Response to Interactive comment Anonymous Referee #1

General comments

Comment: *This manuscript presents an interesting analyse of the performance of hydrological ensemble predictions. The skills are screened according the regime (low and high streamflow) and the generating processes (snow melt, short rain, long rain floods etc.). This study further disentangles hydrological model errors and errors from meteorological forcing. The methodology is applied to a mountainous catchment. The combination of existing methodologies is pertinent and is worth being published in HESS.*

However the reading is not easy and a major revision is necessary. Some information is redundant in the introduction, methodology and results sections and long lists of references are not always necessary. The focus should be made on the main contribution of the paper i.e. the analysis of the skill for different hydro-meteorological conditions and skip or shorten secondary experiments. Some validation methodologies are described but their results are not shown. A balance should be found: either shorten the description or include those results. Some suggestions are given in the specific comments. The English should be improved.

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

The reviewer's suggestion to improve the flow of the paper is valuable, and the specific comments contain many relevant points for this.

With respect to the comment to increase the focus of the paper on the main scientific innovation, we will leave out the additional updating experiment, which has also not been used because it was unsuccessful (P5 Line 31 – P6 Line 2, P11 Line 3 – P11 Line 5).

Regarding the experiments on pre- and post-processing of the ensemble forecasts we consider this important and propose not to remove it from the paper. The procedure is common, so removing it will presumably result in doubts about why we have not applied a correction procedure. The results of this experiment are quite striking and we will add a figure with CRPS values for the different pre- and post-processing strategies showing this finding (see Figure 1).

Further replies to this comment follow below in response to the specific comments.

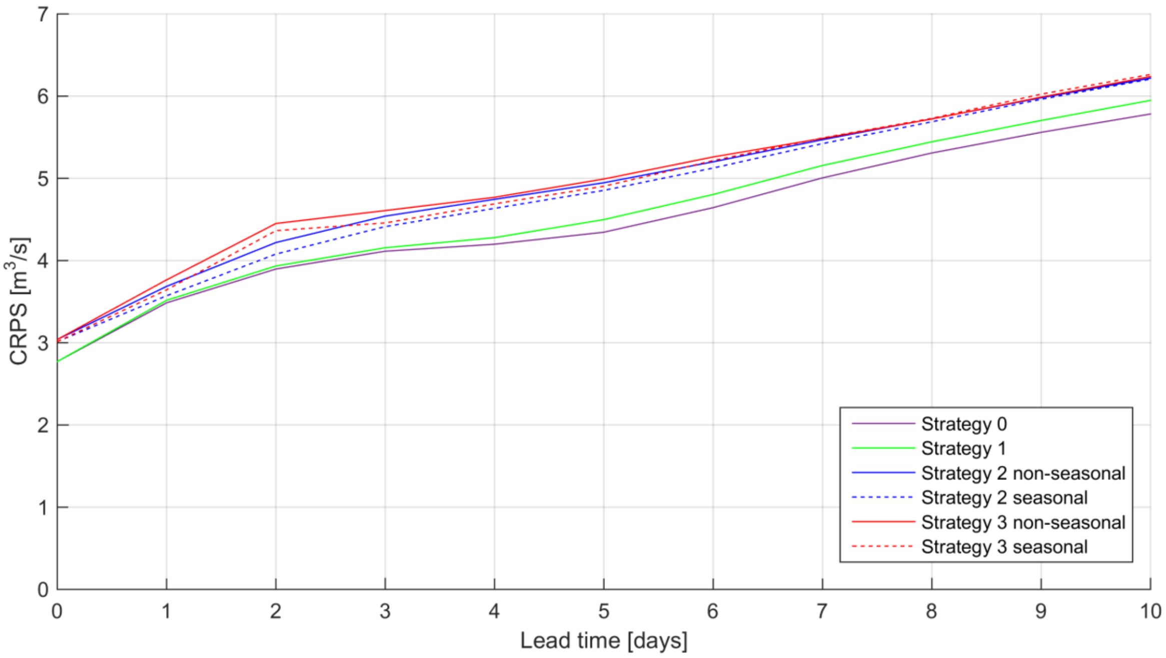

Figure 1: CRPS of the post-processing strategies over the validation period 2008-2011

