Author response HESS-2016-584 8 May 2017

Title: Performance of ensemble streamflow forecasts under varied hydrometeorological conditions

Authors: Harm-Jan F. Benninga, Martijn J. Booij, Renata J. Romanowicz, Tom. H.M. Rientjes

Dear editor,

Thank you for your consideration of the paper. Based on the comments by the three reviewers on the first version of the manuscript we have revised the manuscript regarding explanation of methods, explanation of results and English writing. We have added Fig. 5 and additional background figures in a supplement. The revised manuscript is uploaded.

This document contains a point-by-point reply to the comments and a marked-up manuscript. Page and line numbers refer to the first version of the manuscript. We have updated revised texts in the point-by-point reply with the text in the uploaded revised manuscript (minor changes compared to the responses on 18 March 2017). In the cases that our response to a comment has changed compared to 18 March 2017, we have indicated this by keeping the original reply (18 March 2017) and adding an additional reply below it (8 May 2017).

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.

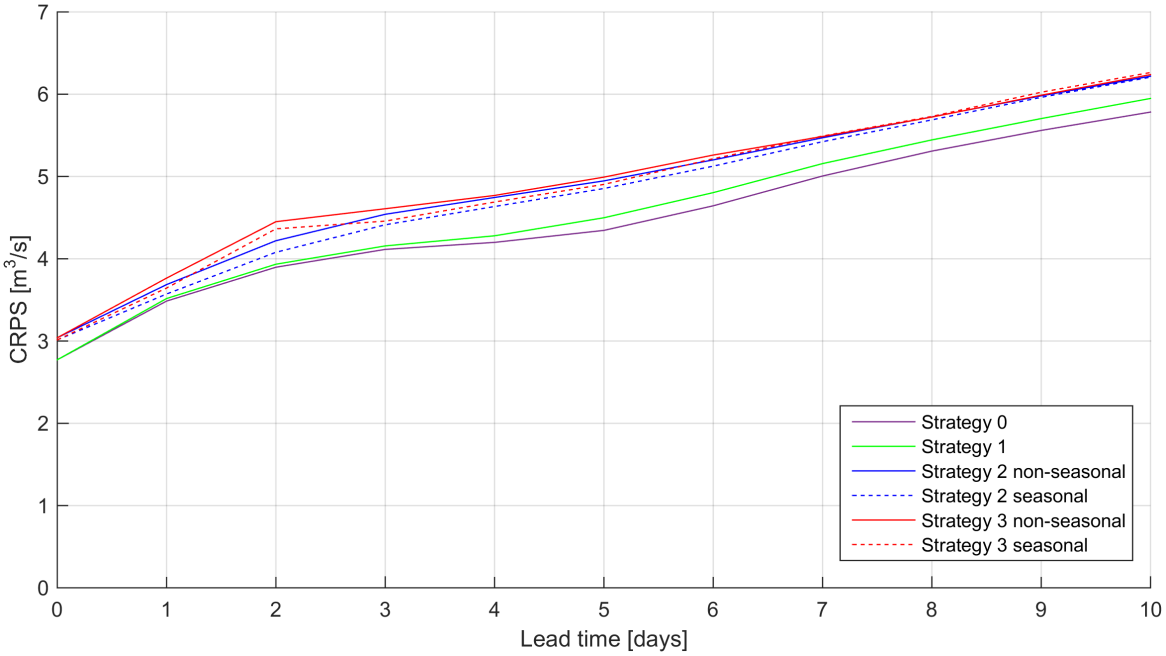

Figure 1: CRPS of the post-processing strategies over the validation period 2008-2011

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,~~ It is recommended to calibrate the hydrological model specifically on low streamflow conditions and high streamflow conditions. ~~the hydrological model should be improved. The study~~ It is further recommend~~ed~~s ~~improving~~ that the ~~reliability~~ dispersion (reliability) of the ensemble streamflow forecasts is enlarged by including the uncertainties in hydrological model parameters and initial conditions, and by ~~improving~~ enlarging 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 and~~,~~ temperature ~~and streamflow~~ measurement series are available from five meteorological stations and streamflow measurement series are available from one discharge gauging station, at a daily time interval for the period 1 January 1971 to 31 October 2013, and 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 meteorological 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~~gradient 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~~ small to accurately determine precipitation and temperature gradients, the calculated precipitation gradients are used because of the clear difference between the 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 measurements from each station are corrected for the difference between the elevation of the station and the mean elevation its respective Thiessen polygon. To represent catchment averages, the corrected measurements are

weighted based on the relative coverage of their Thiessen polygon (Fig. 1). 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 employ the histograms that show the sample size of the forecast probability bins used to establish the reliability diagrams (Ranjan, 2009; Renner et al., 2009; WMO, 2015)."

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 to the measurement from the pool of ensemble members 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 18 March 2017: We will delete P9 Line 24-26.

Reply 8 May 2017: On second thought we consider P9 Line 24-26 valuable in the text. The first sentence explains that the streamflow measurement error and the hydrological model error are eliminated by evaluation against observed meteorological input forecasts, whereas the second sentence explains how this is used to investigate the contribution of error sources.

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~~ generating 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 used in 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 is typical 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 18 March 2017: 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."

Reply 8 May 2017: The explanation of evaluation scores in Sect. 3.2 is shortened and the reference is omitted there. Therefore we keep the reference in this section.

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

Reply 18 March 2017: 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."

Reply 8 May 2017: We remove this sentence, because at this point in the paper it is clear that skill is generated by the ECMWF meteorological forecasts compared to 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 meteorological 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 18 March 2017: This is an interesting question and we will investigate how often this occurs.

Reply 8 May 2017: We have further investigated this question. There are two possible cases if the ensemble forecast is closer to the measured streamflow than to the observed meteorological input forecasts: 1. the observed meteorological input forecast is closer to the measured streamflow (example in Fig. 1), and 2. the ensemble forecast is closer to the measured streamflow (example in Fig. 2). The second case indicates a hydrological model deficiency: in the rainfall-runoff relation or in the flood peak timing. Table 1 lists the numbers associated with both cases, based on CRPS calculations for each day classified as high streamflow.

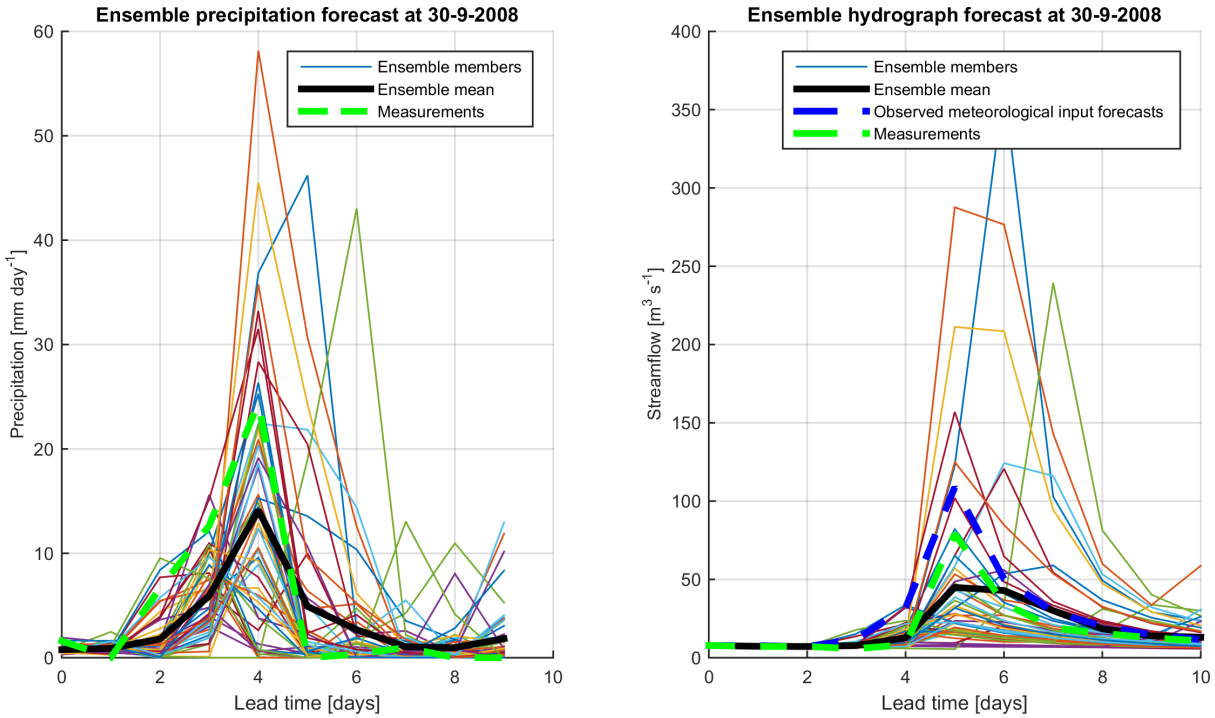

Figure 2: Example in which ensemble forecast set is closer to the measured streamflow than to the observed meteorological input forecast

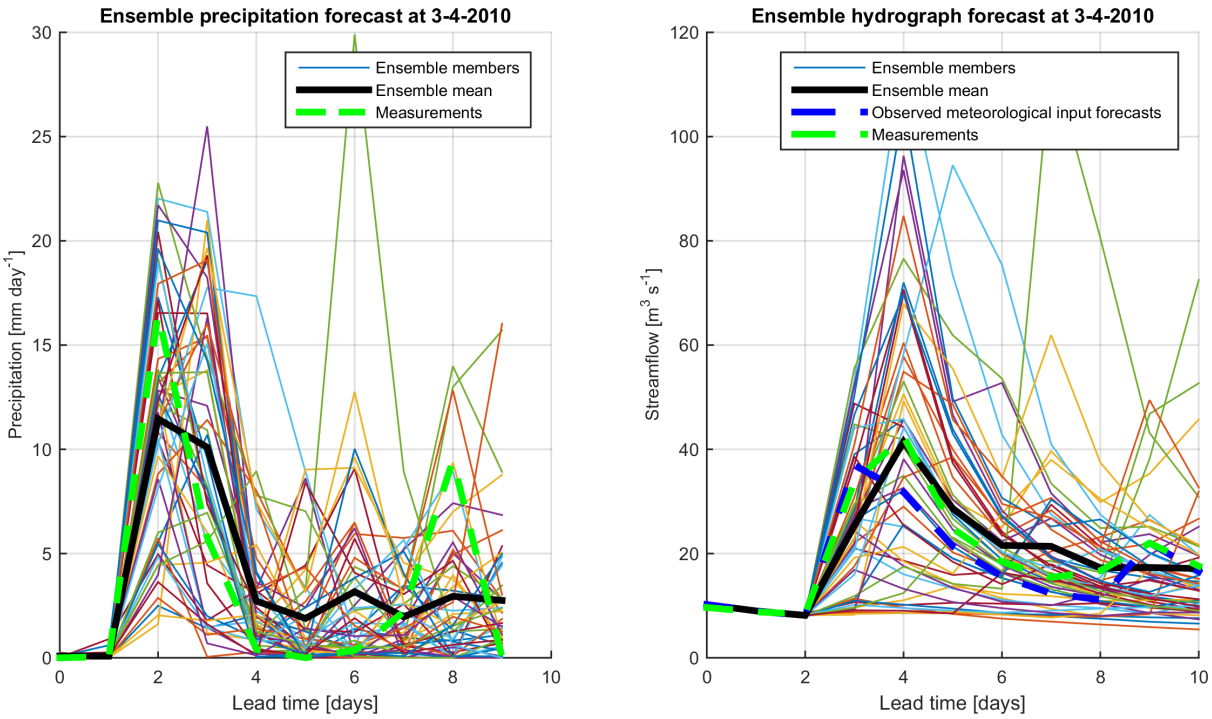

Figure 3: Example in which the ensemble forecast is closer to the measured streamflow than to the observed meteorological input forecast due to a shifted peak

Table 1: Numbers of days that observed meteorological input forecast is closer to measured streamflow and number of days that ensemble forecast is closer to measured streamflow, in case the ensemble forecast is closer to the measured streamflow than to the observed meteorological input forecast. This is based on CRPS calculations.

Lead time	# Observed meteorological input forecast closer to measured streamflow	# Ensemble forecast closer to measured streamflow
0	0	0
1	59	96
2	75	114
3	81	142
4	85	165
5	94	173
6	96	173
7	95	169
8	105	159
9	115	152
10	130	132

As mentioned in the paper one cause of the ensemble forecasts being closer to the observed streamflow than to the measured streamflow is:

“The precipitation peak in the measurements and the precipitation peak in the meteorological forecasts can be shifted one day with respect to each other and this can cause that the timing of the peak of the streamflow forecasts better corresponds to the streamflow measurements than to the peak of the perfect streamflow forecasts.” (Page 16, Line 8-10)

We have further investigated this by comparing the days of the peak streamflow of the observed streamflow series, observed meteorological input forecast series and the mean of the ensemble forecasts. Days of peak streamflow are defined as days with the highest streamflow in periods of 5

days and the peaks must be separated by at least 4 days in between, resulting in 97 peak streamflow days. Table 2 lists the results.

Table 2: Peak day correspondence between observed streamflow, observed meteorological input forecasts and ensemble forecasts. The total number of peak days is 97.

Lead time	# Peak day observed meteorological input forecast matches to peak day observed streamflow	# Peak day mean ensemble forecast matches to peak day observed streamflow	# Peak day observed meteorological input forecast does not match to peak day observed streamflow, but peak day ensemble forecast matches to peak day observed streamflow
0	47	47	0
1	47	42	17
2	49	28	12
3	52	26	8
4	55	23	9
5	53	28	13
6	55	20	6
7	54	26	13
8	54	20	9
9	53	22	9
10	54	19	8

These examples and the numbers illustrate that hydrological model deficiencies have a large effect on both the observed meteorological input forecasts and the ensemble forecasts. To improve the ensemble forecast system, the study outcomes show that the hydrological model needs to be improved, with a special attention to flood peak timing. We will change the text in the paper accordingly:

"The results in Fig. ~~8b~~ 9b show that the ratio between the CRPS against ~~perfect forecasts~~observed meteorological input forecasts and the CRPS against streamflow measurements is above 100% for ~~short-rain floods~~high streamflows, and short-rain floods in particular. This means that these forecasts are closer to the measurements than to the ~~perfect forecasts~~observed meteorological input forecasts. ~~The precipitation peak in the measurements and the precipitation peak in the meteorological forecasts can be shifted one day with respect to each other and this can cause that the timing of the peak of the streamflow forecasts better corresponds to the streamflow measurements than to the peak of the perfect streamflow forecasts.~~Analyses show that on high streamflow days on which the forecasts are closer to the measurements than to the observed meteorological input forecasts (28% at lead time of 1 day to 48% of the days at lead time of 10 days), depending on lead time, on 50% to 66% of the days the forecasts are closer to the measurement than the observed meteorological input forecast. This indicates a hydrological model deficiency, either from the rainfall-runoff relation or the flood peak timing. The precipitation peak in the measurements and the precipitation peak in the meteorological forecasts can be shifted one day with respect to each other and this ~~can~~may cause that the timing of the peak of the streamflow forecasts better corresponds to the streamflow measurements ~~than to the peak of the perfect streamflow forecasts.~~ Of the 97 separate peak streamflow days, on 6 days (lead time of 6 days) to 17 days (lead time of 1 day) the flood peak day of the observed meteorological input forecasts does not match to the peak day of the measurement but the peak day of the mean of the ensemble forecast does match to the peak day of the measurements. This illustrates that hydrological model deficiencies have a considerable effect on the observed meteorological input forecasts and the ensemble forecasts."

