust results.

AR: We think that the systematic analysis that we have carried out and reported in the paper is of relevance for the community, mainly since several efforts have been recently put into improving meteorological seasonal forecasting systems to better quantify risks and impacts in hydrology. In this regard, we think that our paper provides useful insights to how hydrological seasonal forecasts can benefit from this information (when used directly as input to a hydrological model or when providing conditioning indices to select ensemble traces). It must be noted that the work carried out in this paper is based on previously bias corrected System 4 precipitation forecasts, as stated on line 7, Page 5 (see also our paper Crochemore et al., 2016). We thus investigate the performance of the conditioning considering an improved ensemble precipitation prediction system. 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. We also think that our investigation on the utility of applying a conditioning method is useful to the community. We have evaluated the conditioning based on several major qualities of ensemble forecasts: overall performance, sharpness and reliability. We have demonstrated how these respond differently to the conditioning approaches. Again, it seems important to us that a developer or a user of seasonal forecasting systems be aware of the different impacts on forecast quality. We illustrate the performance of our ensembles with a low-flow forecasting case. We think that this illustration can also be useful to other users, with other preferences or operational focusses. Mainly, our study 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.

We agree with the reviewer that many other additional experiments could be done. For instance, the analysis of conditioning approaches based also on temperatures would be very interesting, although the impact of the potential evapotranspiration (used as input to the hydrological model) is expected to be of a second-order (as measures on the streamflow forecasts), comparatively to the impact of the rainfall conditioning, given that rainfall-runoff models are more sensitive to errors in rainfall than in potential evapotranspiration. The use of other indices could also be interesting for further studies. We could not add more experiments to our paper or it would become too long and lose its focus. We mentioned however numerous perspectives for further studies on Page 16 (line 10 onwards).

RC: Lastly, the underlying standard of this manuscript is the stated inherent reliability of historical observation-based ensembles, but this is a bit misleading. In true forecasting (not hindcasting), climatology-driven predictions may not be all that reliable. Several decades worth of historical information is often sought to build an ensemble forecasting system, but the climatic regime of the forecast area may be changing too rapidly for this. Thus, the distribution functions of actual forecasts and their corresponding observations may be offset from one another (i.e. not fall on a 1:1 line). Perhaps the authors should frame the goal more along the lines of using the conditioning to sharpen the ensembles, and less along the lines of marrying the reliability of historical ensembles with the sharpness of GCMs.

AR: We thank the reviewer for this interesting comment. We focused on the search for sharper ensembles while maintaining reliability, since this is a widespread notion in forecast verification. However, we also agree that the main message to convey is on using the conditioning to sharpen the ensembles (without deteriorating reliability). We will have this comment in mind when producing the revised version of the paper.

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