Authors' responses to the comments of anonymous Reviewer 1

We would like to thank Reviewer 1 for the important and constructive criticisms and the suggestions made. We have substantially revised the manuscript in accordance with the suggestions. The main revisions are in the following:

(1) we have completely re-written Introduction to highlight the existing gaps in the area of interest and thereby to clarify motivation and objective of the study;

(2) we have included 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 have substantially revised the Results and Discussion section to emphasize our contribution, and

(4) we have considerably revised English.

Below, we are responding to the Reviewer's comments, point-by-point.

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 have revised Introduction to highlight the motivation of the study and our original contribution. We have included new subsection 3.1 ("Operational data-driven forecast of spring inflow into the Cheboksary reservoir: a current practice") describing the current operational forecasting method and new subsection 4.3.1 ("Ensemble (model-based) and operational (data-driven) deterministic forecasts") 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. Without an attempt to characterise this result as a strong one (since it is of course not very strong), we'd like to note here that the values of RPSS<0.5 are not infrequent in the 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 4.3.1 ("Ensemble (model-based) and operational (data-driven) deterministic forecasts").

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 1st (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 incidental. 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.

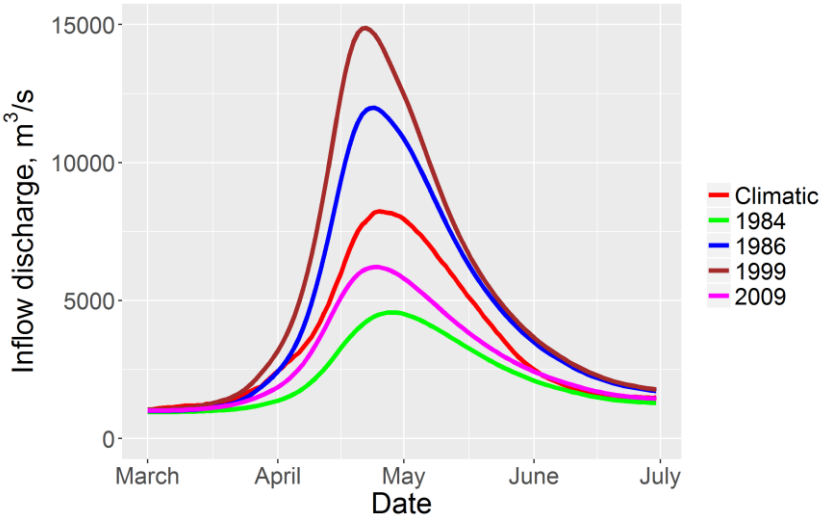

Figure 1R Forecast of daily inflow into the Cheboksary reservoir during March – June in selected years compared to the climatic mean inflow

2. Some methodological/conceptual elements need clarification

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 this fragment has been revised.

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, ECOMAG actually 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?

Agreed - it was not clear. We have revised the respective paragraph in order to clarify this

(1) Spin-up ECOMAG-based simulations ("warm start") using meteorological observations data prior to the forecast issue date (March, 31) 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.

Good point. We have revised the text adding several fragments related to this issue. First of all, we have included 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 concluded 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. 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. 12 that under the used 51-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.