Comment: *The authors are using ensemble predictions from ECMWF from 2007 to 2013 with a training of the pre- and post-processing during two water years between 2011 and 2013. They associate the failure of the quantile mapping for post-processing method to the short time series of forecasts for training and to the inconsistency of the bias between the training and the validation period. They forget that the ensemble prediction system has undergone many changes during this period including spatial resolution changes. This is why retrospective forecasts are available since long and provide samples of 18 to 20 years back for post-processing purposes. Re-forecasts have been widely used and reported in the literature. These meteorological re-forecasts have also been used for the preparation of hydrological re-forecasts for the statistical postprocessing of hydrological ensemble predictions.*

Reply: It is correct that we used meteorological forecasts from a system that has undergone changes. The TIGGE data portal contains the operational forecasts from meteorological forecast centres. We agree that this affects the pre-processing and post-processing results and we thank the reviewer for this suggestion. We will add a statement to Page 15 Line 15-16 that the joint distribution of measurements and forecasts is nonhomogeneous in time, because the meteorological forecast system has undergone changes during our analysis period (Mladek, 2016):

https://software.ecmwf.int/wiki/display/TIGGE/Model+upgrades#Modelupgrades-ECMWF

Comment: *Figure 5 to 9 are the core of the paper. They will gain value if the plots are associated with confidence intervals.*

Reply: CRPS and CRPSS are the main evaluation scores that we used. In recent literature these scores are commonly applied without associated confidence intervals or statistical tests, by Demargne et al. (2010), Hersbach (2000), Pappenberger et al. (2015), Renner et al. (2009), Verkade et al. (2013), and Ye et al. (2014). We agree to the suggestion that confidence intervals around the CRPS values would add value to the figures, but we consider establishing such confidence intervals outside the focus of this paper.

Comment: *The use of the term "perfect forecast" is questioned because it is neither a forecast nor perfect and, would the future meteorological forcing be known, predictions with the model would include growing errors due to initial conditions as somehow shown in Table 5.*

Reply: We appreciate the comment. The term "perfect forecast" was introduced by Olsson and Lindström (2008), but the term is somewhat misleading. For the same concept, Renner et al. (2009) used the term "baseline simulation", Demargne et al. (2010) used the term "simulated flow", Verkade et al. (2013) used the term "simulated streamflow" and Bennett et al. (2014) used the term "perfect-rainfall-forced forecasts". We propose to use "observed meteorological input forecasts".

Specific comments

Comment: *P1, L20-24 Should be rephrased e.g. too many occurrence of "improve".*

Reply: We agree to the comment and will change it to:

"To improve the performance of the forecasting system for high streamflow events, ~~in particular~~ the meteorological forecasts ~~require improvement~~are crucial. For low streamflow forecasts, ~~the hydrological model should be improved~~it is advised to calibrate a hydrological model specifically on low streamflow events. The study further recommends improving the reliability of the ensemble streamflow forecasts, by including the uncertainties in hydrological model parameters and initial conditions, and by ~~improving~~ increasing the dispersion of the meteorological input forecasts."

Comment: *P3 L23-P4, L3 How do you correct measurement? Do you correct each station for the difference between the elevation of the station and the average of the elevation in the area defined by the intersection of the Thiessen polygon corresponding to the station and the watershed? Then average the corrected values of the stations using their relative contribution to the catchment area as weights?*

Reply: The assumption of the reviewer is correct: this is the procedure that we used. We will revise the text to make this clear:

"Precipitation, temperature and streamflow measurement series are available at a daily time interval for the period 1 January 1971 to 31 October 2013, provided by the Polish Institute of Meteorology and Water Management. Precipitation and temperature data from 5 measurement stations (Fig. 1) have been selected because of their distribution over the catchment and data series completeness. ~~The data are spatially interpolated based on Thiessen polygons (Fig. 1) to represent catchment averages.~~ Given that stations are mostly located in valleys and precipitation and temperature vary with elevation, the catchment averages ~~are~~ may be biased (Panagoulia, 1995; Sevruk, 1997). Following Akhtar et al. (2009), precipitation measurements are corrected using relative correction factors (in %), whereas temperature measurements are corrected using absolute correction factors (in °C). The precipitation correction factor differs considerably between months. For December–February the mean precipitation gradient is 10.5 % 100 m^{-1}, while for March–November the mean precipitation gradient is 5.4 % 100 m^{-1}. Although the number of stations is limited to accurately determine precipitation and temperature gradients, the calculated precipitation gradients are used because of the clear difference between two periods. The temperature gradient does not vary much over the year and therefore the global standard temperature lapse rate of 0.65 °C 100 m^{-1} is applied. The corrected ~~data~~ measurements are ~~spatially interpolated~~weighted based on the relative coverage of Thiessen polygons (Fig. 1), to represent catchment averages. By the corrections the annual mean precipitation increases from 741.2 mm to 768.4 mm and the annual mean potential evapotranspiration decreases from 695.3 mm to 674.4 mm."

