Response to Interactive comment Anonymous Referee #3

The authors proposed a methodology to give insight in the performance of ensemble streamflow forecasting systems in three streamflow categories (low, medium and high) and related runoff generating processes from lead times of 1 day to 10 day with a case study in a mountainous river catchment of less than 1000 sqr km in Poland. The quantitative precipitation forecasts and temperature forecasts extracted from the European Centre for Medium-Range Weather Forecasts (ECMWF) are averaged with catchment as input of a lumped hydrological (HBV) to generate ensemble streamflow. Several intensively used verification measures (CRPS, CRPSS, Rank histogram, Reliability diagram and ROC) are selected to evaluate the ensemble forecasts. Additionally, the pre-processing, post processing and updating of model initial states are adopted to improve the behavior of the system.

Generally speaking, the study gave an interesting investigation on the assessment of hydrological ensemble prediction system on different runoff processes including snowmelt, short-rain flood and so on, and a further analysis was made on the uncertainty source of these varied hydrometeorological conditions. There I suggest accept this manuscript after a moderate revision.

We thank the reviewer for the assessment. We appreciate the reviewer's opinion about the study and the valuable suggestions to improve the manuscript. Below are our responses to the comments and points raised.

There are a few issues list below that the authors should address:

***Comment:** 1) The logic in Paragraph 2 and 3 of Section 1 needs to be perfect. Some irrelevant statements can be removed, eg. SOME CONTENTS from Line 10 to Line 15 in Page 2 about EFAS are unnecessary to some degree.*

Reply: We agree with this comment. The text below is a revised version of paragraph 2 and 3 of Sect. 1.

"A number of studies investigated the performance of ensemble forecasting systems for different lead times, e.g. ~~Ye et al. (2014) for the European Centre for Medium-Range Weather Forecasts (ECMWF) medium-range ensemble precipitation forecasts,~~ Alfieri et al. (2014) for the European Flood Awareness System (EFAS), and Bennett et al. (2014), Olsson and Lindström (2008), Renner et al. (2009) and Roulin and Vannitsem (2005) for several catchments varying in size and other characteristics. ~~They~~ These studies all found a deterioration of performance with increasing lead time. ~~EFAS serves to provide high streamflow forecasts in large European river catchments for lead times between 3 and 10 days (Thielen et al., 2009). Relative to hydrological persistency the system skilfully forecasts high streamflow events for all lead times up to 10 days, with increasing skill for larger upstream areas (Alfieri et al., 2014). In EFAS critical flood warning thresholds are based on simulated streamflow, because model results and streamflow measurements can largely deviate (Thielen et al., 2009). EFAS is aimed at providing early warnings of possible flooding, instead of providing specific river streamflow forecasts (Demeritt et al., 2013).~~ Most studies on medium-range ensemble streamflow forecasting focused either on flood forecasts (e.g. Alfieri et al., 2014; Bürger et al., 2009; Komma et al., 2007; Olsson and Lindström, 2008; Roulin and Vannitsem, 2005; Thielen et al., 2009; Zappa et al., 2011) or low streamflow forecasts (Demirel et al., 2013; Fundel et al., 2013)~~,~~. ~~in contrast to~~The studies ~~to~~ on general ensemble streamflow forecasting systems (Bennett et al., 2014; Demargne et al., 2010; Renner et al., 2009; Verkade et al., 2013) do not evaluate the performance for different streamflow categories (e.g. low streamflow and high streamflow). Moreover, p~~P~~revious studies did not assess effects of runoff processes, like snowmelt and extreme rainfall events, on the performance of the ensemble forecasts. The only study we found that touches on this is the study by Roulin and Vannitsem (2005), who described that their high streamflow forecasting system is more skilful for the winter period than for the summer period. ~~For two Belgium catchments the high streamflow forecasting system of Roulin and Vannitsem (2005) is more skilful for the winter period than the summer period.~~ ~~Previous studies did not assess effects of runoff processes, like snowmelt and extreme rainfall events, on the performance of the ensemble forecasts.~~

