Authors’ responses to the comments of anonymous Reviewer 1

We would like to thank Reviewer 1 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 Results and Discussion section to emphasize our contribution, and (4) priori to resubmission we’ll have a language check done by a native speaker..

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

1. *The original contribution is not clear/significant enough*

> *To sum up my first major comment: I think the authors should clarify and emphasize their original contribution, which is not clear for me at the moment. One way to do so, besides phrasing it more clearly, is to include a comparison with at least one of the 3 operational forecasting systems described on page 2.*

We revise Introduction to highlight the motivation of the study and our original contribution. 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.

> *In fact, many of the results presented in the manuscript do not appear to me as a clear improvement over climatology. For instance, in Table 5, the skill scores obtained by the WG-based forecasts are all below 0.5. While I do agree that this represents an improvement over climatology, it is not a large one.*

We agree that the RPSS estimates are quite low and do not demonstrate clear improvement over climatology. In the revised Discussion section we consider this result and express our point of view on the possible ways for the forecast improvement. Not in order to justify the rather weak result, we'd like to note here that the values of RPSS<0.5 are not infrequently presented in well-cited publications related to the ensemble streamflow forecast verification (see, for instance, Greel et al., 2016 (Fig. 8c); Yuan et al., 2012 (Fig. 4); Franz et al., 2003 (Fig. 8))

Greuell W., Franssen W. H. P., Biemans Hester, Hutjes Ronald W. A. (2016) Seasonal streamflow forecasts for Europe – I. Hindcast verification with pseudo- and real observations. HESSD, doi:10.5194/hess-2016-603, 2016

Yuan X., Wood E.F., Roundy J.K. and Ming Pan (2012) CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. J. Climate, 26, 4828-4847

Franz J K Hartmann H C Sorooshian S and Bales R 2003 Verification of National Weather Service Ensemble Streamflow Predictions for water supply forecasting in the Colorado River Basin J. Hydrometeorology 4 1105-1118.

> *I don’t have any problem with this (a slight improvement over climatology), but it would probably be much more convincing to see (also) the improvement relative to at least one of the current operational methods mentioned on page 3.*

The improvement over the operational forecast is demonstrated in subsection 4.3.2.

Figures 13 and 14 also support my comment: the forecasts presented on Figure 14 appear only slightly different from the climatology presented on Figure 13. This is especially true for the forecasts issued on March 1^{st} (Figure 14 left) compared to Figure 13.

Indeed, the hydrograph predicted for the spring season of 2017 is close to the climatic one, but this proximity of the hydrographs is largely occasional. In most of 35 years, the predicted hydrographs were significantly different from the climatic one. As an illustration of this statement, Fig. 1R shows difference between forecasted and climatic hydrographs for some years of the verification period.

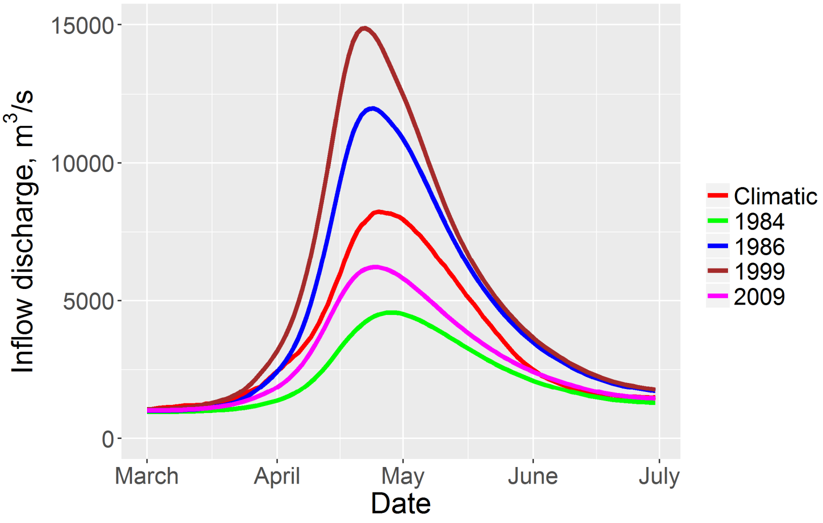

Figure 1R Forecast of daily inflow into the Cheboksary reservoir during March – June in selected years compared to the climatic mean inflow