In the conclusion at bullet 2 we will add:

"Also the hydrological model performance on high streamflow must be improved, by specific calibration on streamflow during high streamflow events and flood peak timing."

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.

Ranjan, R.: Combining and Evaluating Probabilistic Forecasts, PhD thesis, University of Washington, Seattle, Washington USA., 2009.

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.

WMO: Forecast Verification: Issues, Methods and FAQ, [online] Available from: http://www.cawcr.gov.au/projects/verification/ (Accessed 12 March 2015), 2015.

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.

Response to Interactive comment Anonymous Referee #2

General comments:

This paper summarizes the application of the widely used HBV hydrologic model to streamflow forecasting in a Polish mountain river. The project uses ECMRWF ensemble weather forecasts to drive the streamflow model, and explores both pre- and post-processing of the ensembles for bias correction. Useful results are obtained, and the study has significant potential. I recommend that the paper is accepted pending major revisions.

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

Detailed comments:

Comment: *1. The paper repeatedly refers to HBV as a spatially lumped model. This isn't just terminology, as around lines 20-25 of page 15, the manuscript seems to imply that the model assumes a snowpack to be present (or absent) across the entire model domain. There are a few versions of HBV, but it's normally viewed as semi-distributed, using (at a minimum) elevation bands.*

Reply: We appreciate the comment but see little opportunity to soundly add more detail in representing elevation bands. We have chosen to apply a lumped version of the HBV model, without elevation bands, because the available measurement data does not justify to enter multiple bands. For the area only five meteorological stations are available, which cannot be used to represent multiple elevation bands over the complete elevation distribution of the catchment. Following a first analysis on streamflow simulation results, there was no clear signal that model performance is largely affected by lumping so we considered it plausible to rely on the lumped model approach. If requested by the reviewers we could add a comment to Sect. 3.1.1 to address these considerations.

Comment: *2. The manuscript makes a good point on lines 29-30 of page 1 about socio-economic development increasing the impacts of extreme hydrometeorological events. It also probably bears mentioning that climate changes, both natural and anthropogenic, may further exacerbate these impacts. See Perkins, Pagano, and Garen, "Innovative operational seasonal water supply forecasting technologies," Journal of Soil and Water Conservation, 2009; and Fleming, "Demand modulation of water scarcity sensitivities to secular climatic variation: theoretical insights from a computational maquette," Hydrological Sciences Journal, 2016.*

Reply: We thank the reviewer for the comment and refer to P1 Line 29-30:

"Accurate forecasting becomes increasingly more important, ~~since~~ because frequency and magnitude of low and high streamflow events are projected to increase in many areas in the world as a result of climate change (IPCC, 2014). Due to socio-economic development ~~also~~ the impacts of extreme events further increase (Bouwer et al., 2010; Fleming, 2016; Rojas et al., 2013; Wheater and Gober, 2015)."

The first sentence aims to mention that climate change exacerbate both low and high streamflow events. Following the reviewer's comment we will add "as a result of climate change" to make the statement more explicit. The paper by Fleming (Hydrological Sciences Journal, 2016) is a good reference for the second sentence.

Comment: *3. Terms could stand to a little better defined. For example, most flood and water supply forecasters who I know would regard "short-term" forecasts as having lead times of 0-10 days, and "long-term" forecasts as having lead times of weeks to months. So what the authors refer to here as "medium-term" would be referred to as "short-term" by many if not most others working in the field. And no effort is made here to distinguish medium-*

term from short-term hydrologic forecasting. More broadly, some of the wording throughout the manuscript would benefit from a re-think for better clarity and precision.

Reply: To be consistent with respect to forecast windows, we explicitly define "medium-range" forecasts and follow the definition for "medium-range" by the World Meteorological Organization, which is also followed by ECMWF (ECMWF, 2012). WMO defines medium-range as forecasts with lead times from 3 days to 10 days, and we also refer to Olsson and Lindström (2008), Renner et al. (2009), and Roulin and Vannitsem (2005). We note that Bennett et al. (2014) refer to this range of lead times as "short-term" forecasts, so there is ambiguity. We opt to keep the term "medium-range" instead of changing it to "short-range", to remain consistent with definitions commonly used in meteorology.

In this paper the term "medium-range" is just used as a generic term to characterize the forecasting system. We do not explicitly distinguish short-range forecasts and medium-range forecasts, because in the analyses there is always referred to specific lead times.

Comment: *4. Why is only meteorological forecast uncertainty incorporated into the ensemble model? It's commonplace in the research literature for forecast models to include both meteorological uncertainty (NWP ensemble) and hydrologic model parameterization uncertainty (ensemble of hydrologic parameter values). This work is starting to make its way into operational practice too. Providing some justification for this choice might be a good idea.*

Reply: We agree to the comment, but argue that only meteorological forecast uncertainty is incorporated because this study aims to identify effects of the ECMWF meteorological forecasts on the quality and skill of streamflow forecasts. Additionally incorporating hydrological model uncertainty, parameter uncertainty and initial condition uncertainty would (partly) obscure this relation. In addition, Bennett et al. (2014), and Cloke and Pappenberger (2009) state that uncertainties in meteorological forecasts are the largest source of uncertainty beyond 2-3 days, and that only uncertainty in meteorological forecasts is incorporated in many studies (Bennett et al., 2014). We will add the above in Sect. 3.1.

Comment: *5. The description of the model implementation isn't quite adequate. What was the calibration-testing split, and what were the model performances during both phases? And it's stated that the objective function selected for calibration is "Y", which apparently combines the Nash-Sutcliffe efficiency with a volumetric error measure. Objective function selection is a key step in model calibration, and more information needs to be provided, starting with an explicit mathematical definition for "Y".*

Reply: We refer to P5 Line 16-23 where the calibration procedure is explained. The equations for Y, NS and E_{RV} are directly accessible in the cited references and we therefore hesitate to add the equations. The calibration and validation performances are listed in Table 4 and referred to on P10 Line 25-28.

Comment: *6. The updating of initial states was performed here for the slow-runoff and fast-runoff reservoirs. That's interesting and useful, but why was SWE not selected as the object of this data assimilation exercise? It seems like it would be a more rewarding, and certainly more conventional, choice in this northern continental European mountain catchment.*

Reply: We thank the reviewer for this thoughtful comment. If the catchment would have exclusively or mainly a snow regime, we would agree that updating of the snow storage would be a more logical choice. However, the catchment does not have an exclusive snow regime, but it has a mixture of regimes (also represented in Figure 4). Moreover, essential to the success of the updating procedure is the availability and quality of data on snow cover, and we consider this investigation to be out of the focus of this paper.

We have used streamflow measurements on the day preceding the forecast issuing day to update the slow and fast runoff reservoirs. This is possible because in the HBV model there is a direct connection between these reservoirs and discharge. Such a direct connection does not exist with the snow storage reservoir. Daily streamflow measurements commonly have a high autocorrelation, so it can be expected that observed streamflow on day *t-1* provides information about the storage in the slow runoff reservoir and fast runoff reservoir on day *t*. We expect that the correlation between snow water equivalent on day *t* and streamflow on day *t-1* will be much lower, and therefore updating of the snow storage using streamflow measurements will be less effective.

Comment: *7, The literature review of ensemble hydrologic forecasting, pre- and post-processing for bias corrections, and data assimilation and model updating, is a good start but seems a little light. Citing more work would provide valuable context to the paper. A reasonable place to start might be recent work by Dominique Bourdin at the University of British Columbia and Hamid Moradkhani at Portland State University.*

Reply 18 March 2017: We thank the reviewer for suggesting these sources of additional relevant literature, especially the work by Moradkhani about pre- and post-processing (Khajehei and Moradkhani, 2017; Madadgar et al., 2014) and updating and data assimilation (e.g. Liu et al., 2012; Pathiraja et al., 2015; Yan and Moradkhani, 2016), and the work by Bourdin which contains recent developments in ensemble streamflow forecasting (Bourdin et al., 2012; Bourdin and Stull, 2013). We will further study the papers by Moradkhani and Bourdin and use this to further extend the context of the paper on ensemble streamflow forecasting, pre- and post-processing and updating procedures.

Reply 8 May 2017: To extend the context of the paper on ensemble streamflow forecasting, pre- and post-processing and updating procedures, we have added references to Bourdin and Stull (2013) and Krzysztofowicz (2001), Bennett et al. (2014), Khajehei and Moradkhani (2017), Verkade et al. (2013), Hashino et al. (2007), Clark et al. (2004), Boé et al. (2007) and Wetterhall et al. (2012), and Houser et al. (2012) and Liu et al. (2012), respectively.

Comment: *8. Some of the specific conclusions seem a little surprising. That's great, but it also means they'd benefit from additional discussion. In particular, the paper concludes in section 4.2.1 that the quality of the forecasts at lead times of less than 3 days is dominated by hydrologic initial conditions, and the weather forecasts become the dominant source of predictive skill after that. This would be a reasonable conclusion for a large or flat basin, but for a small, steep mountain river it seems a little surprising – these are typically flashy systems that respond to rain or snowmelt inputs within a day or so. Indeed, a few pages later near the end of section 5, the paper states that "in the hydrological model the lag time between a rainfall event and the streamflow peak is set to 1 day." It also seems that conclusions like this, which attempt to attribute predictive skill (and therefore also predictive error) to various different sources, might be difficult to make convincingly without using a more statically sophisticated and exhaustive data assimilation procedure, incorporating ensembles of hydrologic models and/or model parameters, etc.*

Reply: We thank the reviewer for this comment and will further explain the observations of the reviewer. Regarding the comment that "the quality of the forecasts at lead times of less than 3 days is dominated by hydrologic initial conditions, and the weather forecasts become the dominant source of predictive skill after that" is "a little surprising": our results show that this depends on the streamflow category and the streamflow generating process. Short-rain generated high streamflows, snowmelt generated high streamflows and snow accumulation generated low flows are skillfully forecasted by the meteorological forecasts after 1 or 2 days, which could be expected for these fast processes and confirms the expectations of the reviewer. For long-rain generated high streamflows, medium streamflows and precipitration deficit generated low streamflows the maximum skill is observed at larger lead times, because for these processes both the forecasts and the alternative forecasts are dominated by the initial conditions at small lead times.

References

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.

Boé, J., Terray, L., Habets, F. and Martin, E.: Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies, Int. J. Climatol., 27(12), 1643–1655, doi:10.1002/joc.1602, 2007.

Bourdin, D. R. and Stull, R. B.: Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow, J. Hydrol., 502, 77–88, doi:10.1016/j.jhydrol.2013.08.028, 2013.

Bourdin, D. R., Fleming, S. W. and Stull, R. B.: Streamflow Modelling: A Primer on Applications, Approaches and Challenges, Atmosphere-Ocean, 50(4), 507–536, doi:10.1080/07055900.2012.734276, 2012.

Bouwer, L. M., Bubeck, P. and Aerts, J. C. J. H.: Changes in future flood risk due to climate and development in a Dutch polder area, Glob. Environ. Chang., 20(3), 463–471, doi:10.1016/j.gloenvcha.2010.04.002, 2010.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R.: The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields, J. Hydrometeorol., 5(1), 243–262, doi:10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2, 2004.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol., 375(3–4), 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

ECMWF: Describing ECMWF's forecasts and forecasting system, edited by B. Riddaway, ECMWF Newsl., 133, 11–13 [online] Available from: http://old.ecmwf.int/publications/newsletters/pdf/133.pdf, 2012.

Fleming, S. W.: Demand modulation of water scarcity sensitivities to secular climatic variation: theoretical insights from a computational maquette, Hydrol. Sci. J., 61(16), 2849–2859, doi:10.1080/02626667.2016.1164316, 2016.

Hashino, T., Bradley, A. A. and Schwartz, S. S.: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, Hydrol. Earth Syst. Sci., 11(2), 939–950, doi:10.5194/hess-11-939-2007, 2007.

Houser, P. R., De Lannoy, G. J. M. and Walker, J. P.: Hydrologic Data Assimilation, in Approaches to Managing Disaster - Assessing Hazards, Emergencies and Disaster Impacts, edited by J. Tiefenbacher, pp. 41–64, InTech., 2012.

IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Core Writing Team, R. K. Pachauri, and L. A. Meyer, IPCC, Geneva, Zwitzerland. [online] Available from: http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full.pdf, 2014.

Khajehei, S. and Moradkhani, H.: Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach, J. Hydrol., 546, 476–489, doi:10.1016/j.jhydrol.2017.01.026, 2017.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249(1–4), 2–9, doi:10.1016/S0022-1694(01)00420-6, 2001.

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H. J., Kumar, S., Moradkhani, H., Seo, D. J., Schwanenberg, D., Smith, P., Van Dijk, A. I. . J. M., Van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O. and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities, Hydrol. Earth Syst. Sci., 16(10), 3863–3887, doi:10.5194/hess-16-3863-2012, 2012.

Madadgar, S., Moradkhani, H. and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrol. Process., 28(1), 104–122, doi:10.1002/hyp.9562, 2014.

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.

Pathiraja, S., Marshall, L., Sharma, A. and Moradkhani, H.: Hydrologic modeling in dynamic catchments: A data assimilation approach, Water Resour. Res., 52(5), 3350–3372, doi:10.1002/2015WR017192, 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.

Rojas, R., Feyen, L. and Watkiss, P.: Climate change and river floods in the European Union: Socio-economic consequences and the costs and benefits of adaptation, Glob. Environ. Chang., 23(6), 1737–1751, doi:10.1016/j.gloenvcha.2013.08.006, 2013.

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

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.

Wetterhall, F., Pappenberger, F., He, Y., Freer, J. and Cloke, H. L.: Conditioning model output statistics of regional climate model precipitation on circulation patterns, Nonlinear Process. Geophys., 19(6), 623–633, doi:10.5194/npg-19-623-2012, 2012.

Wheater, H. S. and Gober, P.: Water security and the science agenda, Water Resour. Res., 51(7), 5406–5424, doi:10.1002/2015WR016892, 2015.

Yan, H. and Moradkhani, H.: Combined assimilation of streamflow and satellite soil moisture with the particle filter and geostatistical modeling, Adv. Water Resour., 94, 364–378, doi:10.1016/j.advwatres.2016.06.002, 2016.

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. These studies~~y~~ 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).~~ However, m~~M~~ost 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 on~~to~~ ~~general~~ non-specific ensemble streamflow forecasting systems (Bennett et al., 2014; Demargne et al., 2010; Renner et al., 2009; Verkade et al., 2013) did not evaluate the performance for different streamflow categories (i.e. for low streamflow and high streamflow events). Moreover, previous studies did not assess effects of runoff processes, like snowmelt and extreme rainfall events, on the performance of ensemble forecasts. The only study we found that touches on this is the study by Roulin and Vannitsem (2005). This study concluded that the developed 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, ~~I~~information 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. The extent~~is 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~~ whereas only 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~~Contrarily, 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~~underdispersion ~~of the spread~~ of ensemble flood forecasts, ~~to an extent that~~which decreases with lead time. ~~They conclude that the~~ 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 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~~ are 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 3: 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.