Comment: *P5, L20-21 Equations would be appropriate here in order to define Y, NS and E_RV.*

Reply: We hesitate to add the equations since Y, NS and E_{RV} are defined in the given references.

Comment: *P5, L28 preceding the first forecast day.*

Reply: We agree. We will change it to: “the day preceding the forecast issuing day” (from comment on P11, L21).

Comment: *P5, L32-P6, L2 I would suggest to skip this experiment or, if impossible to skip, tell already that it failed (according to P11, L3-5). This is to lighten the methodologies to keep in mind until the result section.*

Reply: We agree to leave this out. Also see the response to the first general comment.

Comment: *P6, L31-P7, L11 Some information (and references) is redundant with the sub-sections.*

Reply: We agree. Also looking at comment 3 by Reviewer 3 we will omit general information about the evaluation scores, but focusing on what aspect on forecast quality each score evaluates and citing the relevant references.

Comment: *P7, L1 Three properties of probabilistic forecast quality* ...

Reply: We do not understand this comment.

Comment: *P7, L8 “The histograms accompanying ...” the histograms of what?*

Reply: We will change this sentence to:

“~~The histograms accompanying reliability diagrams are used to evaluate sharpness.~~ To evaluate sharpness we use histograms which show the sample size over the range of forecast probability bins used to establish the reliability diagrams.”

Comment: *P7, L20-21 “CRPS approaches the average value of the evaluated variable” What do you mean with “approaches”?*

Reply: We will change “approaches” to “converges to”.

Comment: *P7, L24-27 “and compares the forecasts with a relevant alternative forecast” somehow redundant with the beginning of the sentence.*

Reply: We will change this sentence to:

“~~Normalizing the CRPS against the CRPS of alternative forecasts eliminates the effect of the magnitude of the investigated variable and compares the forecasts with a relevant alternative forecast (i.e. skill), used by e.g. Bennett et al. (2014), Demargne et al. (2010), Renner et al. (2009), Velázquez et al. (2010) and Verkade et al. (2013).~~

To eliminate the magnitude of the investigated variable we normalize the CRPS against the CRPS of a relevant alternative forecast, a principle which is also used by Bennett et al. (2014), Demargne et al. (2010), Renner et al. (2009), Velázquez et al. (2010) and Verkade et al. (2013) to evaluate forecast skill.”

Comment: *P8, L1-2 “... argue that this” choice ... these two lines should be rephrased. I would prefer a positive phrasing saying that the choice of another alternative forecast may result in a more robust estimation of forecast skill.*

Reply: We propose to delete P7 Line 30 – P8 Line 2, because it is not really relevant to explain the procedure that we followed. It explains why we have not applied hydrological persistency or hydrological climatology as alternative forecast set, but we can focus the text on what we have done: using the forecast set with the lowest CRPS values as alternative forecast set, because this set is most difficult to beat in performance.

Comment: *P8, L22-23 Either provide an equation for the "numerical indicator delta" if it adds to the understanding of the adopted methodology or skip any reference to delta.*

Reply: We will remove the reference to delta.

Comment: *P8, L30-31 "... contain a random element ..." explain how it works for the flatness coefficient.*

Reply: We will add further explanation about the random element:

"In this case a random rank is assigned from the set of ensembles and the measurement that have the same value."

Comment: *P9, L3 "... for a certain event ..." It would be useful to define "event" and refer to sub-section 3.4 or Table 1.*

Reply: We will specify "certain event" as "for low streamflow events and high streamflow events (defined in Sect. 3.4)".