<u>2. Some methodological/conceptual elements need clarification</u>

Page 3 line 30: I disagree with the formulation: "(. . .) incorporating a stochastic weather generator (WG) that will allow for reproduction of a hydrological system response to a large variety of possible weather conditions (. . .)". I think you might want to say that "(. . .) incorporating a stochastic weather generator (WG) that will allow for a large variety of possible weather conditions that can then be provided to the hydrological model (. . .)".

We agree with the Reviewer and revise the fragment in accordance with the suggestion.

Page 6, line 11: Is ECOMAG really taking daily precipitation intensities as inputs? As in mm/hour? All the models I know rather use total daily precipitation. Although it is true that mm/day can be seen as an intensity (since it is a quantity over time), it seems a bit unusual to me.

Precipitation intensity (L/T) as well as other flows (evaporation, infiltration, streamflow, etc.) is contained in the ECOMAG governing equations (see Motovilov et al., 1999). Since these equations are numerically integrated under 1-day time step, than ECOMAG really takes daily precipitation intensity in mm/day.

Motovilov, Yu., Gottschalk, L., Engeland, K., and Belokurov, A.: ECOMAG – regional model of hydrological cycle. Application to the NOPEX region. Department of Geophysics, University of Oslo, Institute Report Series no. 105. 1999.

Page 8, line 29: How many months is "several"? Is it at least one full year?

We have edited the respective paragraph in order to clarify this

(1) Spin-up ECOMAG-based simulations ("warm start") using meteorological observation data prior to the forecast issue date in order to calculate the initial watershed hydrological state (soil, snow and channel water contents, groundwater level, soil freezing depth, etc.) that initializes the forecast. The simulations start from the end of the previous freshet, i.e. 8-9 months before the forecast issue date

Page 9, Figure 4: On which basis did you chose to generate 1000 members from the WG while there are 50 in the ESP system? I suggest either setting the WG to issue the same number of ensemble members as EPS or at least justifying the choice of 1000 members and discussing the impact of ensemble size on performance assessment metrics.

We add several fragments relating to this issue into the revised manuscript. First of all, we include additional literature review in the Introduction (Buizza and Palmer, 1998; Richardson 2001; Müller et al. 2005; Weigel et al., 2007; Ferro et al. 2008; Najafi et al. 2012) and conclude that the forecast skill is improved "as the ensemble size increases, wherein degree of improvement depends on the verification measure used". In the Result section we highlight that the ranked probability skill score (RPSS) is strongly dependent on ensemble size and negatively biased. Then we add estimations of RPSS bias into the corresponding table (Table 5 in the first version of the manuscript) and show that the bias of the ESP-based forecast is two orders of magnitude larger than that of the WF-based forecast. In the revised manuscript, dependence of the RPSS bias on sample size is analyzed and the illustrating figure is added (see below as Fig. 2R). One can see from this Fig. 2R that under the used 35-member ensemble (i.e. the ESP-based ensemble) the bias can reach tens of percent depending on the RPSS estimate. Under the used 1000-member ensemble, the bias is close to zero.