Performance of ensemble streamflow forecasts under varied hydrometeorological conditions

Harm-Jan F. Benninga[1,*], Martijn J. Booij[1], Renata J. Romanowicz[2], Tom H.M. Rientjes[3]

[1]Water Engineering and Management, Faculty of Engineering Technology, University of Twente, 7500 AE Enschede, The Netherlands
[2]Institute of Geophysics, Polish Academy of Sciences, 01-452 Warsaw, Poland
[3]Water Resources, Faculty of Geo-Information Science and Earth Observation, University of Twente, 7500 AE Enschede, The Netherlands
[*]Present address: Water Resources, Faculty of Geo-Information Science and Earth Observation, University of Twente

Correspondence to: Harm-Jan F. Benninga (h.f.benninga@utwente.nl)

Abstract. The paper presents a methodology to give insight in the performance of ensemble streamflow forecasting systems. We have developed an ensemble forecasting system for the Biała Tarnowska, a mountainous river catchment in southern Poland, and analysed the performance for lead times ranging from 1 day to 10 days for low, medium and high streamflow and ~~related~~ different ~~runoff generating processes~~hydrometeorological conditions. Precipitation and temperature forecasts from the European Centre for Medium-Range Weather Forecasts serve as inputs to a deterministic lumped hydrological (HBV) model. Due to an inconsistent bias, ~~the best streamflow forecasts were obtained without~~ pre- and post-processing of the meteorological and streamflow forecasts is not effective. The b~~B~~est forecast skill, relative to alternative forecasts based on historical meteorological measurements~~ of precipitation and temperature~~, is shown for high streamflow and for snow accumulation low streamflow events. Forecasts of medium streamflow events and low streamflow events generated by a precipitation deficit show less skill. 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,~~ It is recommended to calibrate the hydrological model specifically on low streamflow conditions and high streamflow conditions. ~~the hydrological model should be improved. The study~~ It is further recommend~~ed~~s ~~improving~~ that the ~~reliability~~ dispersion (reliability) of the ensemble streamflow forecasts is enlarged by including the uncertainties in hydrological model parameters and initial conditions, and by ~~improving~~ enlarging the dispersion of the meteorological input forecasts.

1 Introduction

Accurate flood forecasting (Cloke and Pappenberger, 2009; Penning-Rowsell et al., 2000; Werner et al., 2005) and low streamflow forecasting (Demirel et al., 2013a; Fundel et al., 2013) are important to mitigate the negative effects of extreme events by enabling early warning. Accurate forecasting becomes increasingly more important, ~~since~~ because frequency and magnitude of low and high streamflow events are projected to increase in many areas in the world as a result of climate

change (IPCC, 2014). Due to socio-economic development ~~also~~ the impacts of extreme events further increase (Bouwer et al., 2010; Fleming, 2016; Rojas et al., 2013; Wheater and Gober, 2015).

Hydrological forecasting systems are often implemented as ensemble forecasting systems (Cloke and Pappenberger, 2009). Ensemble forecasts provide information about the possibility that an event occurs (Krzysztofowicz, 2001; Thielen et al., 2009), and allow quantification of the forecast uncertainty (Krzysztofowicz, 2001; Zappa et al., 2011). Uncertainties in streamflow forecasts originate from meteorological inputs, and hydrological model parameters, initial conditions and model structure (Bourdin and Stull, 2013; Cloke and Pappenberger, 2009; Demirel et al., 2013a; Zappa et al., 2011).

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. These studies~~y~~ 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).~~ However, m~~M~~ost 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., 2013a; Fundel et al., 2013)~~, .~~ ~~in contrast to~~The studies on~~to~~ ~~general~~ non-specific ensemble streamflow forecasting systems (Bennett et al., 2014; Demargne et al., 2010; Renner et al., 2009; Verkade et al., 2013) did not evaluate the performance for different streamflow categories (i.e. for low streamflow and high streamflow events). Moreover, previous studies did not assess effects of runoff processes, like snowmelt and extreme rainfall events, on the performance of ensemble forecasts. The only study we found that touches on this is the study by Roulin and Vannitsem (2005). This study concluded that the developed 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, ~~I~~information 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. The extent~~is~~ ~~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~~ whereas only 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~~Contrarily, 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~~underdispersion ~~of the spread~~ of ensemble flood forecasts, ~~to an extent that~~which decreases with lead time. ~~They conclude that the~~ 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. (2013a) concluded that uncertainty of hydrological model parameters has the largest effect, whereas 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~~ are more important.

The objective of this study is to investigate the performance and limitations of ECMWF meteorological forecasts based ensemble streamflow forecasting for lead times up to 10 days for low, medium and high streamflow in a catchment with seasonal variation in runoff generating processes. We aim to evaluate whether performance of the forecasting system can be related to specific runoff generating processes based on hydrometeorological conditions. Further, we assess whether the main source of forecast error relates to the meteorological inputs or to deficiencies of the hydrological model, for the different streamflow categories and runoff generating processes.

2 Study catchment and data

2.1 Study area and measurement data

The Biała Tarnowska catchment in Poland serves as study area. This catchment is selected because of its large variation in streamflow~~, with~~ and seasonal variation in runoff generating processes. The catchment (Fig. 1) is located in a mountainous part of southern Poland. Napiorkowski et al. (2014) further describe the catchment. The ~~River~~ Biała Tarnowska River discharges into the ~~River~~ Dunajec River, which is a tributary of the ~~River~~ Vistula River. The length of the river is 101.8 km with a catchment area of 956.9 km^2. The mean streamflow ~~discharge (1972–2013)~~ is 9.4 m^3 s^{-1} (1972–2013). ~~Streamflow is characterized by large variation and extreme high flows with highest measured streamflow of 611 m^3 s^{-1}.~~ The highest measured streamflow is 611 m^3 s^{-1}. During winter and spring snow(melt) plays an important role. Comparison of the time series of precipitation and streamflow ~~reveals~~ shows that the lag time between intense precipitation events and related peaks in streamflow varies between 1 and 3 days.

Precipitation and~~,~~ temperature ~~and streamflow~~ measurement series are available from five meteorological stations and streamflow measurement series are available from one discharge gauging station, at a daily time interval for the period 1

January 1971 to 31 October 2013, and 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 meteorological 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~~gradient 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~~ small to accurately determine precipitation and temperature gradients, the calculated precipitation gradients are used because of the clear difference between the 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 measurements from each station are corrected for the difference between the elevation of the station and the mean elevation its respective Thiessen polygon. To represent catchment averages, the corrected measurements are weighted based on the relative coverage of their Thiessen polygon (Fig. 1). 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.

2.2 Meteorological forecast data

The meteorological ensemble forecasts ~~data from~~ by ECMWF are used, because of the good performance compared to other meteorological ensemble forecast sets (Buizza et al., 2005; Tao et al., 2014) and because the~~se~~ ECMWF forecasts are frequently used in hydrological ensemble forecasting (Cloke and Pappenberger, 2009). Persson and Andersson (2013) and ECMWF (2012) describe how ECMWF generates the meteorological ensemble forecasts. The ensemble forecasts consist of one control forecast (no perturbation) and 50 ensemble members. The ensemble members should represent initial condition and meteorological model uncertainty (Leutbecher and Palmer, 2008; Persson and Andersson, 2013).

The THORPEX Interactive Grand Global Ensemble (TIGGE) project, developed by The Observing System Research and Predictability Experiment (THORPEX), provides historical meteorological forecast data from 1 October 2006 onwards (Bougeault et al., 2010). The resolution of the ensemble and control forecasts is 32 km × 32 km (ECMWF, 2012). Using the TIGGE data portal we interpolated the forecasts to a regular grid (Bougeault et al., 2010) with a resolution of 0.25° × 0.25° (~17.9 km × 27.8 km at this latitude). In this study a maximum lead time of 10 days is used, following the World Meteorological Organization (WMO) that defines medium-range as forecasts with lead times from 3 days to 10 days (ECMWF, 2012). We also refer to Alfieri et al. (2014), Bennett et al. (2014), Demirel et al. (2013a), Olsson and Lindström (2008), Renner et al. (2009), Roulin and Vannitsem (2005) and Verkade et al. (2013) that use 9 or 10 days as maximum lead time for hydrological forecasting. Because we use a lumped hydrological model with a daily time step (Sect. 3.1.1), we averaged daily ECMWF forecasts according to the relative area coverage of the seven grid cells that overlay the catchment.

According to Persson and Andersson (2013) ECMWF forecasts may apply to a land elevation that significantly differs from the actual elevation in a grid and this can lead to biases. ~~In this study c~~Correction for such elevation errors is ignored, ~~since~~ because any systematic bias is accounted for in the pre-processing step (Sect. 3.1.3). ECMWF provides temperature forecasts at 00:00 hr. or 12:00 hr. This means that temperature forecasts cannot be considered as representative for one day. To obtain representative daily average temperature forecasts, we weight the temperature forecasts at 00:00 hr., 12:00 hr and 24:00 hr by 25%, 50% and 25% respectively.

3 Method~~sology~~

3.1 The ensemble streamflow forecasting system

The ensemble streamflow forecasting system consists of multiple components, ~~presented~~ shown in Fig. 2. Uncertainties in meteorological forecasts, model parameters, model initial conditions and model structure affect ~~ensemble~~ streamflow forecasts (Bourdin and Stull, 2013; Cloke and Pappenberger, 2009; Demirel et al., 2013a; Zappa et al., 2011). To capture the full range of predictive uncertainty, uncertainties arising from all sources of error must be incorporated (Bourdin and Stull, 2013; Krzysztofowicz, 2001; Zappa et al., 2011). Bennett et al. (2014) and Cloke and Pappenberger (2009) describe that uncertainties in meteorological forecasts are the largest source of uncertainty beyond 2–3 days, and therefore only meteorological forecast uncertainty is incorporated in many studies (Bennett et al., 2014). We only include uncertainty in the meteorological forecasts to focus on the effect of ensemble meteorological forecasts on streamflow forecasts. As a consequence, underdispersion of the streamflow forecasts may be expected. ~~By considering only uncertainty of the meteorological forecasts we focus on the effect of ensemble meteorological forecasts on streamflow forecasts.~~

3.1.1 Hydrological model

The hydrological model we use is a lumped Hydrologiska Byråns Vattenbalansavdelning (HBV) model that we run at daily time step ~~by available hydrometeorological time series data for streamflow, gauged precipitation and temperature, and ECMWF meteorological forecasts~~. The model has 14 parameters and includes a snow accumulation and melting routine (Lindström et al., 1997; Osuch et al., 2015). Daily potential evapotranspiration rates are based on air temperature ~~following~~ using the method of Hamon (Lu et al., 2005). The HBV model has wide application in studies on ensemble streamflow forecasting (e.g. Cloke & Pappenberger, 2009; Demirel et al., 2013a, 2015; Kiczko et al., 2015; Olsson & Lindström, 2008; Renner et al., 2009; Verkade et al., 2013). The choice for a lumped model with a daily time step is ~~basically~~ the result of the spatial and temporal resolution of the available data. The available measurements of precipitation and temperature from five meteorological stations and streamflow from one discharge gauging station do not justify application of a spatially distributed hydrological model. The River Rhine forecasting suite also adopts the HBV model at a daily time step ~~that is applied~~ as a semi-distributed model to 134 sub catchments (Renner et al., 2009). The catchment area of Biała Tarnowska (~1000 km^2) is comparable to the area of the sub catchments in the River Rhine forecasting suite.

~~The HBV model is calibrated using the differential evolution with global and local neighbourhoods (DEGL) method~~To calibrate the HBV model we used Differential Evolution with Global and Local neighbourhoods (DEGL), described by Das et al. (2009). The settings ~~that we used~~ are adopted from the best performing variant of Das et al. (2009) (maximum number of model runs is set to 50000). ~~The model is calibrated over the period 1 November 1971 to 31 October 2000 with the time series of precipitation and temperature as input and streamflow measurements as reference output. The validation period is 1 November 2000 to 31 October 2013.~~ The model parameters were drawn uniformly from predefined parameter ranges (Osuch et al., 2015). The objective function selected for calibration is *Y*, which combines the Nash–Sutcliffe coefficient (NS) and the relative volume error (E_{RV}) (Akhtar et al., 2009; Rientjes et al., 2013). According to Rientjes et al. (2013), values of *Y* below 0.6 indicate poor to satisfactory performance. The model is calibrated over the period 1 November 1971 to 31 October 2000, with the time series of precipitation and temperature as inputs and streamflow measurements as reference output. The validation period is 1 November 2000 to 31 October 2013. Initialization periods of 10 months and 1 year respectively ensure realistic initial conditions at the first day of the calibration and the validation period. ~~The model parameters were drawn uniformly from predefined parameter ranges (Osuch et al., 2015).~~

3.1.2 Updating of initial states

~~Hydrological forecasting often relies on the updating of hydrological model storages to best represent the hydrological conditions in the catchment at the forecast day (e.g. .~~ To best represent the hydrological conditions in the catchment at the forecast issuing day, hydrological forecasting system often rely on the updating of hydrological model storages by combining simulations with real-time data (Demirel et al., 2013a; Liu et al., 2012; Werner et al., 2005; Wöhling et al., 2006). A number of sophisticated techniques have been developed for data assimilation and model state updating (Houser et al., 2012; Liu et al., 2012). We apply the fairly simple and direct storage updating procedure introduced by Demirel et al. (2013a), which relies on the autocorrelation of streamflow to update model storages. ~~For storage updating we follow Demirel et al. (2013a) and apply a procedure based on measured streamflow on the day preceding the forecast day.~~ The measured streamflow of the day preceding the forecast issuing day is divided in a fast and a slow runoff component to update the fast runoff reservoir and the slow runoff reservoir ~~in~~ of the HBV model. To determine the ratio between the fast and slow runoff components, a relationship between total simulated streamflow and the fraction of fast runoff is established based on historical simulations. ~~However, this relationship contains large uncertainty. For example, for a total simulated streamflow of 10 $m^3 \cdot s^{-1}$ the fraction varies between 0 and 0.6 and for a streamflow of 20 $m^3 \cdot s^{-1}$ it varies between 0.3 and 0.7. To reduce uncertainty in the fraction of fast runoff the storage of the fast runoff HBV reservoir and net inflow in the fast runoff reservoir are both tested as an additional descriptor of the relationship between streamflow and the fraction of fast runoff.~~

3.1.3 Pre- and post-processing

Errors in the meteorological forecasts ~~as well as~~ and in the hydrological models introduce biases in the mean and errors in the spread of ensemble streamflow forecasts (Cloke and Pappenberger, 2009; Khajehei and Moradkhani, 2017; Verkade et

al., 2013). Several studies suggest that post-processing of streamflow forecasts is more effective to improve the forecast ~~skill~~ quality than pre-processing of meteorological input data (Kang et al., 2010; Verkade et al., 2013; Zalachori et al., 2012). Verkade et al. (2013) and Zalachori et al. (2012) found that corrections made to meteorological forecasts lose their effect when propagated through a hydrological model ~~(Verkade et al., 2013; Zalachori et al., 2012)~~. Results by Zalachori et al. (2012) indicate that combined pre- and post-processing results in the best forecast quality. In this study both pre-processing of the meteorological input forecasts and post-processing of the streamflow forecasts are tested.

Many studies used (conditional) quantile mapping (QM) for pre-processing (Boé et al., 2007; Déqué, 2007; Kang et al., 2010; Kiczko et al., 2015; Verkade et al., 2013; Wetterhall et al., 2012) and post-processing (Hashino et al., 2007; Kang et al., 2010; Madadgar et al., 2014; Shi et al., 2008) to correct for bias and dispersion errors. According to Kang et al. (2010), QM generally performs well in both pre- and post-processing. Hashino et al. (2007) advise to use QM, because of the good performance regarding sharpness and discrimination and the simplicity of the method. ~~The principle of QM is that~~With QM the cumulative distribution function (CDF) of the forecasts over a ~~control~~ training period is matched to the CDF of the measurements over the same period, after which a correction function is generated (Boé et al., 2007). This means that the correction is conditional on the value of the forecasted variable itself. Boé et al. (2007), Déqué (2007) and Madadgar et al. (2014) further explain QM. The empirical CDFs of the measurements and forecasts are established on the training period 1 November 2011 to 31 October 2013 (two hydrological years) and validated ~~over~~ on the period 1 November 2007 to 31 October 2011.

