Authors' responses to the comments of anonymous Reviewer 2

We would like to thank Reviewer 2 for the important and constructive criticisms and suggestions made to our manuscript. We substantially revise the manuscript in accordance with the suggestions. The main revisions are the following: (1) we fully re-write Introduction to point out the existing gaps in the area of interest and thereby to clarify motivation and objective of the study; (2) we include subsections containing description of the current operational forecast of inflow to the Cheboksary reservoir and comparison of the developed model-based forecast with the operational one; (3) we substantially revise the Result and Discussion section to stress our contribution, and (4) priori to resubmission we'll have a language check done by a native English speaker..

Below, we respond to the Reviewer's comments in a point-by-point manner.

1. Introduction: The authors state that "the purpose of this paper is to present the performance assessment of a long-term ensemble forecasting system of water inflow into the Cheboksary reservoir of the VKRC". I suggest re-formulating that purpose based on one or two science questions, whose answers could be found using the aforementioned system.

We revise Introduction to highlight motivation of the study. The corresponding fragments of the revised text are below:

> …utilizing the process-oriented hydrological models results in strengthening of physical adequacy of the forecast and, potentially, in improving forecast accuracy in comparison with the operational practice. However, this potential is rarely studied; to our knowledge the only example is the comprehensive experiment presented by Mendoza et al. (2017) and comparing the ESP model-based forecasts with the operational data-driven forecasts for a multi-year historical period. Our paper partly bridges this gap. We present development and verification of the ESP-based forecasts of water inflow into the Cheboksary reservoir of the VKRC and compare them against the operational forecasts for 35 years.
>
> …The observed weather scenarios that are used within the ESP framework do not encompass all of the possible weather conditions for the forecast period. … Hence the ensemble size is limited to the number of the historical years, statistical problems can appear stemming from large sample errors. For instance, Buizza and Palmer (1998) demonstrate improvement of the weather forecast skill as the ensemble size increases, wherein degree of improvement depends on the verification measure used. Particularly, the ranked probability skill score is strongly dependent on ensemble size and negatively biased (see also Müller et al. 2005, Weigel et al., 2007). Different aspects of the ensemble size effect on statistical properties of the ensemble weather forecast and verification scores are studied by Richardson (2001), Ferro et al. (2008), Najafi et al. (2012).The problem, can be solved by incorporating a stochastic weather generator (WG) into the ESP procedure… In this paper, we compare the ESP-based forecast with the WG-based forecast and assess possible advantage of the latter approach in forecasting rare hydrological events in the study basin and estimating verification measures.
>
> Thus, the motivation of this study is to answer two questions: (1) Does the model-based ESP technique allow one to improve reliability and skill of the operational forecast of spring inflow into the Cheboksary reservoir? (2) Does the enlarged ensemble size lead to any appreciable advantage when using the WG-simulated ensemble compared to the ESP-based ensemble?

2. ...Most of the text refers to the VKRC, with limited connection with recent literature on long-range hydrological forecasting (e.g., Schepen and Wang 2015; Mendoza et al. 2017; Beckers et al. 2016; Najafi and Moradkhani 2015; Demirel et al. 2015; Yossef et al. 2013; DeChant and Moradkhani 2014). A better link with current approaches will help readers to understand what is the contribution of this study.

In the revised manuscript, review of the recent literature is added, in particular using the listed publications

3. Methods. The authors mention that the first long-term forecasts for the VKRC are dated back to the 1930s and 1940s (P2, L8). In my opinion, the authors should include one or two benchmark methods – e.g., direct water balance methods, or index-based methods – to understand the added value of the proposed methodologies, ideally for several forecast initialization dates

We include new subsection 3.3.1 describing the current operational forecasting method and new subsection 4.3.2 describing comparison of the developed model-based forecast with the operational one

4. ...it is really hard for this reviewer to understand – from the information provided in the supplement section – the differences in forecast ensemble spread between ESP and WG-based technique. I think it would be helpful to see WG results contrasting boxplots or CDFs with observations for monthly precipitation amounts or temperature averages

In accordance with the Reviewer's suggestion, we add the corresponding boxplots to the supplement section (Fig. 8S).

5. The authors include both BS and BSS (having climatology as a reference) in Table 4, although they don't need both metrics to conclude that the WG-based approach is better for the event occurrence analyzed (similar to RPS and RPSS in table 5). Also, I strongly suggest to include some metric and/or graphic device for the assessment of forecast ensemble spread, since this is something that the authors point to without a solid quantitative basis (e.g., P16, L14). This could be done, for instance, using QQ plots (e.g., Thyer et al. 2009; Renard et al. 2010) or rank histograms (e.g., Hamill 2001; Delle Monache et al. 2006). The authors could further assess the ability of their forecasting system to distinguish between occurrence and non-occurrence by using discrimination diagrams (e.g., Clark and Slater 2006).

Discrimination diagrams and Q-Q plots are presented and analyzed in the Result and Discussion section of the revised manuscript.

Minor comments

1. P6, L14: I think that the authors should provide a short description of the calibration method, since the paper should be self-contained. Also, the authors state in P6-L9 that "most of the parameters are physically meaningful". I think that statement should be re-visited, because even measurable parameters have uncertainties associated with (i) observational errors, and (ii) their applicability at spatial scales that are different to those for which physically-based equations were developed.

The calibration method is described briefly in the sub-section 3.1 of the revised manuscript

> The ECOMAG calibration procedure is described in detail by Gelfan et al. (2015). Here, we emphasize two issues concerning the procedure. First, values of a few key-parameters pre-assigned from literature of from available measurements are considered as the initial approximation of the optimal values and the latter are quested within the closest neighborhood of the initial, pre-assigned values. Second, in the process of calibration, the ratios between the initial values of the distributed parameter relating to different soils, landscapes and vegetation are conserved.

2. P8, L16: Please provide a reference for the Cholesky's decomposition method

Reference is included

> Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. Numerical recipes: the art of scientific computing: 2.9 Cholesky decomposition, 3rd Edition, Cambridge University Press, 2007.

6. Verification metrics: it would be helpful to condense them in a table, including equation, possible range of values and references.

Verification metrics are summarized in new Table 1S added into the supplement section

7. P10, L27: The authors state that maximum inflow discharge is well simulated by the hydrologic model, although the plot (Fig. 6) still shows considerable spread around the 1:1 line.

The authors apologize for the fact that this particular panel was constructed erroneously. We revised the figure and added a one-standard-deviation-wide confidence band to the 1:1 line so one can see that there are only a few points outside this band.

8. Figure 6: Instead of using "lower-left", "lower-right", etc., I suggest using panels (a), (b), (c) and (d)

Corrected according to the reviewer's recommendation.

9. P12, L15: "1000-year Monte Carlo generated time series". Do you mean 1000- member ensemble? Please re-word.

Corrected according to the reviewer's recommendation.

10. Please clarify forecast initialization dates and forecasting approach in the caption of tables and figures.

Corrected according to the reviewer's recommendation.

11. Figure 9: In my opinion, the results from this figure could be better communicated using time series with ensemble forecasts as boxplots, including a line with observations (e.g., Bracken et al. 2010)

We'd prefer keeping this figure as is. In our opinion, CDFs better demonstrate performance of probabilistic forecast than boxplots

12. P20, L2-25: This should be moved to the Methods section.

Section 4.4 is removed from the revised manuscript because the forecastability issues turned to be out of the main framework of the study after the revisions

13. Forecast example for 2017: although this is a very interesting demonstration, I strongly encourage the authors to include verification metrics in their analyses.

Additional verification metrics are included