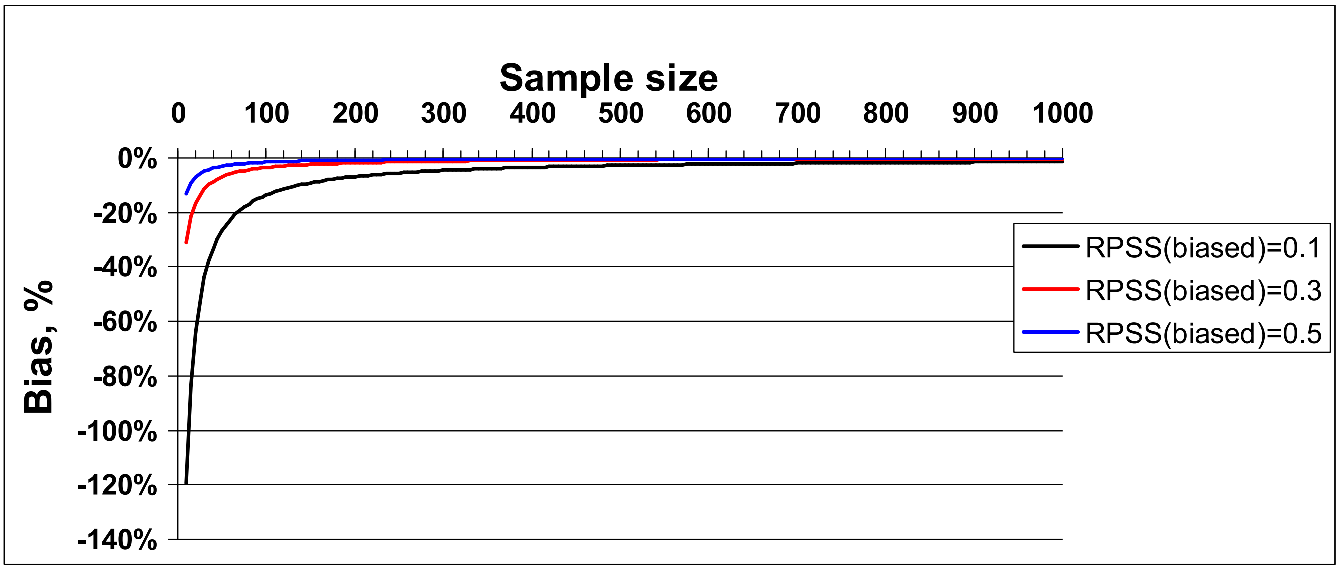

Fig. 2R. Negative bias of the RPSS-estimate in dependence on an ensemble size

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. Mon. Wea. Rev., 126, 2503–2518.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 19–24, 2008.

Müller WA, Appenzeller C, Doblas-Reyes FJ, Liniger MA. 2005. A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. Journal of Climate 18: 1513–1523.

Najafi, M. R., Moradkhani, H., and Piechota, T. C.: Climate signal weighting methods vs. Climate Forecast System Reanalysis, J. Hydrol., 442–443, 105–116, 2012.

Richardson, D. S.: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, Q. J. Roy. Meteorol. Soc., 127, 2473–2489, 2001

Weigel AP, Liniger MA, Appenzeller C. 2007. The discrete Brier and ranked probability skill scores. Monthly Weather Review 135: 118–124.

Page 14 line 17-18: According to Murphy (1973), "Hedging is said to occur whenever a forecaster's forecast r does not correspond to his judgement p (...) ". I don't understand how you associate your results to hedging. Hedging, by definition, arise from human intervention. Since your research does not involve human forecasters, I don't think hedging is the appropriate term here. Perhaps you want to refer to a systematic over forecasting bias in the forecasting system? In my opinion, this overforecasting is to be expected if the historical database includes many years with "higher than usual" precipitations. EPS (and WG) are very much dependent on the sample of data you have.

We revise the following sentence according to the reviewer's comment:

The hindcasts of Qmax show perfect detection estimates for both methodologies, but, as the frequency Bias is very high, this might be an outcome of overprediction, so with the high values of False Alarm Ratio and Hansen-Kuipers score.

Page 15 line 1: What do you mean by "forecast by chance"? Please define

We clarify the respective fragment as follows

However, the forecast accuracy with Heidke Skill Score of more than 60% is significantly better tan the accuracy of random chance

Page 15 line 10: "(. . .) comparing forecasts to climatology." I suspect you mean "streamflow climatology"? If so, this should be explicitly mentioned in the text everywhere applicable.

Changed in accordance to the suggestion.

Page 18 line 12: When you write "confidence bands", do you mean "confidence intervals"? If so, please provide the level of confidence and if not, please define what you mean by "confidence bands"

Indeed, confidence bands are closely related to confidence intervals and in some cases they are synonyms (see, for instance, Owen, 1995). Thus we use "confidence interval" term in the revised text and explicitly define the level of confidence in Figs. 10-11.