Distributions ~~can~~ may be different for different lead times and weather patterns or seasons (Boé et al., 2007; Wetterhall et al., 2012), so three QM set-ups are tested with ~~or~~ and without distinguishing ~~different~~ lead times and seasons. Combining the options for pre-processing and post-processing results in four processing strategies. In strategy 0, no pre- and post-processing are applied. In strategy 1 and 2, QM is applied to pre-process the meteorological forecasts, ~~respectively~~ without post-processing and with post-processing respectively. In strategy 2, the post-processing is performed based on the correction between '~~perfect forecasts~~observed meteorological input forecasts' (streamflow simulations with inputs from meteorological measurements) and streamflow measurements to account for hydrological model uncertainties (Verkade et al., 2013). In strategy 3, only post-processing is applied, based on the correction between streamflow forecasts generated with uncorrected meteorological forecasts and measured streamflow. In this strategy meteorological and hydrological model uncertainties are treated together (Verkade et al., 2013).

3.2 Evaluation scores of the ensemble forecasts

To measure the overall performance, we employ the frequently-used Continuous Ranked Probability Score (CRPS) (Bennett et al., 2014; Demargne et al., 2010; Hamill et al., 2000; Hersbach, 2000; Khajehei and Moradkhani, 2017; Pappenberger et al., 2015; Velázquez et al., 2010; Verkade et al., 2013). To evaluate forecast skill, we use the Continuous Ranked Probability Skill Score (CRPSS), which is the CRPS of the forecasts relative to the CRPS of alternative forecasts. The alternative forecast set is selected in Sect. 3.2.1. ~~To measure general quality and skill of the streamflow forecasts, the continuous ranked~~

probability score (CRPS) and the continuous ranked probability skill score (CRPSS) are used. According to Demargne et al. (2010) and Hamill et al. (2000) a single evaluation score is inadequate to evaluate the overall performance of a forecasting system. Three properties of forecast quality are reliability, sharpness and resolution (Wilks, 2006; WMO, 2015).

Reliability refers to the statistical consistency between measurements and simulations (Candille & Talagrand, 2005; Velázquez et al., 2010) and whether uncertainty is correctly represented in the forecasts (Bennett et al., 2014). We evaluate reliability by rank histograms (Sect. 3.2.2) and reliability diagrams (Bröcker and Smith, 2007; Ranjan, 2009; Wilks, 2006; WMO, 2015). The five forecast probability bins that we use to establish the reliability diagrams are 0%–20%, 20%–40%, … and 80%–100%, which were also used by Demirel et al. (2013a) and Bennett et al. (2014), and the low streamflow and high streamflow thresholds considered are defined in Sect. 3.4.

Sharpness is defined as the tendency to forecast probabilities of occurrence near 0 or 1, as opposed to values clustered around the mean (climatological) probability (Ranjan, 2009; Wilks, 2006; WMO, 2015). If an ensemble forecasting system always forecasts an event probability of occurrence close to climatological probability, instead of close to 0 or close to 1, this the forecasting system is not useful, although it might be well calibrated (Ranjan, 2009; Wilks, 2006). To evaluate sharpness, we employ the histograms that show the sample size of the forecast probability bins used to establish the reliability diagrams (Ranjan, 2009; Renner et al., 2009; WMO, 2015). The histograms accompanying reliability diagrams are used to evaluate sharpness.

Resolution is the ability of the forecast model to correctly forecast the occurrence or and nonoccurrence of events (Demirel et al., 2013a; Martina et al., 2006). We employ relative Relative operating Operating Ceharacteristics (ROC) curves to evaluate resolution (Fawcett, 2006; Khajehei and Moradkhani, 2017; Velázquez et al., 2010; Wilks, 2006; WMO, 2015). The Area Under the ROC Curve (AUC) provides a single score of performance regarding resolution (Fawcett, 2006; Wilks, 2006). A perfect ensemble forecasting system has an area of 1 under the ROC curve (100% hit rate, 0% false alarm rate for all probability thresholds), while a forecasting system with zero skill has a diagonal ROC curve with an area of 0.5 (coincides with diagonal) (Fawcett, 2006; Velázquez et al., 2010; WMO, 2015).

3.2.1 Continuous ranked probability score

The CRPS is an overall, single-number score for judging the quality of probabilistic forecasts (Hamill et al., 2000). CRPS measures the error of the ensemble forecasts by integrating the squared distance between the CDFs of the forecasts and a reference streamflow (Bennett et al., 2014; Demargne et al., 2010; Verkade et al., 2013). The score is frequently used in atmospheric (Velázquez et al., 2010) and hydrological sciences (Bennett et al., 2014; Pappenberger et al., 2015; Velázquez et al., 2010) and in most cases it is the recommended evaluation score for ensemble forecasts (Pappenberger et al., 2015). CRPS is sensitive to the entire range of the variable of interest and it does not require the introduction of predefined classes (Hersbach, 2000). A CRPS of 0 indicates a perfect simulation, which can only be achieved in the case of a perfect deterministic forecast (Hersbach, 2000). Because in practice CRPS approaches the average value of the evaluated variable

~~(with the same unit), the score cannot directly be compared among different areas, seasons or streamflow categories (Ye et al., 2014). Comparison between different lead times is possible, as average streamflow values do not change with lead time.~~

3.2.~~2~~ 1 ~~Continuous ranked probability skill score~~Alternative forecast set

Because in practice the CRPS converges to the average value of the evaluated variable (with the same unit), the score cannot be compared among different areas, seasons or streamflow categories (Ye et al., 2014). 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. ~~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).~~ The CRPSS is defined as:

$$CRPSS = 1 - \frac{CRPS_{forecasts}}{CRPS_{alternative}}, \quad (1)$$

A system with perfect skill results in a CRPSS of 1 and a negative CRPSS indicates that the forecasting system performs worse than the alternative forecasts (Demargne et al., 2010; Ye et al., 2014). ~~To evaluate skill of the forecasting system we define the alternative forecast set as forecasts that are generated without using meteorological forecasts.~~ Forecasts that are generated without meteorological forecasts provide the alternative forecast set. ~~It is common practice to apply hydrological persistency or hydrological climatology as alternative forecast set (Bennett et al., 2014). However, Pappenberger et al. (2015) argue that this can result in an overestimation of forecast skill because other alternative forecast sets might be more difficult to beat in performance. Following Bennett et al. (2013), Bennett et al. (2014) and Pappenberger et al. (2015) the most appropriate alternative forecast set is selected based on their CRPS results. We use a single alternative forecast set for all streamflow categories, so one $CRPS_{alternative}$ is calculated.~~ It is common practice to apply hydrological persistency or hydrological climatology as alternative forecast set (Bennett et al., 2014). With hydrological persistency the most recent streamflow measurement available (i.e., from the day preceding the forecast issuing day) serves as forecast for all lead times. Regarding hydrological climatology, the average measured streamflow, after a smoothing window of 31 days, on the same calendar day over the last 20 years is used, following Bennett et al. (2013). For streamflow forecasts based on an ensemble of historical meteorological measurements ~~of precipitation and temperature~~, measurements on the same calendar day over the past 20 years are used, after Pappenberger et al. (2015).

The alternative forecast set with the lowest CRPS will serve as alternative forecast set to evaluate skill (Bennett et al., 2013, 2014; Pappenberger et al., 2015). We use a single alternative forecast set for all streamflow categories, so one $CRPS_{alternative}$ is calculated. The results in Fig. 3 ~~indicate~~ show that the forecasts based on meteorological climatology result in the best CRPS scores and thus imply to be the most appropriate alternative streamflow forecasts, as also found in the studies of the Bennett et al. (2013), Bennett et al. (2014) and Pappenberger et al. (2015).

3.2.~~.3~~ 2 Rank histogram

~~Rank histograms enable to diagnose average errors in the mean and spread (under- or overdispersion) of the ensemble forecasts (Hamill, 2001; Hamill et al., 2000) and according to Wilks (2006) they are commonly used to evaluate the reliability (or consistency) of ensemble forecasts.~~ The consistency condition states that the reference streamflow is just one more member of the ensemble and it should be statistically indistinguishable from the ensemble forecast (Wilks, 2006). ~~To construct a rank histogram, the reference streamflow is added to the ensemble forecast set and the histogram is constructed from the ranks of the reference streamflow (Velázquez et al., 2010).~~ In an ensemble forecasting system with perfect spread each ensemble member is equally likely, so all reference streamflow ranks are equally likely and the rank histogram is uniform (Hamill, 2001; Hersbach, 2000; Wilks, 2006; WMO, 2015; Zalachori et al., 2012). For backgrounds of the rank histogram, readers are referred to Hamill (2001), Wilks (2006), Velázquez et al. (2010), WMO (2015) and Zalachori et al. (2012). ~~To indicate the flatness of rank histograms Candille and Talagrand (2005) propose a numerical indicator δ. Because δ is proportional to the length of the time series (Velázquez et al., 2010), we~~ We use the Mean Absolute Error as flatness coefficient ε of the rank histogram, with the uniform distribution as reference:~~:~~

$$\varepsilon = \frac{1}{n+1}\sum_{z=1}^{z=n+1}|f(z) - y| \,, \qquad (2)$$

$f(z) =$ Relative frequency of reference streamflow in rank z [-]

$y = \frac{1}{n+1} =$ Theoretical relative frequency (uniform distribution) [-]

$n =$ Number of ensemble members [-]

~~In a perfectly consistent forecasting system the relative frequency in each rank is equal to the relative frequency according to uniform distribution. This gives an optimum value of ε equal~~ The rank histogram and flatness coefficient contain a random element if multiple ensemble members and the measurement have the same value, like 0 mm precipitation (Hamill and Colucci, 1998). In this case, a random rank is assigned to the measurement from the pool of ensemble members and the measurement that have the same value.

~~3.2.4 Reliability diagram~~

~~The reliability diagram is a common way to summarize and evaluate reliability of probabilistic forecasting systems (Bröcker and Smith, 2007). The diagram plots observed relative frequency against the predicted probability for a certain event (Bröcker and Smith, 2007; Demirel et al., 2013a). For a well calibrated forecasting system the reliability diagram is close to the 1:1 diagonal (Ranjan, 2009; WMO, 2015). The five forecast probability bins that we use are 0%–20%, 20%–40%, … and 80%–100%, which were also used by Demirel et al. (2013a) and Bennett et al. (2014). Following Bröcker and Smith (2007) the observed frequencies are plotted against the average of forecast probabilities per bin instead of the bin centre %. Plotting against bin centres (so 10%, 30%, etc.) can cause substantial deviations from the diagonal.~~

~~The histogram showing sample size in each probability bin indicates the sharpness of forecasts (Ranjan, 2009; Renner et al., 2009; WMO, 2015).~~

~~3.2.5 Relative operating characteristic~~

~~Contingency tables and ROC curves analyze whether the forecast model correctly forecasts the occurrence and nonoccurrence of events. To establish the ROC a set of contingency tables is made, one for each examined probability threshold and these form a hit rate/false alarm rate graph for one predefined flow threshold (Atger, 2001; Buizza et al., 1999; Fawcett, 2006; WMO, 2015). The area under the ROC curve (AUC) can be used to obtain a single score for performance (Fawcett, 2006; Wilks, 2006). A perfect ensemble forecasting system has an area of 1 under the ROC curve (100% hit rate, 0% false alarm rate for all probability thresholds), while a forecasting system with zero skill has a diagonal ROC curve with an area of 0.5 (coincides with diagonal) (Fawcett, 2006; Velázquez et al., 2010; WMO, 2015). Buizza et al. (1999) state that it is common practice to consider an area of more than 0.7 as indicative for useful prediction systems and 0.8 for good prediction systems.~~

3.3 ~~Investigation of error contributors~~Contribution of error sources

The evaluation of ensemble streamflow forecasts is affected by errors from the meteorological forecasts, the hydrological model (including errors in the initial conditions) and errors in the measurements that serve as reference streamflow (Renner et al., 2009). By evaluation against ~~perfect forecasts~~observed meteorological input forecasts, the streamflow measurement error and the hydrological model error are eliminated, because both the ensemble streamflow forecasts and the reference streamflows contain these errors (Demargne et al., 2010; Olsson and Lindström, 2008; Renner et al., 2009). If we neglect measurement errors, evaluation against streamflow measurements ($CRPS_{meas}$) contains errors from the meteorological forecasts and the hydrological model and evaluation against ~~perfect streamflow~~observed meteorological input forecasts ($CRPS_{sim}$) exclusively contains errors from the meteorological forecasts (Demargne et al., 2010; Olsson and Lindström, 2008; Renner et al., 2009). If the ratio in Eq. (3) is low, the hydrological model errors are dominant, and if this ratio is high, the meteorological forecast errors are dominant.

$$\frac{CRPS_{sim}}{CRPS_{meas}} \sim \frac{meteorological\ forecast\ errors}{meteorological\ forecast\ errors + hydrological\ model\ errors}, \quad (3)$$

~~A low $CRPS_{sim}$/$CRPS_{meas}$ ratio means that the hydrological model errors are dominant and a high ratio means that the meteorological errors are dominant.~~

3.4 Evaluation of streamflow categories

We evaluate the forecasting system for the different streamflow categories ~~as~~ that are defined in Table 1. A low streamflow threshold of Q_{75} (exceedance probability of 75%) guarantees that a sufficient number of events are considered in the evaluation of this streamflow category, while streamflow below this threshold still affects river functions (Demirel et al., 2013b). Similarly, we use~~d~~ Q_{25} as high streamflow threshold.

3.5 Evaluation of runoff generating processes

The high streamflow forecasts and low streamflow forecasts are evaluated for the ~~various hydrometeorological conditions~~specific runoff processes that can generate these events, based on hydrometeorological conditions. Medium flows are not evaluated for different runoff generating processes ~~since~~ because these events commonly result from a combination of runoff generating processes under non-extreme hydrometeorological conditions.

3.5.1 High streamflow generating processes

Various runoff ~~contributing~~ generating 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 used in the HBV model. ~~The classification rules are based on fluxes and storages at one day before the event, because in the HBV model it takes one modelling time step before the rainfall and snowmelt fluxes end up in the fast runoff and slow runoff reservoirs and can form runoff.~~

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

3.5.2 Low streamflow generating processes

Processes that result in low flows are snow accumulation and the combination of low rainfall and high evapotranspiration over a period (precipitation deficit). Table 3 further characterizes and defines these processes.

Figure 4b shows that these ~~classification~~ rules for classification result in a distribution of low streamflow generating processes over the year that is typical for this region.

~~result in a reliable distribution of low streamflow generating processes over the year for this region.~~

4 Results

4.1 Ensemble streamflow forecasting system

4.1.1 Calibration and validation of the hydrological model

~~In~~ Table 4 lists the calibration and validation results. ~~The hydrological model performs better with corrected input data as compared to uncorrected input data. This implies that the systematic underestimation of precipitation and systematic overestimation of temperature (Sect. 2.1) are not fully captured in the calibration.~~ The validation performance is satisfactory, indicating that the lumped model approach is plausible in this case. The updating of initial states of the fast runoff reservoir and slow runoff reservoir (Sect. 3.1.2) results in an improvement of Y from 0.75 to 0.82 over the validation period. This

effect decreases with lead time, but it is still noticeable at a lead time of 10 days. ~~Relating the fraction of fast runoff additionally to the storage of the fast runoff reservoir storage or net inflow does not result in a significant improvement of *Y* compared to the original updating model. Therefore the original updating model, introduced by Demirel et al. (2013a), is used.~~

Simultaneous measurements and ECMWF forecasts are available ~~over~~ for the period 1 November 2006 to 31 October 2013. In the hydrological year 2007 (1 November 2006 to 31 October 2007) the agreement between streamflow measurements and simulations is poor. Also with another model (~~data~~ Data ~~based~~ Based ~~mechanistic~~ Mechanistic methodology (DBM)), ~~with the same measurement data~~ the performance was worse during this year (Kiczko et al., 2015). This ~~is~~ must be the result of measurement errors and/or human influence, because it is unlikely that in this period different hydrological processes are taking place that are not captured well by both the HBV model and the DBM model. Therefore, the period 1 November 2006 to 31 October 2007 is excluded from the evaluation period.