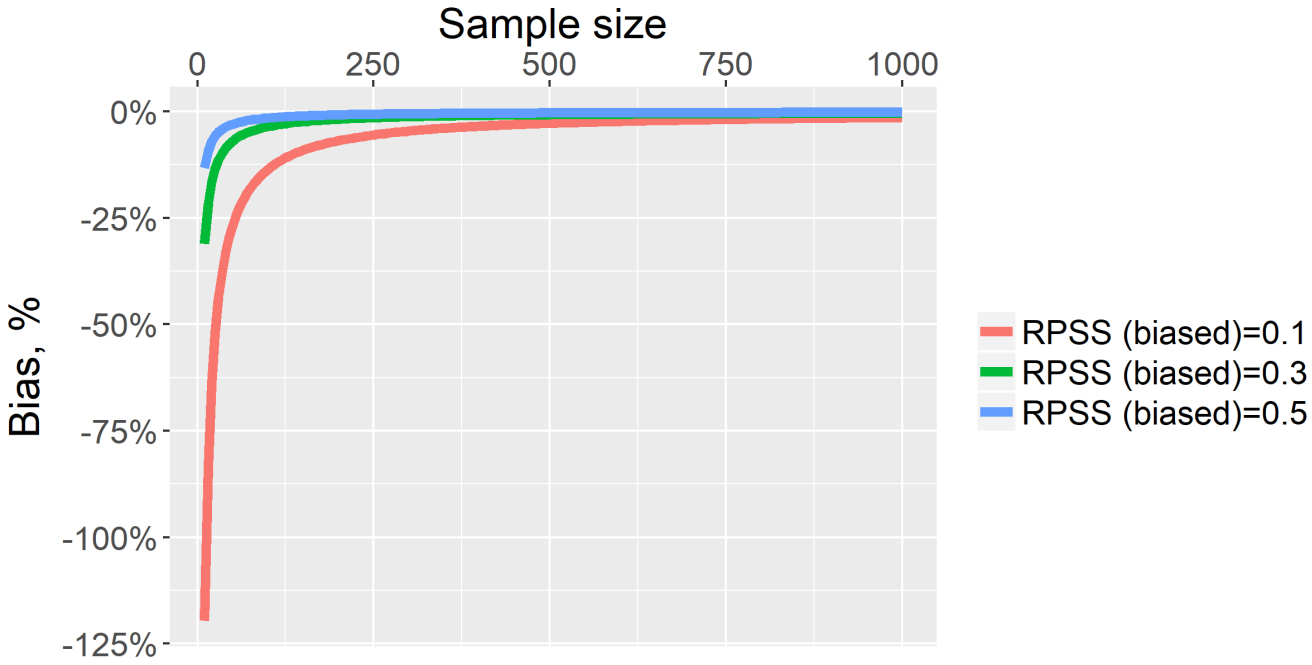

Figure 2R: Negative bias of the RPSS-estimate in dependence on the ensemble size and the RPSS value

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.

Good point. We have revised the following sentence according to the reviewer's comment:

> The forecasts show good detection estimates (even perfect for Q_{max}) for both model-based methodologies. However, as the frequency Bias is high, this might be the result 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 have clarified the respective fragment as follows

> For W and Qmax, the forecast accuracy with the Heidke Skill Score (HSS) of around 60% is better than 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.

Agrred. Changed in accordance to this 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

Agreed. Section 4.4 is removed from the revised manuscript because the predictability issues turned to be out of the main framework of this 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 this Table (replaced in the Supplementary Material) presents the meaning and the values of the WG parameters

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

In the revised supplement section, all the verification measures used 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 now substantially reworked and enhanced in accordance with the Reviewer's recommendation. Particularly, former Fig. 15 (Fig. 7 in the revised paper) is analyzed, as well as new Fig. 8 is added to compare the ensemble forecasts with the current operational forecasts.

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.

Agreed. We have included the new subsection 3.1 ("Operational data-driven forecast of spring inflow into the Cheboksary reservoir: a current practice") describing the current operational forecasting method and new subsection 4.3.1 ("Ensemble (model-based) and operational (data-driven) deterministic forecasts") 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.

Agreed. Discrimination and reliability diagrams are added into the Supplimentary material and are discussed 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 has been 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 has been 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.

(Fig. 2 in the revised paper) 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")

(Fig. 3 in the revised paper) 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)?

(Fig. 4 in the revised paper) 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.

(Fig. 11 in the revised paper) 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 has been removed from the revised manuscript

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

Figure 13 has been removed from the revised manuscript. Fig. 14 (Fig 6 in the revised version) has been 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).