Next to an assessment of performance of forecasts, i~~I~~nformation on the relative importance of uncertainty sources in forecasts is ~~helpful~~ essential to improve the forecasts effectively (Yossef et al., 2013). A number of studies report on how errors in the meteorological forecasts and the hydrological model contribute to errors in medium-range hydrological forecasts. Demargne et al. (2010) show that hydrological model uncertainties (initial conditions, model parameters and model structure) are most significant at short lead times. However, t~~T~~his also depends on the streamflow category~~.~~: h~~H~~ydrological model uncertainties significantly degrade the evaluation score up to a lead time of 7 days for all flows and up to a lead time of 2 days for the very high streamflow events. Renner et al. (2009) found an underprediction of low forecast probabilities (few ensemble members over a high streamflow threshold), which they attribute to the meteorological forecasts having ~~(~~insufficient variability~~)~~. On the other hand, the high forecast probabilities (low threshold) are overpredicted, which Renner et al. (2009) attribute to both the hydrological model and the meteorological input data. Olsson and Lindström (2008) found an ~~underestimation of the spread~~underdispersion of ensemble flood forecasts, to an extent that decreases with lead time. They conclude that the meteorological forecasts and the hydrological model have a comparable contribution to this ~~underestimation~~. In addition, Olsson and Lindström (2008) show overprediction of forecast probabilities over high thresholds, which they ~~mainly~~ primarily attribute to the meteorological forecasts. ~~Regarding low streamflow forecasts,~~ Demirel et al. (2013) concluded that uncertainty of hydrological model parameters has the largest effect~~, whereas~~ and meteorological input uncertainty has the smallest effect on low streamflow forecasts. Based on those studies we can say that for high streamflow forecasts uncertainties in the meteorological forecasts are dominant, whereas for low streamflow forecasts the uncertainties in the hydrological model become more important."

Comment: *2) Lines18-20 Page 6: A further explanation is expected why the training period is defined from 2011-2013 while the years previous to 2011 is used to validation.*

Reply: Our approach was triggered by practical considerations. We have serious doubts about the quality of the observation data in 2007: for the hydrological year 2007 (1 Nov 2006 – 31 Oct 2007) the agreement between observed discharge and simulated discharge with observed precipitation and temperature is poor (see table below). Therefore the hydrological year 2007 was excluded from further analysis.

The performance of the hydrological model for the hydrological year 2008 also raised some doubts about the quality of the observation data during this year. For this reason we started the pre- and post-processing with 2012-2013 (just two hydrological years to have a sufficiently long evaluation period left) as the training period, and we validated the pre- and post-processing procedures on both 2008-2011 and 2009-2011. There was no significant difference in validation performance of the pre- and post-processing procedures between these two periods and also the hydrographs of observations and simulations do not indicate poor quality of observation data for 2008, so in the end we included 2008 in the validation period.

Table 1: Validation performance per hydrological year

Hydrological year	NS	E_{RV} [%]	Y
2007	-1.34	43.41	-0.94
2008	0.22	17.14	0.19
2009	0.53	-4.67	0.51
2010	0.93	0.07	0.93
2011	0.59	6.20	0.55
2012	0.62	19.47	0.52
2013	0.46	12.79	0.41

We noticed that the validation performance numbers in Table 4 of the paper do include the hydrological year 2007. We will recalculate these numbers after excluding 2007.

Comment: *3) In Section 3.2, it is not necessary to introduce all the evaluation scores in details, for the CRPS, CRPSS, Reliability diagram and ROC can be regarded as "industry standards" in ensemble forecasting, so simply citing the relevant references.*

Reply: We agree to the comment and will omit general information about the evaluation scores (P7 Line 13-15, Line 18-20, P8 Line 14-21, P9 Line 2-5, Line 12-15, Line 16-18). In Sect. 3.2 we will address what aspect of forecast quality a score evaluates and refer to other studies for further details.

Comment: *4) In Section 4.1.2, it is confusing that since the QM pre-processing brings improvement to the precipitation and temperature forecasts, why the conclusion is that the strategy 0 results in the best CRPS.*

Reply: We agree to the reviewer that this is a remarkable result. The results indicate that the slight improvement of the meteorological forecasts by the pre-processing procedure loses its effect after propagating through the hydrological model. We will add this finding to the conclusion of the paper (P17 Line 12).

Comment: *5) The figures about rank histograms and reliability diagrams are missing or not shown intentionally?*

Reply: The figures about rank histograms, reliability diagrams and ROC curves are not shown by intention to keep the paper short. However, we think that these results are relevant and therefore we described them in words. We agree that this makes reading of the paper difficult and the results nontransparent. We could make the figures available by a supplement to the paper.

Comment: *6) The catchment area is less than 1000km2 and the data used are daily. For flood forecasting in such catchment area, is it daily data too coarse? Perhaps 3h or 6h subdaily data are more useful for flood forecasting in such area. Please make it an elaborate story.*

Reply: We thank the reviewer for the comment but note that discharge measurements are available at a daily resolution. For this reason we applied and evaluated the forecasting system at a daily time step. When focusing on short-range forecasts (lead times of 0-2 days), we agree that smaller time steps are preferred for a mountainous catchment of about 1000 km^2 like the Biala Tarnowska catchment. We focus on medium-range forecasts (0-10 days), for which the very quick streamflow response is less important.