Table 5 presents the performance of the hydrological model for different lead times and streamflow categories, including the Relative Mean Absolute Error (E_{RMA}). The NS values for the low and medium streamflow categories are negative, meaning that the averages of streamflow measurements in these categories are a better approximation of the measurements than the simulations. ~~All measures~~ The scores highlight that the calibration is skewed to high streamflow ~~situations~~conditions, which is the result of the selected objective function that includes NS (Gupta et al., 2009). Gupta et al. (2009) also found that model calibration with NS tends to underestimate the low and high streamflow peaks.

The results in Table 5 improve considerably as a result of the updating of initial storages, especially for the low streamflow simulations. The effectiveness of the updating procedure depends on the autocorrelation of daily streamflow~~, because the updating is based on streamflow measurements of the preceding day~~. In low streamflow periods there is usually a high autocorrelation of daily streamflow, in contrast to high streamflow periods.

4.1.2 Pre- and post-processing strategy results

The best precipitation forecasts are obtained ~~when~~ if QM is applied separately to each lead time, whereas the best temperature forecasts are obtained if, in addition, separate relationships for the summer and winter season are applied. The CRPS and ~~Relative Mean Absolute Error (E_{RMA})~~ of the precipitation and temperature forecasts improve slightly and the flatness coefficients improve considerably as a result of the pre-processing.

However, for ~~Regarding~~ the combined pre- and post-processing strategies, ~~the results (not shown in the paper)~~the results in Fig. 5 show ~~indicate~~ that strategy 0 (no pre- and post-processing) results in the best CRPS ~~and flatness coefficients of streamflow simulations~~.

4.2 Forecast performance

4.2.1 Forecast skill

The streamflow forecasts are evaluated over the period 1 November 2007 to 31 October 2013, for lead times from 1 day to 10 days and for the different streamflow categories (defined in Table 1). The results are ~~presented~~ shown in Fig. ~~5~~6. The CRPS increases with lead time for all streamflow categories (Fig. ~~5a~~6a), so the performance of the streamflow ~~forecasts~~ forecasting system deteriorates with lead time. For all streamflows ~~together~~ aggregated, the CRPSS is positive for all lead times (Fig. ~~5b~~6b), so on average the streamflow forecasts are better than the alternative forecasts. This forecast skill is generated by the ECMWF forecasts compared to historical meteorological measurements.

Fig. ~~5b~~ 6b shows that the forecast skill is very different for the low, medium and high streamflow forecasts. The low skill of low streamflow forecasts, especially for small lead times, can be explained by the important role of the initial hydrological conditions ~~in the hydrological model~~. In low streamflow situations runoff is mainly generated by available resources in the catchment instead of precipitation input. Since the same initial model conditions are used to simulate the alternative forecasts, ~~it is difficult to generate skilful~~ the result is that low streamflow forecasts cannot skilfully be forecasted for small lead times (<3 days). Also the origin of the alternative forecasts plays a role. Since low streamflow events normally occur in the same period of the year due to climatic seasonality, it can be expected that historical meteorological measurements ~~of precipitation and temperature~~ on the same calendar day provide ~~functional~~ plausible inputs. After all, the performance of the meteorological forecasts preceding these events contributes to the low skill. The negative skill at small lead times indicates that historical meteorological measurements ~~of precipitation and temperature~~ are even better forecasts than the meteorological ensemble forecasts ~~from~~ by ECMWF for this category of flows. From a lead time of 3 days the accumulated ~~effects of the~~ meteorological forecasts are more skilful than historical meteorological measurements.

The medium streamflow forecasts do not have clear positive skill for all lead times. This can be explained by the fact that ~~historical~~ streamflow ~~measurements are~~ is most often ~~around~~ close to the medium streamflow, so forecasts based on historical meteorological measurements ~~of precipitation and temperature~~ will be a good approximation for these flows.

The system has a high positive skill in forecasting high streamflow. In general initial conditions are relatively less important ~~in~~ for these events, because of the amount of water usually added to the system. However, we note that this depends on the responsible runoff generating process (see results in Sect. 4.4.1). As a result the streamflow forecasts and reference forecasts can easier deviate. In addition, ~~these~~ high streamflow events will be~~are~~ less well captured ~~in~~ by historical meteorological 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.~~

4.2.2 Forecast quality

Figure~~.~~ ~~6~~ 7 ~~presents~~ shows the flatness coefficients. The high values indicate that the rank histograms are far from flat, especially for small lead times and low streamflow events. The rank histograms (in supplement Fig. S1) are U-shaped, which indicates underdispersion and/or conditional bias in the streamflow forecasts (Hamill, 2001). ~~The rank histograms of the meteorological forecasts show that t~~The ECMWF forecasts are also underdispersed, so this is one cause why the streamflow forecasts are underdispersed. In Sect. 5 the consequences of ~~neglecting~~ ignoring uncertainties in the hydrological model and initial conditions are further discussed.

The rank histograms of the different streamflow categories (Fig. S2) show that the streamflow forecasts contain a conditional bias. In general, high streamflow is underestimated by the forecasting system and this ~~underestimation~~ increases with lead time. ~~On the other hand, low~~ Low streamflow is generally overestimated. ~~This~~ Both observations can be the result of too coarse spatial and temporal model resolution. Using a lumped model and aggregating the meteorological inputs spatially over the catchment and temporarily over the day flattens the extreme flow events.

Also the reliability diagrams (Fig. S3) indicate low reliability of the streamflow forecasts, especially for small lead times. It appears that for low streamflow forecasts the observed relative frequencies are underestimated. Regarding the high streamflow forecasts the observed relative frequencies are overestimated, ~~although~~ whereas the rank histograms indicate that high streamflow is underestimated. This is possible because in ~~a~~ the rank histogram the measurements and forecasts are compared directly, whereas in ~~a~~ the reliability diagram the measurements and forecasts are compared to a streamflow threshold.

Histograms showing the sample size in each probability bin of the reliability diagrams indicate that the sharpness of the forecasts is good, because forecast probabilities of low and high streamflow are most often close to 0 or 1, instead of forecast probabilities close to the mean probability. The sharpness decreases with lead time.

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 forecasting system. Buizza et al. (1999) state that, for meteorological forecast systems, it is common practice to consider an area of more than 0.7 as indicative for useful prediction systems and 0.8 for good prediction systems. The ROC curves are included in Fig. S4.

4.3 Dominant error contributors

Figure~~.~~ ~~7~~ 8 shows that the relative contribution of meteorological forecast errors increases and the relative contribution of hydrological model errors decreases with lead time, although the performance of the hydrological model also deteriorates with lead time (~~see~~ Table 5). Two effects contribute to this. ~~In the first place~~First, the meteorological forecasts get worse with lead time (Fig. 5) and errors in the meteorological forecasts accumulate in the hydrological forecasting system ~~with lead time~~. ~~In the second place~~Second, the effect of the initial hydrological conditions ~~in the hydrological model~~ at the forecast issuing day becomes smaller at larger lead times, because more water is added to the system.

~~In~~ For high streamflow forecasts the contribution of meteorological forecast errors is relatively more important, while in low streamflow forecasts the contribution of hydrological model errors is relatively more important. Initial conditions have relatively less influence on high streamflow (discussed in Sect. 4.2.1). In addition, the hydrological model performs better for high streamflow than for low streamflow ~~situations~~ conditions (Table 5), so meteorological forecast errors are relatively more important in high streamflow ~~situations~~conditions.

4.4 Forecast skill for the runoff generating processes

4.4.1 High streamflow generating processes

The highest skill is obtained for short-rain floods (Fig. ~~8a~~9a), at small lead times (1-5 days). Two effects ~~explain this observation~~contribute to this. First, long-rain floods and snowmelt floods are essentially driven by the water storage conditions in the catchment whereas ~~in~~ for short-rain floods meteorological input has more influence. Figure ~~8b~~ 9b confirms the relative importance of meteorological forecasts in these events. This results in higher potential to generate forecast skill, already at small lead times. At larger lead times the accumulation of rainfall in the forecasting system becomes important, which is confirmed by the increasing contribution of meteorological forecast errors in long-rain floods and snowmelt floods. ~~Long-rain floods are skilfully forecasted from a lead time of 3 days and snowmelt floods are skilfully forecasted from a lead time of 2 days.~~ Second, the short and heavy rain events preceding short-rain floods ~~are~~ will be less well captured in historical meteorological measurements than the longer term processes ~~underlying~~ generating long-rain floods and snowmelt floods. Long-rain floods are skilfully forecasted from a lead time of 3 days and snowmelt floods are skilfully forecasted from a lead time of 2 days. ~~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.~~ The forecast skills of short-rain floods and snowmelt floods decrease considerably ~~again at larger lead times~~ from lead times of 6 days and 9 days respectively. This is the result of a decreased performance of the meteorological forecasts preceding these events. The skill of short-rain flood forecasts decreases the most ~~and at the shortest lead time~~.

4.4.2 Low streamflow generating processes

Figure ~~9a~~ 10a shows that the low forecast skill of low streamflow is caused by the precipitation deficit process, whereas the forecast skill of low streamflow events that are generated by snow accumulation is rather high. The low forecast skill of the low streamflow events generated by precipitation deficit ~~precipitation deficit generated low streamflow events~~ can be explained by the fact that low rainfall periods often occur in the same period of the year, due to climatic seasonality, and are therefore well captured by historical meteorological measurements. Also the performance of meteorological forecast models may play a role. Meteorological models tend to forecast drizzle instead of zero precipitation (Boé et al., 2007; Piani et al., 2010) and pre-processing has not been applied to correct for this. The skill increases for larger lead times, so for larger lead times the ECMWF meteorological forecasts accumulated in the forecasting system give better predictions than historical

meteorological measurements. ~~are better model inputs than historical measurements for larger lead times.~~ The fact that the contribution of initial hydrological conditions at the forecast issuing day decreases for larger lead times (~~also see~~reflected in Fig. ~~9b~~10b) adds to this skill.

The forecast skill for both snowmelt floods and snow accumulation generated low streamflow events decreases from a lead time of 8 days, which indicates a decreasing skill of ECMWF temperature forecasts for large lead times.

~~For low streamflow generated by snow accumulation and precipitation deficits, errors from the HBV model and initial conditions make up a large part of the total error (Fig. 9b).~~

5 Discussion

The developed methodology of analysing an ensemble streamflow forecasting system has been applied to the Biała Tarnowska catchment for a 6 year period. By this, findings ~~by~~ of this study do not allow direct generalisation but serve ongoing discussions on improving streamflow forecasting. Also, a longer evaluation period would allow evaluation of more extreme definitions of high and low streamflow.

The best streamflow forecasts are obtained without pre- and post-processing. The effectiveness of QM depends on whether during the validation period the same bias is present between the CDF of the measurements and the CDF of the forecasts as during the training period. Figure ~~10~~ 11 shows large differences in biases between different years and between the training period and the validation period, suggesting that bias is affected by randomness. The relatively short time series of forecasts constrains the pre- and post-processing procedures, because different weather patterns cannot be well identified and with a longer period a more consistent bias distribution could be obtained. In addition, limitations of QM, as described by Boé et al. (2007) and Madadgar et al. (2014), are expected to play a role in the ineffectiveness of the pre- and post-processing. In spite of the limitations of QM, over the training period the pre- and post-processing strategies result in an improvement of the evaluation scores (strategy 3 with seasonal distinction gives the best performance), which indicates the potential of processing with QM if a consistent bias is present. A problem in pre- and post-processing ~~in general~~of forecasts is that the joint distribution of measurements and forecasts is often nonhomogeneous in time ~~by~~due to, for example, an improvement of forecasting systems over time (Verkade et al., 2013). The ECMWF meteorological forecasts in TIGGE, containing historical operational forecasts, have also undergone changes (Mladek, 2016).

Uncertainties in the hydrological model and model initial conditions have been ignored in the forecasting system. Considering the rank histogram results this ~~may have affected~~ affects the reliability of streamflow forecasts of short lead times and low streamflow in particular. Regarding the main effect on short lead times, Bennett et al. (2014) and Pagano et al. (2013) discuss similar findings. The lower flatness coefficients of high streamflow forecasts compared to low streamflow forecasts reflect that for high streamflow forecasts the meteorological inputs ~~is~~are relatively more important.

The classification of low and high streamflow generating processes is based on hydrometeorological information that is available from the HBV model and measurement ~~data~~ series. This provides more insight in the performance of the

forecasting system than a seasonal characterisation. However, s~~S~~ome assumptions must be kept in mind when interpreting the results. It is assumed that snow accumulation before an event is embedded in the snowpack storage of the HBV model. If a snowpack is present the event is classified as snowmelt flood or snow accumulation low streamflow. The lumped model causes a simplification here, because when there is a snowpack present in the model there is not necessarily a snowpack that covers the whole catchment. If no snowpack is present, it is assumed that the low streamflow event or high streamflow event is caused by low or high rainfall. The threshold of 10 mm day^{-1} (see Table 2) is a~~n unrefined~~ simplification to distinguish between short-rain floods and long-rain floods. The simple character of the classification rules especially has consequences for the classification of events that are caused by a combination of processes, which often occur in practice and result in the highest floods. Another point is that only short-term information (from the day preceding the forecast issuing day) is used to classify the processes. The lag time between precipitation peaks and streamflow peaks does not necessarily match with the HBV model calculation time step and the classification rules used. Consequently, a streamflow at the day following a high rainfall event is classified as a short-rain flood, whereas the real streamflow peak might come one day later.

In the hydrological model the lag time between a rainfall event and the streamflow peak is set to 1 day. However, the timing of a rainfall event ~~during the~~on a day ~~is very~~ will be important, especially in a small catchment. Evaluation of forecast performance in this paper indicates that the lag time is critical in the forecasting system, especially for short-rain floods. The results in Fig. ~~8b~~ 9b show that the ratio between the CRPS against ~~perfect forecasts~~observed meteorological input forecasts and the CRPS against streamflow measurements is above 100% for ~~short rain floods~~high streamflows, and short-rain floods in particular. This means that these forecasts are closer to the measurements than to the ~~perfect forecasts~~observed meteorological input forecasts. ~~The precipitation peak in the measurements and the precipitation peak in the meteorological forecasts can be shifted one day with respect to each other and this can cause that the timing of the peak of the streamflow forecasts better corresponds to the streamflow measurements than to the peak of the perfect streamflow forecasts.~~Analyses show that on high streamflow days on which the forecasts are closer to the measurements than to the observed meteorological input forecasts (28% at lead time of 1 day to 48% of the days at lead time of 10 days), depending on lead time, on 50% to 66% of the days the forecasts are closer to the measurement than the observed meteorological input forecast. This indicates a hydrological model deficiency, either from the rainfall-runoff relation or the flood peak timing. The precipitation peak in the measurements and the precipitation peak in the meteorological forecasts can be shifted one day with respect to each other and this ~~can~~may cause that the timing of the peak of the streamflow forecasts better corresponds to the streamflow measurements ~~than to the peak of the perfect streamflow forecasts~~. Of the 97 separate peak streamflow days, on 6 days (lead time of 6 days) to 17 days (lead time of 1 day) the flood peak day of the observed meteorological input forecasts does not match to the peak day of the measurement but the peak day of the mean of the ensemble forecast does match to the peak day of the measurements. This illustrates that hydrological model deficiencies have a considerable effect on the observed meteorological input forecasts and the ensemble forecasts.