Comment: *P9, L24-28 Almost the same thing is repeated.*

Reply: We will delete P9 Line 24-26.

Comment: *P9, L29 At a first reading, it was tempting to replace this ratio with a CRPSS of sim against meas but the purpose is different and since it is a major tool in this paper, this paragraph should be written with much care.*

Reply: We will add the equation below (see also comment 8 by Reviewer 3):

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

If this ratio is low, the hydrological model errors are dominant and if this ratio is high, the meteorological forecast errors are dominant.

Comment: *P10, L11-12 Are the rules given also by Merz and Blöschl or defined for this catchment based for instance on data from both simulation and observations during the training period?*

Reply: The study by Merz and Blöschl (2003) is used to characterize the high streamflow generating processes in Table 2. The rules for classification are defined specifically for the study catchment and are based on observations and model simulations. We will change the text to:

"Various runoff contributing processes can result in high flows. Table 2 defines the processes and ~~classification~~ rules for classification ~~we use in this study, based on the processes Merz and Blöschl (2003) distinguish~~. The rules for classification are based on rainfall observations and snowpack model simulations; at one day before the event because of the time step of the HBV model."

Comment: *P10, L16 Do you mean that the distribution of the generating processes shown in the figure is like we can expect for this region?*

Reply: The reviewer's interpretation is correct. We will change this to:

"Figure 4a presents the distribution of high streamflow generating processes over the year following the ~~classification~~ rules for classification in Table 2. ~~The figure shows an expected distribution of~~

~~processes for this region.~~ The distribution of processes over the year is like we can expect for this region."

Likewise we will change P10 Line 20-21.

Comment: *P10, L19, Table 3 What is the rule for precipitation deficit?*

Reply: The rule used for classifying an event as a precipitation deficit generated low streamflow is that if there is a low streamflow event and if there is no snowpack present (based on model simulations) we assume that the low streamflow event is caused by a precipitation deficit. We think that the definition in Table 3 is clear.

Comment: *P11, L21 "preceding day" the day before the forecast issuing day.*

Reply: We agree and we will change "preceding day" to "day preceding the forecast issuing day" accordingly in the paper (also see comment P5 Line 28).

Comment: *P11, L28-29 "not shown in the paper " therefore, going back to section 3.1.3, the methodology description should be simpler and not encumber with strategy numbers.*

Reply: This comment is discussed in the response to the first comment.

Comment: *P10 L20 What do you mean by "reliable distribution"?*

Reply: See response to comment P10, L16.

Comment: *P12, L13 with more skill instead of "skilful"*

Reply: Skill is defined as the performance of the streamflow forecast relative to the performance of alternative forecasts. Here we do not mean 'with more skill', but skilful relative to the alternative forecasts.

Comment: *P12, L16 "functional" what do you mean?*

Reply: We will change "functional" to "plausible".

Comment: *P12, L28 "... are in general less predictable by historical measurements ..." please re-phrase*

Reply: P12 Line 28-29 is partly a repetition of the preceding sentence, so this sentence will be deleted. We will change P12 Line 27-29 to:

"In addition, ~~these events are~~ high streamflow events will be less well captured ~~in~~ by historical measurements, and thus ~~in~~ the alternative forecasts will have lower quality for these events. ~~This is because high streamflow periods are in general less predictable by historical measurements, in particular in small catchments.~~"

Comment: *P12, L32 "not shown" a figure is missing with the rank histograms for the low streamflow forecasts and for the high streamflow forecasts, two lead times. Apparently, for high flow, the rank histogram is not exactly U-shaped but skewed according to P13, L12-13.*

Reply: To keep the paper short we chose not to include these figures in the paper. However, we think that the 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: *P13 L10-13 Difficult to figure out ... Please add a figure with the reliability diagrams and corresponding sharpness histograms for the low streamflow forecasts and for the high streamflow forecasts two lead times.*

Reply: See response to comment P12 L32.

Comment: *P13 L15-17 Note that good sharpness without reliability is useless.*

Reply: We agree. We will emphasize this in the conclusion (bullet 1).

Comment: *P13, L18 reference already given, please re-phrase.*

Reply: We agree. We will change this to:

"All AUC values are above 0.85,~~ whereas Buizza et al. (1999) consider 0.8 as indicative for good prediction systems~~ which indicates a good resolution of the streamflow forecast system."