Owen, A.B. (1995). Nonparametric likelihood confidence bands for a distribution function. *J. American Stat. Association*. 90(430): 516–521

In section 4.4: are the results for ESP or WG? Globally, the explanations in this section (page 20) are difficult to follow. In my opinion, a schematic representation of the methodology would be helpful. And since this portion is, I think, more methodological, it should be moved to section 3.3

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

3. The analysis and discussion of the results is too shallow

Section 4 of the manuscript is labelled "Results and discussion". I was therefore expecting results to be discussed (rather than simply presented) in this section. However, I find this is not the case for many figures and tables. Specifically:

Table 2 (what do does values mean?),

In the revised version of the manuscript, it is clearly pointed out that Table 2 demonstrates meaning and values of the WG parameters

Table 3 (the "first" Table 3 on page 13) and Figure 8.

In the revised supplement section, all used verification measures are condensed in a Table 1S, including equation, possible range of values and references.

I think than Table 5 could also be discussed more, since, as I mention above, the improvement over climatology is still modest. However, without any other basis of comparison (such as the current operational forecasting system), it is hard to put the results in perspective. This is also related to Figures 13 and 14, which are just presented but not discussed. Those figures show that the improvement over climatology is very modest and hence, it is difficult to appreciate the authors' contribution. Similarly, Figure 15 should be analyzed more deeply (i.e. the explanations behind the results, not simply describing the figure).

The discussion section is substantially enhanced in accordance with the Reviewer's recommendation

In the conclusion (page 24 line 19-20), the authors implicitly mention a comparison with "the deterministic forecasts of inflow into reservoir that are used in common practice in Russia (. . .)", but this comparison is not explicitly shown in the manuscript.

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

Another thing that struck me is that the authors are not discussing the performance of their systems in terms of the relative importance of resolution and reliability. For instance on page 15 line 2, it is mentioned that the forecasts are "capable of detecting the occurrence of rare extreme events (. . .)" This is an indication that points toward forecast reliability, but what about resolution? If the forecasts are very widely dispersed, they will likely include any events but with very low power of discrimination. This should be studied and could help to improve the discussion.

Discrimination and reliability diagrams are presented and analyzed in the Results and Discussion section of the revised manuscript.

4. There are numerous spelling, orthographic and typographic errors throughout the manuscript

4.1 All figures except the first one and Figure 8 need reworking:

Figure 2: the resolution of the right hand side figure is very poor. All the small grey pixels should be removed.

Figure 2 is removed from the revised manuscript

Figure 3: I don't think this figure brings much information to the manuscript. It has no legend, and I think most readers are familiar with the requirements of a distributed, process oriented model. I suggest removing this figure.

Figure 3 is removed in accordance with the Reviewer's suggestion

Figure 4: The three small figures in the center of each middle box (representing plots of time series) are much too small and of poor resolution. I suggest either modifying them to make them readable, or removing them

Removed in accordance with the Reviewer's suggestion

Figure 5: The x-axis is labeled 'years' while the text says "daily inflow discharge". The label of the axis should reflect what is plotted on the figure. Labeling it "years" means that you would plot yearly values, not daily.

Corrected according to the reviewer's recommendation.

Figure 6: The legend is missing and the y-axis for 3 of the 4 panels need to be completed ("Simulated inflow volume, km^3" rather than just "Simulated")

Corrected according to the reviewer's recommendation.

Figure 7: Why is the Taylor diagram elliptical? Should it not be more spherical (a portion of a circle)?

Corrected according to the reviewer's recommendation.

Figure 9: The axes should be labeled (titles) ! Since all panels will all likely have the same axis titles, I would suggest writing axis titles only once for each: the x axis at the bottom of the figure, centered and the y axis completely on the left, also centered.

Corrected according to the reviewer's recommendation.

Figures 10-11: The text on the figures (labels, ticks, etc.) is so small, it is absolutely impossible to read anything. It should be made readable, both by increasing character sizes and figure resolution. In addition, the labeling "1", "2", etc under each panel is quite unusual. I advice labeling sub-figures (a), (b), . . . above each panel, as it is usually *done.*

Corrected according to the reviewer’s recommendation.