It is not trivial to compare the CRPS results to results in other studies, because the value depends on the magnitude of the evaluated variable (Ye et al., 2014). A similarity between the results in this study and previous studies is that the

performance of the streamflow forecasts decreases with lead time. ~~Since~~ Because Bennett et al. (2014) use the same alternative forecast set, the CRPSS results can be compared. Although Bennett et al. (2014) use a ~~very~~ different forecasting system and apply it to different ~~situations~~conditions, the forecast skills are comparable to the forecast skills obtained in this study.

6 Conclusions

We developed a methodology to analyse an ensemble streamflow forecasting system. For the case study of the Biała Tarnowska catchment we conclude:

- There are large differences in forecast skill, ~~-~~compared to alternative forecasts based on historical meteorological measurements, for different runoff generating processes~~, compared to alternative forecasts based on historical measurements of precipitation and temperature~~. The system skilfully forecasts high streamflow events, although the skill depends on the runoff generating process and lead time. Also low streamflow events that are generated by snow accumulation are skilfully forecasted. Since the hit rates are high compared to the false alarm rates, the system has potential to generate forecasts for these streamflow categories. Sharpness of the forecasts is good, although it decreases with lead time. Medium streamflow events and low streamflow events that are generated by a precipitation deficit are not skilfully forecasted.
- When this or any other forecasting system is (further) developed with the objective to generate more accurate high streamflow forecasts, it is recommended to focus on improving the meteorological forecast inputs because errors from the meteorological forecasts are dominant in high streamflow forecasts. This can be achieved by improving the meteorological forecasts (e.g. using the higher resolution forecasts from COSMO-LEPS (Renner et al., 2009)) or by improving the pre-processing step. Also the hydrological model performance on high streamflow must be improved, by specific calibration on flood peak timing and high streamflow conditions. To improve low streamflow forecasts, it is recommended to focus first on the hydrological model performance. In this study the calibration of the hydrological model is skewed to high streamflow ~~situations~~conditions. An easy improvement of the forecasts can be achieved by calibrating the hydrological model specifically on low streamflow ~~events~~conditions. Besides improvement of the hydrological model, further research should be done to improve the meteorological forecasts, especially the precipitation forecasts (problem of forecasting of drizzle). When the forecasting system is applied exclusively on low or high streamflow forecasts, the alternative forecast set should be reconsidered.
- The ensemble streamflow forecasting system shows good resolution and sharpness, but the reliability of the streamflow forecasts must be improved. Therefore, ~~To improve the reliability of the ensemble streamflow forecasts~~ it is recommended to include uncertainties in hydrological model parameters and initial conditions in the forecasting system. Particularly for low streamflow forecasts this is essential. ~~The uncertainty in the relationship between the fraction of fast runoff and total streamflow to update initial states might be utilized to incorporate initial condition~~

~~uncertainty. Since~~ Because the precipitation and temperature forecasts are also underdispersed, we recommend to investigate how the reliability of the precipitation and temperature forecasts can be improved, potentially by adding meteorological forecasts from other forecasting systems ('super-ensembles') (Bennett et al., 2014; Bougeault et al., 2010; Fleming et al., 2015; He et al., 2009) or by improved pre-processing.

- Pre-processing with QM slightly improves the meteorological forecasts, but this loses its effect after propagating through the hydrological model. Post-processing of streamflow forecasts was not effective either. ~~and post-processing with QM was not effective. In the discussion several limitations of QM have been described.~~ A longer time series of forecasts and a retrospective forecast set would possibly promote the success of pre- and post-processing. Moreover, techniques such as a Bayesian joint probability approach (Bennett et al., 2014; Khajehei and Moradkhani, 2017), regression techniques (Verkade et al., 2013; Hashino et al. 2007), Schaake shuffle to ascribe realistic space-time variability (Clark et al., 2004), and weather typing (Boé et al., 2007; Wetterhall et al., 2012) or hydrological process typing, can improve the effectiveness of pre- and post-processing procedures.
- It is recommended to extend the study to other catchments and (if possible) with longer forecast datasets, to investigate the generality of the results and to test more extreme high and low streamflow thresholds.

The findings only apply to the study catchment and the developed system set-up, but the presented methodology of analysing an ensemble streamflow forecasting system is generally applicable. The methodology provides valuable information about the forecasting system, in which ~~situations~~ conditions it can be used, and how the system can be improved effectively.

Acknowledgements

Marzena Osuch (Institute of Geophysics Polish Academy of Sciences) and Adam Kiczko (Warsaw University of Life Sciences) are thanked for valuable discussions and support on methods.

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.

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.

~~Atger, F.: Verification of intense precipitation forecasts from single models and ensemble prediction systems, Nonlinear Process. Geophys., 8(6), 401–417, doi:10.5194/npg-8-401-2001, 2001.~~

Bennett, J. C., Robertson, D. E., Shrestha, D. L. and Wang, Q. J.: Selecting reference streamflow forecasts to demonstrate the performance of NWP-forced streamflow forecasts, in MODSIM 2013, 20th International Congress on Modelling

and Simulation, edited by J. Piantadosi, R. S. Anderssen, and J. Boland, pp. 2611–2617, Modelling and Simulation Society of Australia and New Zealand, Adelaide, Australia. [online] Available from: http://www.mssanz.org.au/modsim2013/L8/bennett.pdf, 2013.

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.

Boé, J., Terray, L., Habets, F. and Martin, E.: Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies, Int. J. Climatol., 27(12), 1643–1655, doi:10.1002/joc.1602, 2007.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y. Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L. and Worley, S.: The THORPEX Interactive Grand Global Ensemble, Bull. Am. Meteorol. Soc., 91(8), 1059–1072, doi:10.1175/2010BAMS2853.1, 2010.

Bourdin, D. R. and Stull, R. B.: Bias-corrected short-range Member-to-Member ensemble forecasts of reservoir inflow, J. Hydrol., 502, 77–88, doi:10.1016/j.jhydrol.2013.08.028, 2013.

Bouwer, L. M., Bubeck, P. and Aerts, J. C. J. H.: Changes in future flood risk due to climate and development in a Dutch polder area, Glob. Environ. Chang., 20(3), 463–471, doi:10.1016/j.gloenvcha.2010.04.002, 2010.

Bröcker, J. and Smith, L. A.: Increasing the Reliability of Reliability Diagrams, Weather Forecast., 22(3), 651–661, doi:10.1175/WAF993.1, 2007.

Buizza, R., Hollingsworth, A., Lalaurette, F. and Ghelli, A.: Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System, Weather Forecast., 14(2), 168–189, doi:10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2, 1999.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y.: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems, Mon. Weather Rev., 133(5), 1076–1097, doi:10.1175/MWR2905.1, 2005.

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.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R.: The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields, J. Hydrometeorol., 5(1), 243–262, doi:10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2, 2004.

~~Candille, G. and Talagrand, O.: Evaluation of probabilistic prediction systems for a scalar variable, Q. J. R. Meteorol. Soc., 131(609), 2131–2150, doi:10.1256/qj.04.71, 2005.~~

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol., 375(3–4), 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

Das, S., Abraham, A., Chakraborty, U. K. and Konar, A.: Differential Evolution Using a Neighborhood-Based Mutation Operator, IEEE Trans. Evol. Comput., 13(3), 526–553, doi:10.1109/TEVC.2008.2009457, 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.

~~Demeritt, D., Nobert, S., Cloke, H. L. and Pappenberger, F.: The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management, Hydrol. Process., 27(1), 147–157, doi:10.1002/hyp.9419, 2013.~~

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, 2013a.

Demirel, M. C., Booij, M. J. and Hoekstra, A. Y.: Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times, Hydrol. Process., 27(19), 2742–2758, doi:10.1002/hyp.9402, 2013b.

Demirel, M. C., Booij, M. J. and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, Hydrol. Earth Syst. Sci., 19(1), 275–291, doi:10.5194/hess-19-275-2015, 2015.

Déqué, M.: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values, Glob. Planet. Change, 57(1–2), 16–26, doi:10.1016/j.gloplacha.2006.11.030, 2007.

ECMWF: Describing ECMWF's forecasts and forecasting system, edited by B. Riddaway, ECMWF Newsl., 133, 11–13 [online] Available from: http://old.ecmwf.int/publications/newsletters/pdf/133.pdf, 2012.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognit. Lett., 27(8), 861–874, doi:10.1016/j.patrec.2005.10.010, 2006.

Fleming, S. W.: Demand modulation of water scarcity sensitivities to secular climatic variation: theoretical insights from a computational maquette, Hydrol. Sci. J., 61(16), 2849–2859, doi:10.1080/02626667.2016.1164316, 2016.

Fleming, S. W., Bourdin, D. R., Campbell, D., Stull, R. B. and Gardner, T.: Development and Operational Testing of a Super-Ensemble Artificial Intelligence Flood-Forecast Model for a Pacific Northwest River, J. Am. Water Resour. Assoc., 51(2), 502–512, doi:10.1111/jawr.12259, 2015.

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.

Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Mon. Weather Rev., 129(3), 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hamill, T. M. and Colucci, S. J.: Evaluation of Eta–RSM Ensemble Probabilistic Precipitation Forecasts, Mon. Weather Rev., 126(3), 711–724, doi:10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2, 1998.

Hamill, T. M., Mullen, S. L., Snyder, C., Toth, Z. and Baumhefner, D. P.: Ensemble Forecasting in the Short to Medium Range: Report from a Workshop, Bull. Am. Meteorol. Soc., 81(11), 2653–2664, doi:10.1175/1520-0477(2000)081<2653:EFITST>2.3.CO;2, 2000.

Hashino, T., Bradley, A. A. and Schwartz, S. S.: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, Hydrol. Earth Syst. Sci., 11(2), 939–950, doi:10.5194/hess-11-939-2007, 2007.

He, Y., Wetterhall, F., Cloke, H. L., Pappenberger, F., Wilson, M., Freer, J. and McGregor, G.: Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions, Meteorol. Appl., 16(1), 91–101, doi:10.1002/met.132, 2009.

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.

Houser, P. R., De Lannoy, G. J. M. and Walker, J. P.: Hydrologic Data Assimilation, in Approaches to Managing Disaster - Assessing Hazards, Emergencies and Disaster Impacts, edited by J. Tiefenbacher, pp. 41–64, InTech., 2012.

IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Core Writing Team, R. K. Pachauri, and L. A. Meyer, IPCC, Geneva, Zwitzerland. [online] Available from: http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full.pdf, 2014.

Kang, T. H., Kim, Y. O. and Hong, I. P.: Comparison of pre- and post-processors for ensemble streamflow prediction, Atmos. Sci. Lett., 11(2), 153–159, doi:10.1002/asl.276, 2010.

Khajehei, S. and Moradkhani, H.: Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach, J. Hydrol., 546, 476–489, doi:10.1016/j.jhydrol.2017.01.026, 2017.

Kiczko, A., Romanowicz, R. J., Osuch, M. and Pappenberger, F.: Adaptation of the Integrated Catchment System to On-line Assimilation of ECMWF Forecasts, in Stochastic Flood Forecasting System, edited by R. J. Romanowicz and M. Osuch, pp. 173–186, Springer International Publishing, Cham, Switzerland., 2015.

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.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249(1–4), 2–9, doi:10.1016/S0022-1694(01)00420-6, 2001.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, J. Comput. Phys., 227(7), 3515–3539, doi:10.1016/j.jcp.2007.02.014, 2008.

Lindström, G., Johansson, B., Persson, M., Gardelin, M. and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201(1–4), 272–288, doi:10.1016/S0022-1694(97)00041-3, 1997.

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H. J., Kumar, S., Moradkhani, H., Seo, D. J., Schwanenberg, D., Smith, P., Van Dijk, A. I. . J. M., Van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O. and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities, Hydrol. Earth Syst. Sci., 16(10), 3863–3887, doi:10.5194/hess-16-3863-2012, 2012.

Lu, J., Sun, G., McNulty, S. G. and Amatya, D. M.: A comparison of six potential evapotranspiration methods for regional use in the Southeastern United States, J. Am. Water Resour. Assoc., 41(3), 621–633, doi:10.1111/j.1752-1688.2005.tb03759.x, 2005.

Madadgar, S., Moradkhani, H. and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrol. Process., 28(1), 104–122, doi:10.1002/hyp.9562, 2014.

Martina, M. L. V., Todini, E. and Libralon, A.: A Bayesian decision approach to rainfall thresholds based flood warning, Hydrol. Earth Syst. Sci., 10(3), 413–426, doi:10.5194/hess-10-413-2006, 2006.

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.

Napiorkowski, M. J., Piotrowski, A. P. and Napiorkowski, J. J.: Stream temperature forecasting by means of ensemble of neural networks: Importance of input variables and ensemble size, in River Flow 2014, edited by A. J. Schleiss, G. De Cesare, M. J. Franca, and M. Pfister, pp. 2017–2025, Taylor & Francis Group, London, UK., 2014.

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.

Osuch, M., Romanowicz, R. J. and Booij, M. J.: The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics, Hydrol. Sci. J., 60(7–8), 1299–1316, doi:10.1080/02626667.2014.967694, 2015.

Pagano, T. C., Shrestha, D. L., Wang, Q. J., Robertson, D. and Hapuarachchi, P.: Ensemble dressing for hydrological applications, Hydrol. Process., 27(1), 106–116, doi:10.1002/hyp.9313, 2013.

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.

Penning-Rowsell, E. C., Tunstall, S. M., Tapsell, S. M. and Parker, D. J.: The Benefits of Flood Warnings: Real But Elusive, and Politically Significant, Water Environ. J., 14(1), 7–14, doi:10.1111/j.1747-6593.2000.tb00219.x, 2000.

Persson, A. and Andersson, E.: User guide to ECMWF forecast products. [online] Available from: http://old.ecmwf.int/products/forecasts/guide/user_guide.pdf, 2013.

Piani, C., Haerter, J. O. and Coppola, E.: Statistical bias correction for daily precipitation in regional climate models over Europe, Theor. Appl. Climatol., 99(1–2), 187–192, doi:10.1007/s00704-009-0134-9, 2010.

Ranjan, R.: Combining and Evaluating Probabilistic Forecasts, PhD thesis, University of Washington, Seattle, Washington USA., 2009.

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.

Rientjes, T. H. M., Muthuwatta, L. P., Bos, M. G., Booij, M. J. and Bhatti, H. A.: Multi-variable calibration of a semi-distributed hydrological model using streamflow data and satellite-based evapotranspiration, J. Hydrol., 505, 276–290, doi:10.1016/j.jhydrol.2013.10.006, 2013.

Rojas, R., Feyen, L. and Watkiss, P.: Climate change and river floods in the European Union: Socio-economic consequences and the costs and benefits of adaptation, Glob. Environ. Chang., 23(6), 1737–1751, doi:10.1016/j.gloenvcha.2013.08.006, 2013.

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

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.

Shi, X., Wood, A. W. and Lettenmaier, D. P.: How Essential is Hydrologic Model Calibration to Seasonal Streamflow Forecasting?, J. Hydrometeorol., 9(6), 1350–1363, doi:10.1175/2008JHM1001.1, 2008.

Tao, Y., Duan, Q., Ye, A., Gong, W., Di, Z., Xiao, M. and Hsu, K.: An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin, J. Hydrol., 519(Part D), 2890–2905, doi:10.1016/j.jhydrol.2014.04.040, 2014.