(Fig. 7 in the revised paper) 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).

(Table 1 in the revised paper). Corrected according to the reviewer's recommendation.

4.3 There are two tables labeled "Table 3"

Corrected.

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 compare the ESP-based and the WG-based forecasts, we present them in the form of the Taylor diagram (Fig. 5; Taylor, 2001), which combines three forecast characteristics in one chart, namely, the forecast standard deviation, RMSE and the correlation coefficient between the observed and the forecasted values of the inflow characteristics. The values of all characteristics are normalized by dividing the RMSE by the standard deviation of the observations. This normalization provides a demonstration of the forecast efficiency expressed in fractions of the observed standard deviation. As long as the forecast RMSE is less 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 (. . .)"

The fragment has been removed

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

The fragment has been removed

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 (. . .)".

The fragment has been removed

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

The fragment has been removed

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.

The fragment has been removed.

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 (. . .)".

The fragment has been removed

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

Corrected according to the reviewer’s recommendation.

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 have substantially revised the manuscript in accordance with the suggestions. The main revisions are in the following:

(1) we have completely re-written Introduction to point out the existing gaps in the area of interest and thereby to clarify motivation and objective of the study;

(2) we have included 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 have substantially revised the Results and Discussion section to emphasize our contribution, and

(4) we have considerably revised English.

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.

Agreed. We have revised Introduction to highlight motivation of the study. The corresponding fragments of the revised text follow:

> …utilizing the process-oriented hydrological models results in increasing the physical adequacy of forecasts and, potentially, in improving forecast accuracy in comparison with the methods currently used in operational practice. However, such quantitative comparisons is not a common place; 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.
>
> …The observed weather scenarios that are used within the ESP framework do not encompass all of the possible weather conditions for the forecast period. It is desirable to account not only for the observed weather, but for possible weather condition that might lead to freshet events of rare occurrence. Assessing the magnitude of such an event might be crucial for decision making. Moreover, 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 synthetic, stochastically generated time series of weather variables instead of the historical data used within the ESP framework.
>
> …The studies and examples mentioned above serve as the background and the main motivation for this study. The objective of this study is to contribute to the EPS-related studies, with the focus on the comparative analysis between the data-driven techniques used in operational forecasts, and the ensemble forecasts of sreamflow, using two different weather scenarios: a) based on the historical data, and b) employing the WG-based forecasts. The case study is the Cheboksary reservoir of the VKRC cascade for which the operational forecasts are available since 1982.
> Thus, this study is an attempt to to answer the following two research questions: (1) Does the model-based ensemble methodology allow one to improve reliability and skill of the operational forecast of spring inflow into the Cheboksary reservoir, and to what extent? (2) Does the enlarged ensemble size lead to any noticeable 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 including 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 have included new subsection 3.1 ("Operational data-driven forecast of spring inflow into the Cheboksary reservoir: a current practice") describing the current operational forecasting method and new subsection 4.3.1 ("Ensemble (model-based) and operational (data-driven) deterministic forecasts") 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

Agreed. 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).

> Agreed. 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 the two issues concerning this procedure. First, the values of several key parameters pre-assigned from literature or from the available measurements are considered as the initial approximations of the optimal values and the latter are sought within the neighborhood of the initial, pre-assigned values. Second, during the calibration process, the ratios between the initial values of the distributed parameter corresponding to different soils, landscapes and vegetation are preserved. The Nash and Sutcliffe (1970) efficiency criterion NSE is adopted to represent the goodness of fit of the simulated and measured variables.

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.

Indeed. The authors apologize for the fact that this particular panel contained errors. We have 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 (Fig. 3 in the revised version).

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.

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 appreciate this suggestion, and discussed it, however, still would prefer to keep 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 predictability 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.

