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 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. 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 revise 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 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. 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 classic approach for inflow forecasts and has been tested in many regions. However this is the case for short-term forecasts but not for the long-term ones. Particularly, there are only a few attempts to use synthetic, stochastically generated time series of weather variables instead of the historical data within the ESP framework. 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 by Kuchment and Gelfan (2007) and 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) uses 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 are 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 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 argue that the ranked probability skill score (RPSS) depends on the ensemble size and negatively biased. Then we add estimations of the RPSS bias into the revised Table 5 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. 1R). One can see from this figure 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.

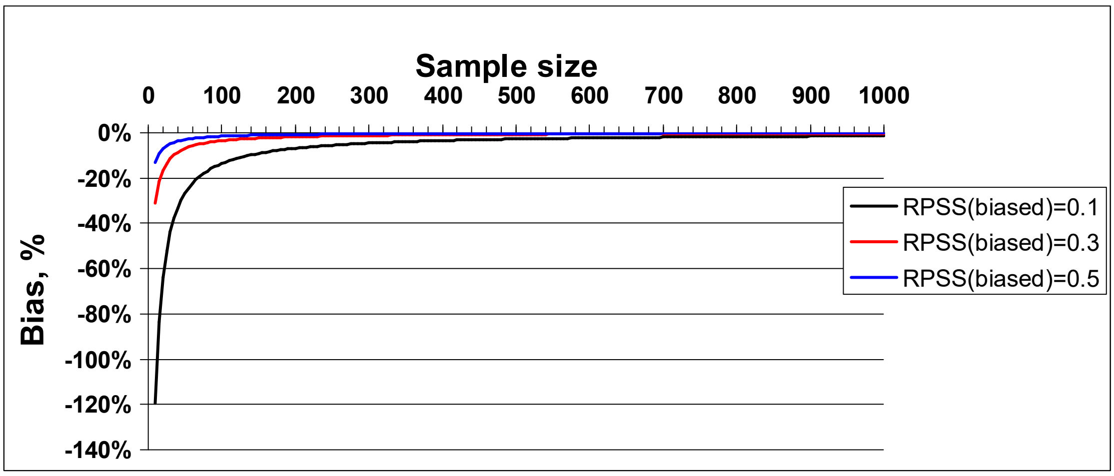

Fig. 1R. Negative bias of the RPSS-estimate in dependence on an ensemble size

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 very interesting to use ESP as a reference forecast and we include the corresponding results into the revised manuscript.