Comment: *7) For flood forecasting, flood peak, volume and peak time are all important. Can these be analyzed in the study?*

Reply: We agree with the reviewer that, in addition to discharge, the peak streamflow, volume and peak time are important, particularly for operational high streamflow forecasting systems. Despite the relevance, we propose not to include the analyses of these aspects in the paper. Looking at the number of pages the paper already has we must be selective in what we can include. Moreover, essential to the topic of the paper is that next to high streamflows we also evaluate the streamflow forecasting system on low streamflows and medium streamflows. In view of the paper length already we cannot evaluate low streamflows, medium streamflows and high streamflows on all relevant aspects, such as duration and discharge deficits regarding low streamflows.

Comment: *8) Page 9: It is not very clear how the errors are contributed in Section 3.3. Why can CRPSsim/CRPSmeans represent the error contribution? Please add more details.*

Reply: Evaluation against observed discharge ($CRPS_{meas}$) is affected by errors from the meteorological forecasts, the hydrological model and measurement errors. By evaluation against simulated discharge based on observed precipitation and temperature ($CRPS_{sim}$), the ensemble streamflow forecasts and the reference streamflow contain similar hydrological model errors and no streamflow measurement errors, so these are eliminated. If we neglect measurement errors we get:

$$\frac{CRPS_{sim}}{CRPS_{meas}} \sim \frac{meteorological\ forecast\ errors}{meteorological\ forecast\ errors + hydrological\ model\ errors}$$

If this ratio is low the hydrological model errors are dominant and if this ratio is high the meteorological forecast errors are dominant. The same approach is used by Demargne et al. (2010), Olsson and Lindström (2008) and Renner et al. (2009).

To clarify this explanation we will add the equation above.

References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D. and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, J. Hydrol., 517, 913–922, doi:10.1016/j.jhydrol.2014.06.035, 2014.

Bennett, J. C., Robertson, D. E., Shrestha, D. L., Wang, Q. J., Enever, D., Hapuarachchi, P. and Tuteja, N. K.: A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days, J. Hydrol., 519, 2832–2846, doi:10.1016/j.jhydrol.2014.08.010, 2014.

Bürger, G., Reusser, D. and Kneis, D.: Early flood warnings from empirical (expanded) downscaling of the full ECMWF Ensemble Prediction System, Water Resour. Res., 45(W10443), doi:10.1029/2009WR007779, 2009.

Demargne, J., Brown, J., Liu, Y., Seo, D. J., Wu, L., Toth, Z. and Zhu, Y.: Diagnostic verification of hydrometeorological and hydrologic ensembles, Atmos. Sci. Lett., 11(2), 114–122, doi:10.1002/asl.261, 2010.

Demirel, M. C., Booij, M. J. and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49(7), 4035–4053, doi:10.1002/wrcr.20294, 2013.

Fundel, F., Jörg-Hess, S. and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, Hydrol. Earth Syst. Sci., 17(1), 395–407, doi:10.5194/hess-17-395-2013, 2013.

Komma, J., Reszler, C., Blöschl, G. and Haiden, T.: Ensemble prediction of floods - catchment non-linearity and forecast probabilities, Nat. Hazards Earth Syst. Sci., 7(4), 431–444, doi:10.5194/nhess-7-431-2007, 2007.

Olsson, J. and Lindström, G.: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden, J. Hydrol., 350(1–2), 14–24, doi:10.1016/j.jhydrol.2007.11.010, 2008.

Renner, M., Werner, M. G. F., Rademacher, S. and Sprokkereef, E.: Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376(3–4), 463–475, doi:10.1016/j.jhydrol.2009.07.059, 2009.

Roulin, E. and Vannitsem, S.: Skill of Medium-Range Hydrological Ensemble Predictions, J. Hydrometeorol., 6(5), 729–744, doi:10.1175/JHM436.1, 2005.

Thielen, J., Bartholmes, J., Ramos, M. H. and De Roo, A.: The European Flood Alert System - Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13(2), 125–140, doi:10.5194/hess-13-125-2009, 2009.

Verkade, J. S., Brown, J. D., Reggiani, P. and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, doi:10.1016/j.jhydrol.2013.07.039, 2013.

Yossef, N. C., Winsemius, H., Weerts, A., Van Beek, R. and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resour. Res., 49(8), 4687–4699, doi:10.1002/wrcr.20350, 2013.

Zappa, M., Jaun, S., Germann, U., Walser, A. and Fundel, F.: Superposition of three sources of uncertainties in operational flood forecasting chains, Atmos. Res., 100(2–3), 246–262, doi:10.1016/j.atmosres.2010.12.005, 2011.