Response to Editor Decision:

Publish subject to technical corrections (11 Dec 2017) by Maria-Helena Ramos

I read carefully your answers to three reviewers and your revised version of the paper. I agree with the reviewers that the paper is worth publishing and I consider that you have addressed well the remarks and suggestions from the reviewers, which greatly improved your manuscript.

Thank you, we agree the reviewer's comments have greatly improved the manuscript.

I have just a few remarks that I think would be interesting to be considered for the final paper (I consider them just technical remarks with no need for a new round of revision, but I would appreciate if you could consider addressing them). They are listed below.

Remarks:

1) I agree with the first remark of reviewer #3 about using a single-value forecast to post-process streamflow predictions. Technically, ensemble dressing with model errors can consider all members of an ensemble without the need for reducing the ensemble to a mean or median value. I can understand your choice not to use the full ensemble members, but I have more difficult to find your arguments convincing. I think the sentences on lines 4-8, Page 10, need some rethinking. I do not think that the applications you cited for aggregating ensemble predictions are used in the correct context. The study of Matte et al. (2017), for instance, focused on a utility model, not a post-processing development approach. Besides, the ESP approach is common in operational seasonal forecasting for water management and it is based on ensembles of historic precipitation used as input to a hydrological model. I would consider writing this sentence differently: "Although the use of aggregation approaches for single-valued streamflow forecast from ensemble predictions has been seen in operational applications (e.g., Lerat et al, 2015…), we note that this approach may result in some information loss."

Then, the sentence on lines 7-8 also raises a question: do you mean further work is needed within your study or in general? I think you could have considered the full ensemble with further work within your study, but I also think that we can find ensemble post-processors in the literature, so it is not as if there is nothing done on the topic. Post-processors that use the full information from ensembles do exist. I suggest that you either delete or you explain better what you mean with this sentence.

These are suggestions to clarify your opinion on the topic as expressed in your paper, so the reader can better apprehend your viewpoints.

Thankyou for the suggestions to clarify this point. The intention was to reflect further work in general could be undertaken to provide clarity on the best approaches to implement ensemble post-processors. However, as pointed out, studies that do use the full information from ensembles do exist. As such, this sentence suggesting further work has been deleted.

Lines 4-8 page 10 have been revised to the constructive suggestion made. Only the Lerat et al, (2015) reference has been retained. Along with Matte et al. (2017) pointed out above, the two other citations used a single streamflow forecast to calibrate and apply a residual error model for streamflow forecasts, but there was not a preceding step of aggregation of an ensemble of streamflow forecasts. The sentence has been changed as follows:

Although the use of aggregation approaches for single-valued streamflow forecast from ensemble predictions has been seen in operational applications (see, for example, Lerat et al, 2015), we note that this approach may result in some information loss.

2. I also have an issue with the calculation of what you call "sharpness" and I think it comes back to a remark of reviewer #1 on using "sharpness versus precision". As indicated in Eq. 10, it is not the same as calculated in forecast verification studies, as presented in the reference books: Jolliffe, I.T., and D.B. Stephenson, 2012: Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd Edition. Wiley and Sons Ltd, 274 pp. , and Wilks, D.S., 2011: Statistical Methods in the Atmospheric Sciences. 3rd Edition. Elsevier, 676 pp.

Sharpness is supposed to be an attribute that measures the degree of variability of the forecasts or concentration of the predictive distributions. In forecast verification, sharpness is a property of the forecast only. In Eq. 10, if I understand well, you calculate a type of coefficient of variation with the standard deviation of the time series of (median?) predictions and the mean of the observations. It could be understood, I guess, as a normalized "standard deviation", but I think it is misleading to call it "sharpness", and I suggest you to call lit otherwise to avoid misunderstanding from the forecast verification community.