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.

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.

Werner, M. G. F., Schellekens, J. and Kwadijk, J. C. J.: Flood early warning systems for hydrological (sub) catchments, in Encyclopedia of Hydrological Sciences, edited by M. G. Anderson and J. J. McDonnell, John Wiley & Sons., 2005.

Wetterhall, F., Pappenberger, F., He, Y., Freer, J. and Cloke, H. L.: Conditioning model output statistics of regional climate model precipitation on circulation patterns, Nonlinear Process. Geophys., 19(6), 623–633, doi:10.5194/npg-19-623-2012, 2012.

Wheater, H. S. and Gober, P.: Water security and the science agenda, Water Resour. Res., 51(7), 5406–5424, doi:10.1002/2015WR016892, 2015.

Wilks, D. S.: Stastistical Methods in the Atmospheric Sciences, 2nd ed., Elsevier Academic Press, Oxford, UK., 2006.

WMO: Forecast Verification: Issues, Methods and FAQ, [online] Available from: http://www.cawcr.gov.au/projects/verification/ (Accessed 12 March 2015), 2015.

Wöhling, T., Lennartz, F. and Zappa, M.: Technical Note: Updating procedure for flood forecasting with conceptual HBV-type models, Hydrol. Earth Syst. Sci., 10(6), 783–788, doi:10.5194/hess-10-783-2006, 2006.

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.

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.

Zalachori, I., Ramos, M. H., Garçon, R., Mathevet, T. and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Adv. Sci. Res., 8, 135–141, doi:10.5194/asr-8-135-2012, 2012.

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.

Figures

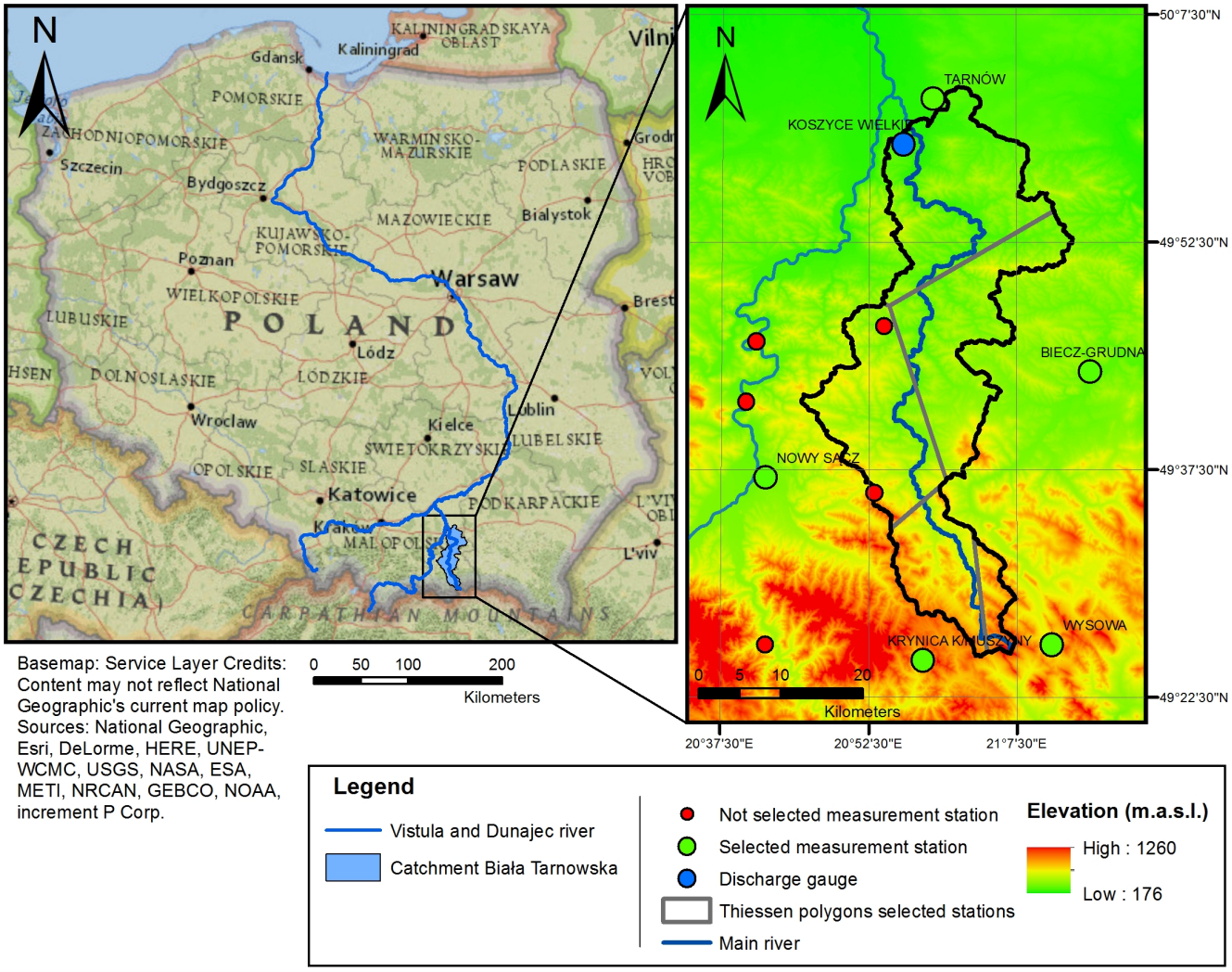

Figure 1: Location and overview of the Biała Tarnowska catchment

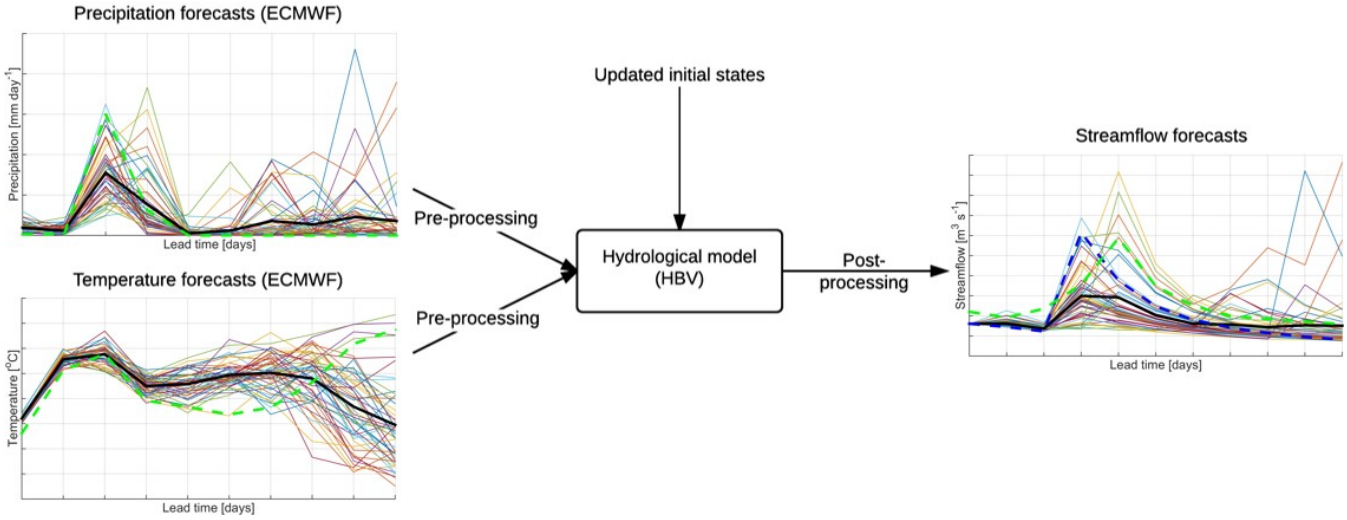

Figure 2: Structure of the ensemble streamflow forecasting system

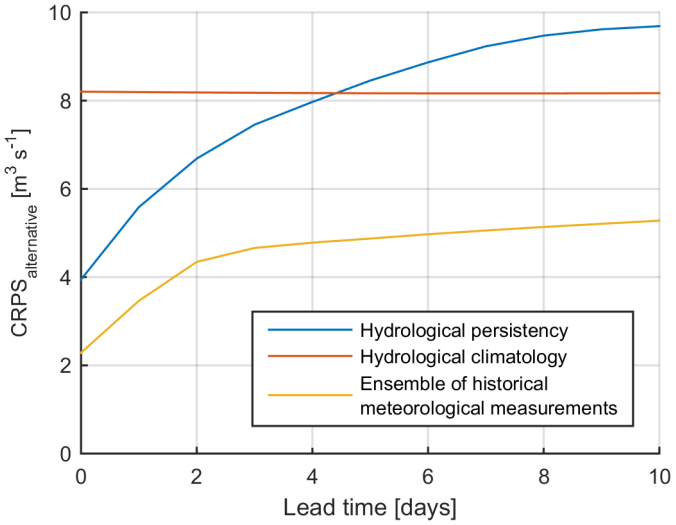

Figure 3: CRPS of three alternative forecast sets, evaluation period 2008-2013

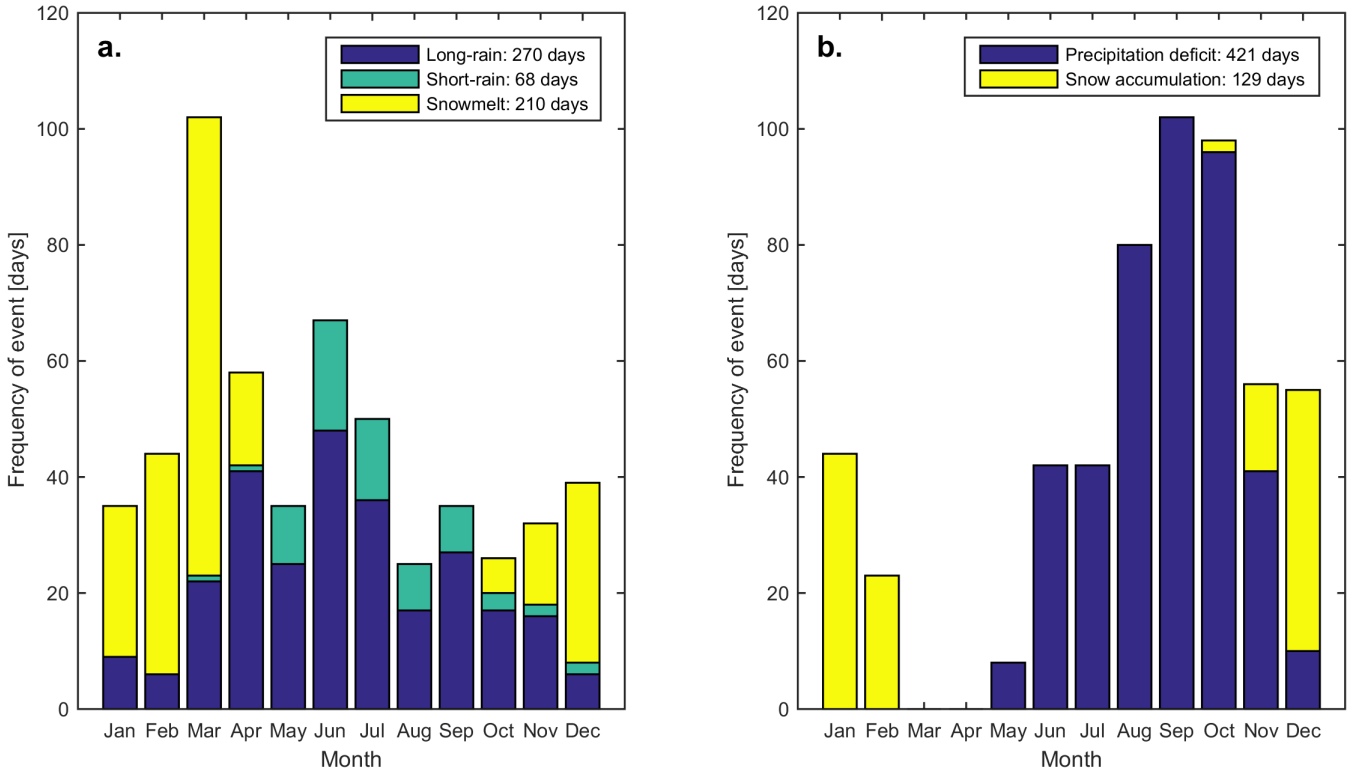

Figure 4: a. High streamflow generating processes over the year b. Low streamflow generating processes over the year, 1-11-2007 to 31-10-2013

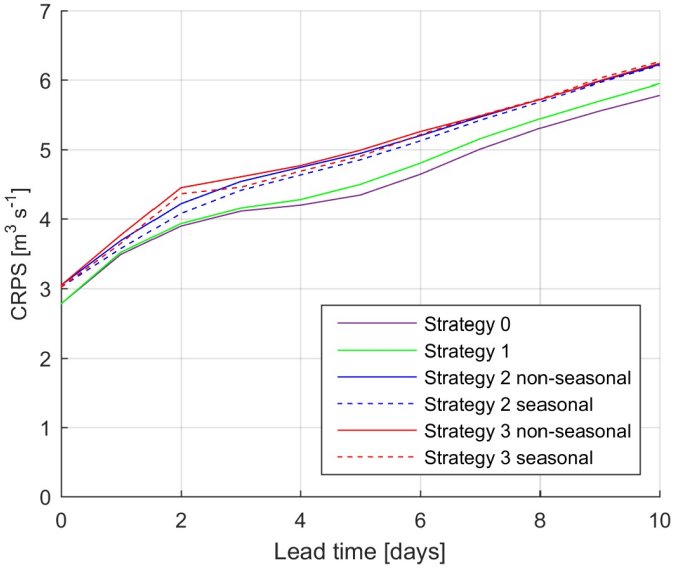

Figure 5: CRPS of streamflow forecasts over the validation period 2008-2011, by applying the post-processing strategies that are introduced in Sect. 3.1.3.