Figure 12: It is also difficult to read, although not as much as figures 10-11. The resolution of the figure could be substantially improved. Again, the labeling of the panels should be placed above, not below, each panel.

Figure 12 is removed from the revised manuscript

Figures 13-14: Same thing: difficult to read. The legend is missing for Figure 14.

Corrected according to the reviewer’s recommendation.

Figure 15: Labels for panels (a, b, . . .) are again misplaced. The axes ticks are very difficult to read (size and resolution).

Corrected according to the reviewer’s recommendation.

4.2. Table 3 on page 13 (the "first" Table 3): the units are all missing in the first column (W, QMax, Nq and Nmax).

Corrected according to the reviewer’s recommendation.

4.3 There are two tables labeled "Table 3"

Corrected according to the reviewer’s recommendation.

4.4 The Taylor diagrams should be explained briefly in the methodology. At the moment, all other performance assessment tools are at least mentioned in the methodology except this one.

Corrected according to the reviewer’s recommendation. The following fragment is added to the text:

> To illustrate forecast performance we used the Taylor diagram (Taylor, 2001) as it combines three forecast characteristics in one chart, namely the forecast standard deviation, RMSE and the correlation coefficient between the observations and the forecasted values. The values of all characteristics are normalized by dividing the RMSE and the standard deviations of the forecasts by the standard deviation of the observations. This normalization provides a vivid demonstration of the forecast efficiency expressed by RMSE fraction of the observed standard deviation (grey circular lines in Fig. 6). As long as the forecast RMSE is lower than the standard deviation of the observations, the forecast can be considered efficient against climatology.

5. English errors and typos

Page 3: line 9 instead of "in (Gelfan and Motovilov 2009)", it should be "in Gelfan and Motovilov (2009)". Similarly, at line 20, remove parenthesis around 2017 in "Arnal et al. (2017)". There are many similar errors with parenthesis around references in the manuscript.

Corrected according to the reviewer’s recommendation.

Page 2 line 31: Change "(. . .) allows forecaster to provide user (. . .)" either to "(. . .) allows the forecaster to provide the user (. . .)" or to "(. . .) allows forecasters to provide users (. . .)"

Corrected according to the reviewer's recommendation.

Page 2 line 33-34: Change "Recent studies illustrating ability of the ensemble (.. .)" to "Recent studies illustrating the ability of the ensemble(. . .)".

Corrected according to the reviewer's recommendation.

Page 3, line 11: remove the "the" from "Water Problems Institute of the Russian Academy of the Sciences

Corrected according to the reviewer's recommendation.

Page 3, line 23: Change "(. . .) but for possible weather condition (. . .)" to "(. . .) but also for possible weather conditions (. . .)".

Corrected according to the reviewer's recommendation.

Page 4, line 15: Replace "Also, analysis of (. . .)" by "Also, an analyse of (. . .)".

Corrected according to the reviewer's recommendation.

Page 8, line 22: Replace "(. . .) leads to increase of the model robustness. List of the (. . .)" by "(. . .) leads to an increase of the model's robustness.

Corrected according to the reviewer's recommendation.

A list of the (. . .)" W, Max, Nq and Nmax are sometimes in italics, sometimes not. Sometimes, the "max" in "Nmax"is in subscript ans sometimes not. Sometimes with a capital "M" and sometimes not. This needs to be uniformed according to the HESS's guidelines.

Corrected according to the reviewer's recommendation.

Page 10 line 11: remove "into" in the sentence "(. . .) in which the observation fell into (. . .)"

Corrected according to the reviewer's recommendation.

Page 12 line 1: Replace "Magnitude of the used metrics and error estimation has led to an assumption that the model is suitable to act as a core component of (. . .)" by "The magnitude of the performance assessment metrics and error estimations lead to the conclusion that the model is suitable as a core component of (. . .)".

Corrected according to the reviewer's recommendation.

Page 12 line 12: Replace "(. . .) tested through its ability" by "(. . .) tested through their ability (. . .)

Corrected according to the reviewer’s recommendation.