Comment: *P14, L11-13 "... the below zero skill ... do not result in positive skill ..."*

Reply: We agree that this sentence is not well written. We will change the sentence to:

"~~The below 0 skill of long-rain and snowmelt flood forecasts indicate that the meteorological forecasts at small lead times do not result in positive skill as compared to forecasts based on historical meteorological measurements.~~

For long-rain floods and snowmelt floods, the meteorological forecasts at small lead times do not result in positive skill as compared to forecasts based on historical meteorological measurements."

Comment: *P14, L23 What is the amount of this fake drizzle?*

Reply: This is an interesting question, but we consider this to be out of the focus of this paper.

Comment: *P14, L24-26 Re-phrase: "... meteorological forecasts accumulated in the forecasting system are better model inputs ..."*

Reply: We agree that this sentence is not well written. We will change the sentence to:

"The skill increases for larger lead times, so for larger lead times ECMWF meteorological forecasts accumulated in the forecasting system ~~are better model inputs~~give better predictions than historical measurements~~ for larger lead times~~."

Comment: *P15, L8 & Figure 10 I would skip this figure which highlights the weakness of drawing such a detailed profile with just a water-year data. The legend is missing for the thin plain lines.*

Reply: We hesitate to skip this figure, because it illustrates why the pre- and post-processing procedures are not working: the training period and validation period show different bias distributions, because of the short time series.

The thin plain lines are showed in the legend as "Single years 2007-2013". We will add an explanation to the caption that each thin line refers to a single year between 2007 and 2013.

Comment: *P16, L8-10 Do you have evidence that such coincidence occurs and is the main explanation for the high ratio for short-rain floods?*

Reply: This is an interesting question and we will investigate how often this occurs.

Comment: *P17, L13-15 "longer time series of forecasts", "longer forecasts datasets" see general comments; "more sophisticated" and first of all more robust.*

Reply: We had to deal with the limitations of available data and to focus on the objective of the study we made choices in the development of the ensemble forecasting system. In the responses to the general comments and in the responses to Reviewer 2 and Reviewer 3 these choices are further explained.

References

Akhtar, M., Ahmad, N. and Booij, M. J.: Use of regional climate model simulations as input for hydrological models for the Hindukush-Karakorum-Himalaya region, Hydrol. Earth Syst. Sci., 13(7), 1075–1089, doi:10.5194/hess-13-1075-2009, 2009.

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.

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.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15(5), 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Merz, R. and Blöschl, G.: Regional flood risk - what are the driving processes?, in Water Resources Systems-Hydrological Risk, Management and Development, edited by G. Blöschl, S. Franks, M. Kumagai, K. Musiake, and D. Rosbjerg, pp. 49–58, International Association of Hydrological Sciences Press, Wallingford, UK. [online] Available from: http://hydrologie.org/redbooks/a281/iahs_281_049.pdf, 2003.

Mladek, R.: Model upgrades, TIGGE [online] Available from: https://software.ecmwf.int/wiki/display/TIGGE/Model+upgrades#Modelupgrades-ECMWF (Accessed 7 March 2017), 2016.

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.

Panagoulia, D.: Assessment of daily catchment precipitation in mountainous regions for climate change interpretation, Hydrol. Sci. J., 40(3), 331–350, doi:10.1080/02626669509491419, 1995.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A. and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble predictions, J. Hydrol., 522, 697–713, doi:10.1016/j.jhydrol.2015.01.024, 2015.

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.

Sevruk, B.: Regional dependency of precipitation-altitude relationship in the Swiss Alps, Clim. Change, 36(3–4), 355–369, doi:10.1023/A:1005302626066, 1997.

Velázquez, J. A., Anctil, F. and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrol. Earth Syst. Sci., 14(11), 2303–2317, doi:10.5194/hess-14-2303-2010, 2010.

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.

Ye, J., He, Y., Pappenberger, F., Cloke, H. L., Manful, D. Y. and Li, Z.: Evaluation of ECMWF medium-range ensemble forecasts of precipitation for river basins, Q. J. R. Meteorol. Soc., 140(682), 1615–1628, doi:10.1002/qj.2243, 2014.