It is agreed that this difference should be highlighted to avoid misunderstanding. The measure of the degree of variability of the predictive distributions is considered to be captured by the standard deviation of the predictive distribution each time step, and then aggregated across time steps. We choose to normalise by the sum of the observed streamflow ($\sum_{t=1}^{N} \widetilde{Q_t}$), as this is consistent between model configurations and rainfall forcings. If this normalisation is not included there can be different baselines for the different model configurations or forcings, which can introduce biases in this metric (see McInerney et al., 2017). The name of the metric has been changed to highlight this normalisation step and avoid confusion, Sharpness$_{Norm}$. The text has been changed as follows:

***Sharpness** refers to the width of the predictive distribution, and can otherwise be known as "resolution" or "precision". Typically, sharpness is a determined using the predicted values only. In this work a measure of sharpness (as the sum of the standard deviation of the predictions each time step), is normalised by the sum of the observed values, to enable a comparison of this metric across catchments with different magnitudes of flow. As such, sharpness is quantified using the following metric from McInerney et al. (2017):*

$$Sharpness_{Norm} = \sum_{t=1}^{N} sdev(\boldsymbol{Q}_t\ ) / \sum_{t=1}^{N} \widetilde{Q_t} \qquad (10)$$

3) Minot issues are:

- Page 1, line 13 (abstract): consider changing to "… compared to a longer calibration period…"

Corrected, "to" included.

- Page 2, line 12: you explain in your reply to the reviewers that you prefer to use "forecast" instead of "forecasted". Maybe you should do the same here for the sake of consistency.

Corrected, thank you for picking this up.

- Page 2, line 15: I would consider changing to "… user need for monthly to seasonal streamflow forecasts…". In fact, the use of "seasonal" alone in the paper is appealing, but the study is dealing with "monthly" forecasts and not seasonal forecasts. I think you should not mix both and should keep to using "monthly" when referring to your study and application. For instance, I suggest to consider this remark in these other parts of the paper: Page 4, line 12; Page 4, line 22; Page 17, line 8.

As pointed out, originally it was considered more streamlined to use “seasonal” in a more generic sense, referring to forecasts longer than a short term forecast (e.g. 7 days). However, it is agreed that this could be ambiguous, particularly as a season typically refers to multiple months. As such, the suggested change has been adopted, and “monthly” is used instead of “seasonal” when referring to the application.

- Page 2, line 20: consider changing to “… calibration; Mount et al., 2016; …)”

Corrected, extra brackets removed.

- Page 2, line 24: consider changing to “…(Wu et al; 2013); however this is not always the case (for example, Brigode et al., 2013)”

Corrected.

- Page 3, line 4: consider changing to “…etc.); see, for example, Westra et al., 2014).”

Corrected.

- Page 3, lines 10-11: consider changing to “…calibration period, which exposes….of a shorter calibration period, which exposes the model…”

Corrected.

- Page 3, line 17: consider changing to “…and Maranzo, 2004) and disaggregation approaches…”

Corrected.

- Page 4, line 4: “… forecasting applications in France.”

Corrected.

- Page 4, line 17: consider changing to “... of this study are to:”

Corrected.

- Page 5, line 3: consider changing to “…(Figure 1), which conveys…”

Corrected.

- Page 8, line 5: consider replacing “big” to “important”

Corrected.

- Page 8, line 28: I would write “10-year calibration period” or “one-year warm-up period”, etc.. Consider checking if that would be correct in English and change over all the uses in the paper.

The paper has been updated to include “-year” where appropriate, following a number.

- Page 11, line 13: with only one reference, I suggest deleting this part of the sentence “and are common in forecasting applications” and only mention “… monthly time scale (see another example in Lerat et al., 2015)”.

Changed to:

These choices of residual error models at the daily and monthly time scales contribute to the study objectives of reliable forecasts at the monthly time scale (see another example in Lerat et al., 2015).

- Page 11, line 19: consider adding a comma: “For all scenarios, observed…”

Comma added.