We have included comparison of the ESP-based and operational forecasts of water inflow into the Cheboksary reservoir for the period of 01/04/2017-30/06/2017 (Fig. 8)

Authors' responses to the comments of anonymous Reviewer 3

We would like to thank Reviewer 3 for the important and constructive criticisms and suggestions made to our manuscript. We have substantially revised the manuscript in accordance with the suggestions. The main revisions are in the following:

(1) we have fully re-written Introduction to point out the existing gaps in the area of interest and thereby to clarify motivation and objective of the study;

(2) we have included 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 have substantially revised the Results and Discussion section to emphasize our contribution, and

(4) we have revised Enlgish.

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

1. In the introduction section, the authors need to re-formulate their research objective based on the current state of literature. The current objective "present the performance assessment of a long-term ensemble forecasting system of water inflow into the Chebokary reservoir of the VKRC " seems too restrictive to a specific area

We have revised Introduction to highlight the motivation of the study and our original contribution. The corresponding fragments of the revised introductory section are below:

> …utilizing the process-oriented hydrological models results in increasing the physical adequacy of forecasts and, potentially, in improving forecast accuracy in comparison with the methods currently used in operational practice. However, such quantitative comparisons is not a common place; 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.
>
> …The observed weather scenarios that are used within the ESP framework do not encompass all of the possible weather conditions for the forecast period. It is desirable to account not only for the observed weather, but for possible weather condition that might lead to freshet events of rare occurrence. Assessing the magnitude of such an event might be crucial for decision making. Moreover, 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 synthetic, stochastically generated time series of weather variables instead of the historical data used within the ESP framework.
>
> …The studies and examples mentioned above serve as the background and the main motivation for this study. The objective of this study is to contribute to the EPS-related studies, with the focus on the comparative analysis between the data-driven techniques used in operational forecasts, and the ensemble forecasts of sreamflow, using two different weather scenarios: a) based on the historical data, and b) employing the WG-based forecasts. The case study is the Cheboksary reservoir of the VKRC cascade for which the operational forecasts are available since 1982.
> Thus, this study is an attempt to to answer the following two research questions: (1) Does the model-based ensemble methodology allow one to improve reliability and skill of the operational forecast of spring inflow into the Cheboksary reservoir, and to what extent? (2) Does the enlarged

ensemble size lead to any noticeable advantage when using the WG-simulated ensemble compared to the ESP-based ensemble?

2. This study compared ESP and weather generator forecasting schemes. These two methods are classic approaches for inflow forecasts and have been tested in many regions...

Indeed, the WG-based scheme is a classic approach for inflow forecasts and has been tested in many regions. However it was mainly done for the short-term forecasts but not for the long-term ones. Particularly, there are not too many attempts to use stochastic weather generator (WG) within framework of long-term ensemble forecasting. Hanes et al. (1977) were probably the first who used Monte-Carlo simulated sequences of daily precipitation to drive the conceptual US Geological Survey hydrological model and provide ensemble seasonal forecast of snowmelt runoff volume. A physically-based distributed hydrological model was used in combination with a weather generator to create a long-term probabilistic forecast of spring runoff of rivers in Central Russia in Kuchment and Gelfan (2007), Gelfan et al. (2015). Caraway et al. (2014) incorporated a stochastic weather generator into the ESP to make a probabilistic seasonal climate forecasts and applied the modified methodology to the San Juan River snowmelt dominated basin. Beckers et al. (2016) used ENSO-conditioned weather generator to compensate for the reduction of ensemble size in the post-processing ensemble forecast scheme presented for the Columbia River basin

In the listed papers, there have been no attempts to compare the ESP-based forecast with the WG-based forecast. Our study bridges this gap.

…The ESP approach is based on the ensemble of historical observed weather data. The weather generator approach generates synthetic weather data based on stochastic models. These two approaches also generated a different number of ensemble members: 50 versus 1000. This is like comparing apple and orange…

In our opinion, comparison of the ESP-based and the WG-based approaches makes sense because allows us to highlight the problem of limited ensemble size when evaluating the first approach.

We have included 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. In the revised manuscript, dependence of the RPSS bias on sample size is analyzed and the illustrating figure is added (see below as Fig. 1R). One can see from this Fig. 12 that under the used 51-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.

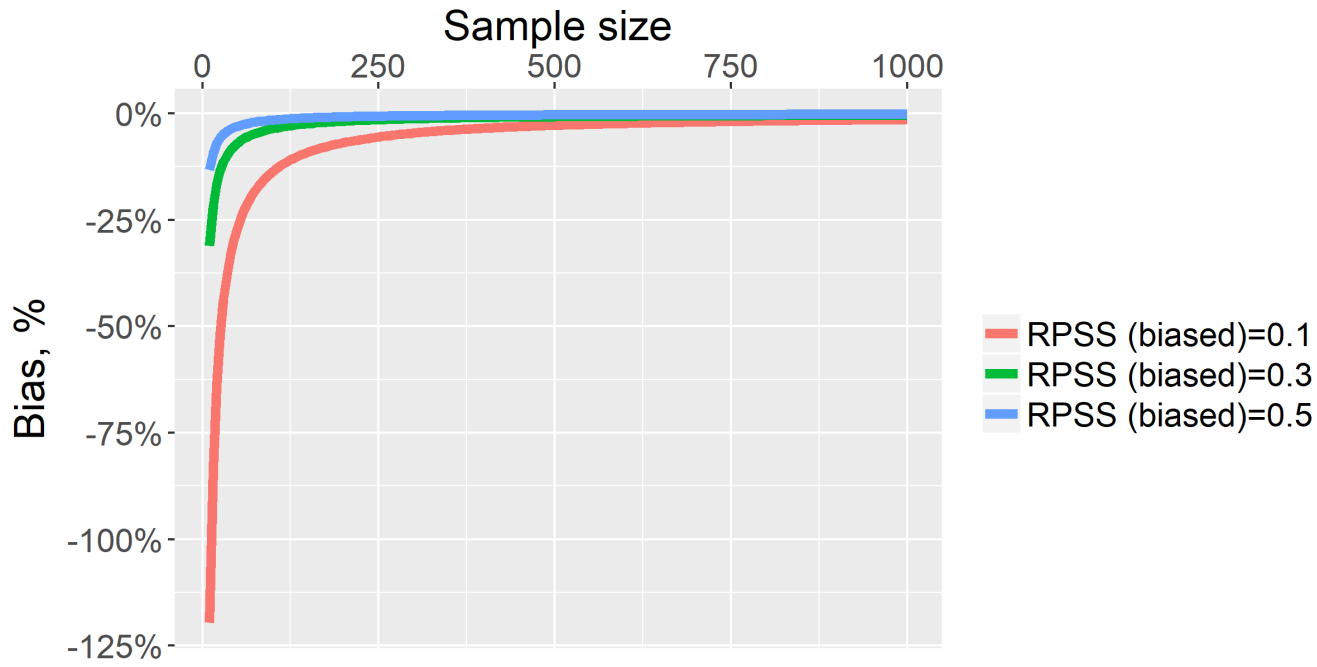

Figure 1R: Negative bias of the RPSS-estimate in dependence on the ensemble size and the RPSS value

The discussion and description of these two methods are too shallow. The authors need to clarify the motivations and implications of comparing these two forecasting schemes, with in-depth analysis and comments on these two methods.

In the revised manuscript, the Discussion section is substantially enhanced in accordance with the Reviewer's recommendation.

3. When evaluating probabilistic forecast, the authors used Brier skill score to compare the two forecast schemes with climatology of the inflow. The weather forecast forcing constructed using ESP is actually climatology of the weather variables. To compare these two forecasting schemes, I think it would be helpful to use ESP as a reference forecast relative to the WG-based forecast.

Thank you for the comment. Indeed, it is interesting to use ESP as a reference forecast and we have included the corresponding results in the revised manuscript.