- Equations on Pages 12-13: in general, I think the terms of the equations are not clearly presented. For instance, “Qt in bold” for predictions and “Qt~” for observations is only clearly stated in line 16 for Eq. 12, while it was already used in Eq. 10. Then in Eq. 14, we have a “Qt Teta”: is it the same as “Qt in bold”?

Further explanation of the notation has been included in Section 3.5 and 3.6, in particular to explain $\boldsymbol{Q}_t$ clearly. $\boldsymbol{Q}_t$ represents the predictive distribution of the forecasts at time t (definition now added Page 11 line 5), and Q_t^{θ} the hydrological model predictions at month t, before application of the residual error model. As such, the different notation is considered necessary.

- Metrics: some metrics seem to sum/average over the time series before computation (Eq. 10 and 11), while other compute at each time step (or forecast monthly lead time) and then take the average (Eq. 12 and Eq. 14). Maybe it is worth untangling this issue when introducing the equations.

The metrics presented in Section 3.7 are generally computed each time step. The standard approach to calculating the volumetric bias has been presented for consistency, however this could be rearranged to take the difference between the expected and observed streamflow each time step, and then take the sum the differences. As outlined at comment 2, the sharpness is considered to be calculated at each time step (as the standard deviation) and then summed over each time step, with the sum over the observed values used to normalise across scenarios. As such, the general approach to compute each metric is to (i) calculate a metric value at each time step, (ii) sum/average over all time steps, and (iii) normalize to improve interpretation. The exception is the reliability metric, where we determine if the observations are considered to be samples from the predictive distribution, which can’t be assessed separately at each time step.

- Page 13, line 16: “These changes…”: what changes exactly are you referring to? Please, clarify.

This sentence (“These changes are considered in detail below”) is considered unnecessary, as the changes are outlined in the following sentences. As such, it has been removed.

- Fig. 5: how do you explain the simulated peak flow at the end of 2002 that is not observed? Isn’t there a problem with the observations? Could you say something about it in the text?

The following has been added to Section 5.4 (Future Work):

These types of catchments are known to be challenging to model (McInerney et al., 2017; Ye et al., 1997). For example, the models predict a streamflow response in 2002 and 2005 in Figure 5 that did not occur in the observations, even when observed rainfall and state updating was used. Some of this difference may be due to errors in the input rainfall data, but this result highlights the difficulty in representing streamflow generation in low yielding, ephemeral catchments, such as those considered.

- Fig. 8: for C1, at least, the parameter “split” seems to evolve around 0.9, therefore, finally, it is worth considering it as a free parameter to calibrate? Could you say something about it in the text?

It is considered further analysis would be required to evaluate the benefits of including the split parameter a free parameter to calibrate. For the longer calibration period length, split~0.95, and the shorter calibration length the value changes over time from 0.95 to 0.8. Further work would be required to investigate if other parameters could have compensated from these changes from the default value of split=0.9, or the trends over time. No change is proposed.

- Discussion: please note that state updating was used for seasonal forecasting in “Crochemore, L., Ramos, M.-H., and Pappenberger, F., 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, Hydrol. Earth Syst. Sci., 20, 3601-3618, doi:10.5194/hess-20-3601-2016”. The authors also discuss about trade-offs between forecast attributes and it could be worth having a look at it.

Thank you for bringing this relevant study to our attention. The state updating approach used in this work is very similar to that used by Crochemore et al. (2016), and as such has been included in the manuscript here:

The approach used for the state updating of GR4J is similar to the approach of Crochemore et al. (2016) and Demirel et al. (2013)

The authors also discuss the trade-off between reliability and sharpness, and as such have also been cited here:

All other metrics (sharpness, bias, CRPS and NSE) show improvements from state updating in catchment C2, suggesting potential trade-offs in performance, similar to that found by Crochemore et al. (2016) and McInerney et al. (2017).

- Page 16, line 26: I am not sure about using the word “appreciably” here. I suggest deleting it or making more explicit what you mean here.

Removed.

- Page 17, line 8: “change to “These catchments were…”

Corrected.