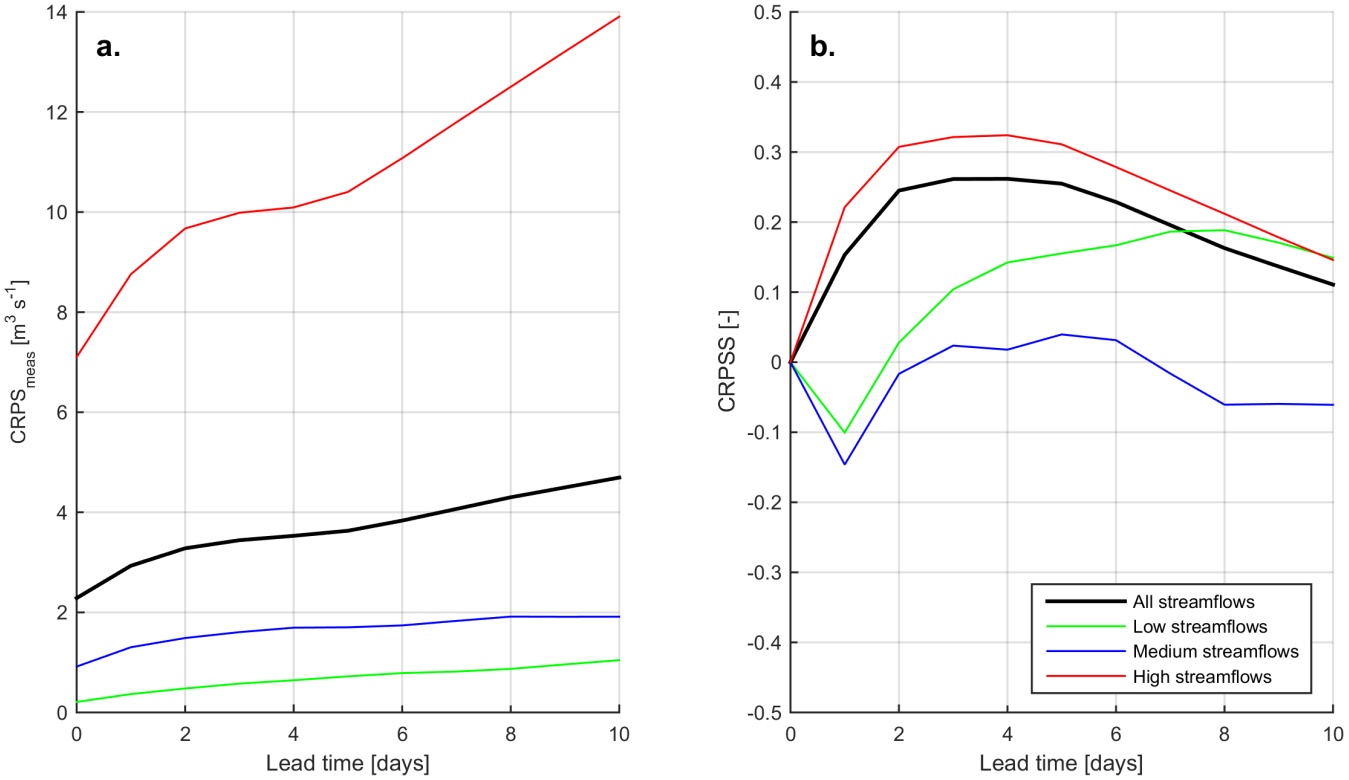

Figure 6: a. Streamflow forecasts evaluated against streamflow measurements b. Skill of the streamflow forecasts, defined in Eq. (1)

5

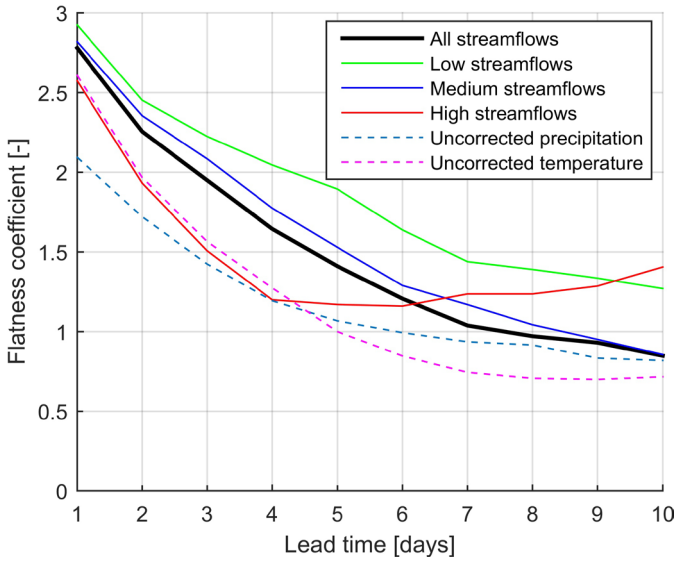

Figure 7: Rank histogram flatness coefficients. The flatness coefficients of the precipitation and temperature forecasts refer to the preceding day.

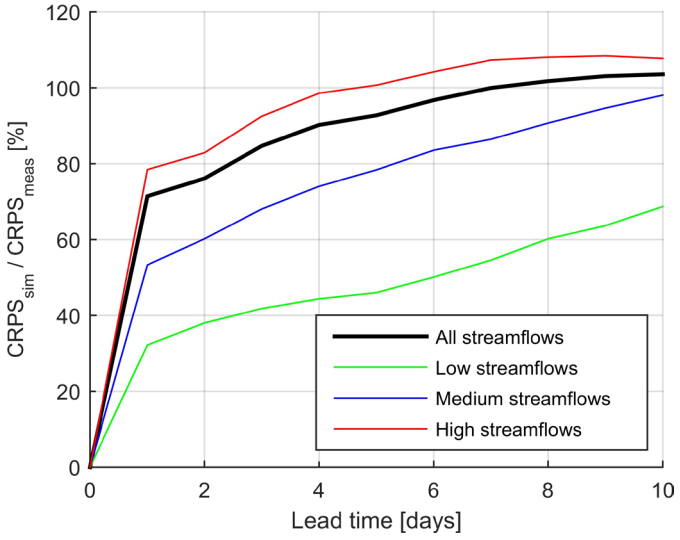

Figure 8: Ratio of errors in meteorological forecasts ($CRPS_{sim}$) to meteorological forecast + model errors ($CRPS_{meas}$)

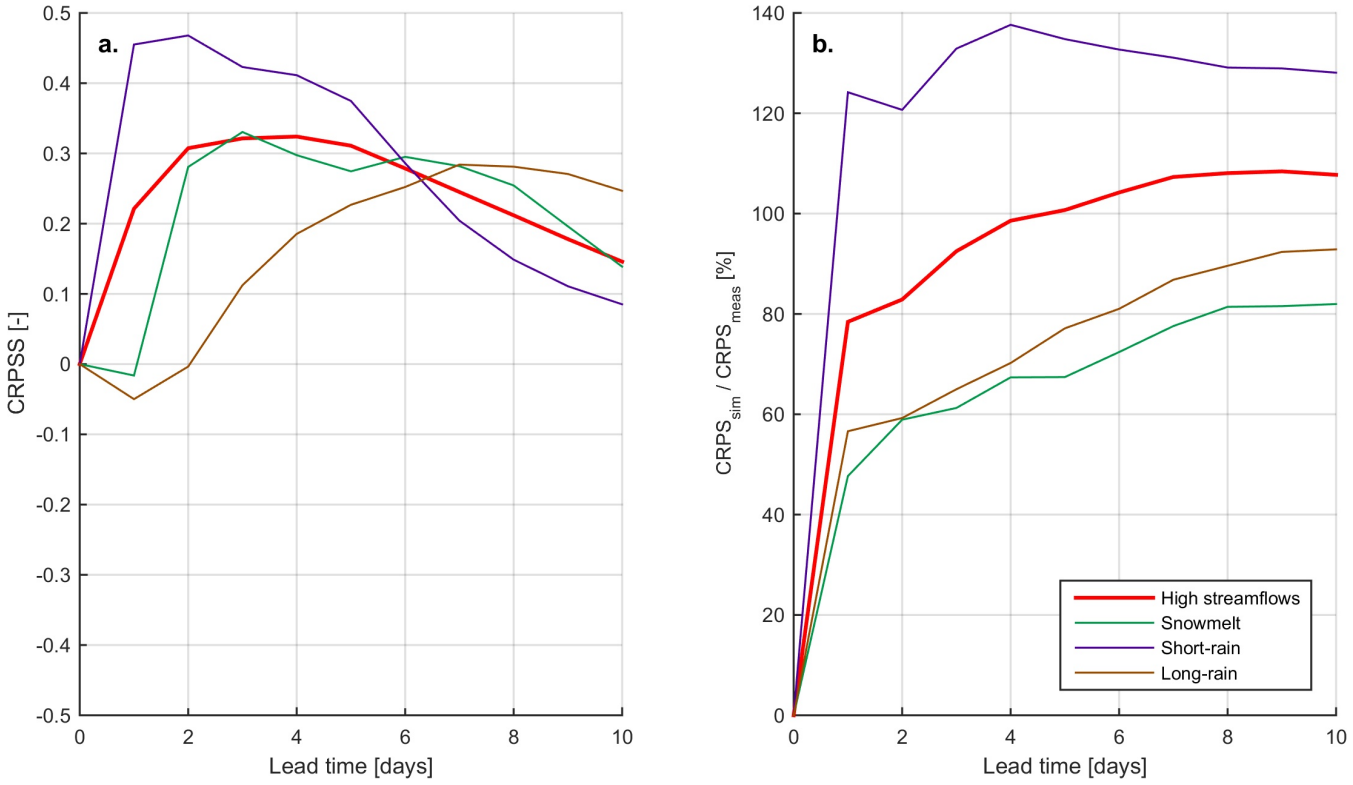

Figure 9: a. Forecast skill of high streamflow generating processes b. Ratio of errors in meteorological forecasts ($CRPS_{sim}$) to meteorological forecast + model errors ($CRPS_{meas}$).

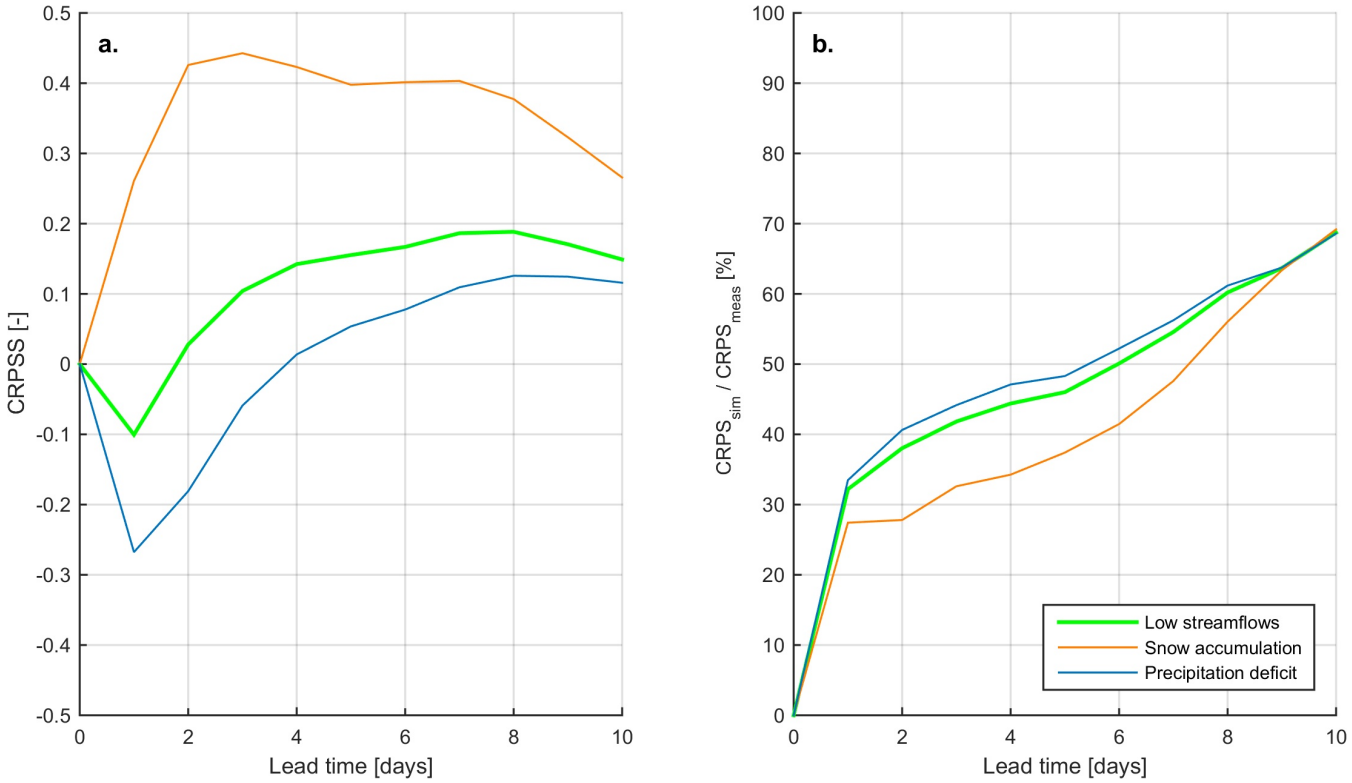

Figure 10: a. Forecast skill of low streamflow generating processes b. Ratio of errors in meteorological forecasts ($CRPS_{sim}$) to meteorological forecast + model errors ($CRPS_{meas}$).

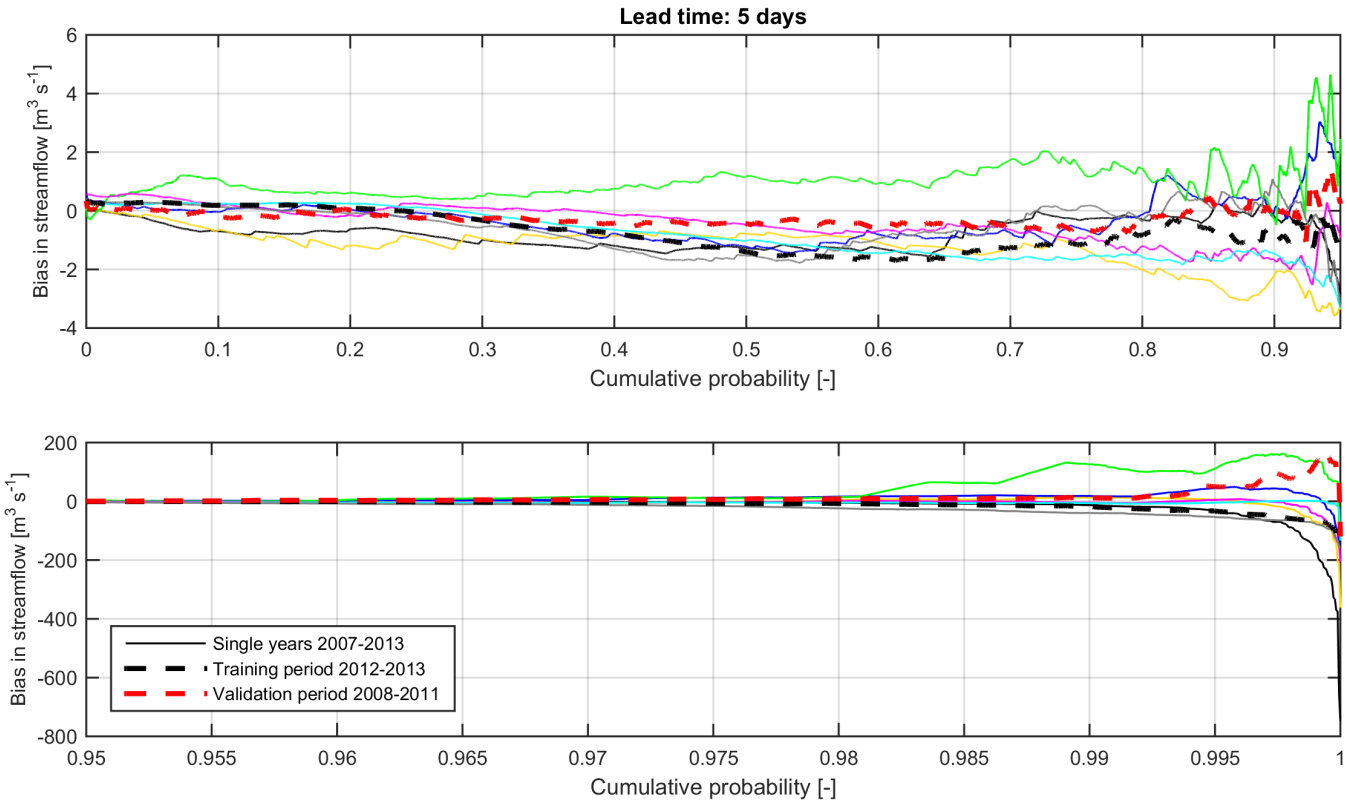

Figure 11: Difference between CDFs of the measurements and CDFs of the uncorrected streamflow forecasts per hydrological year (upper panel cumulative probability 0 – 0.95 and lower panel 0.95 – 1.0). Each thin line refers to a single year between 2007 and 2013. This figure is for a lead time of 5 days.

Tables

Table 1: Definition of streamflow categories

Streamflow category	Thresholds	Streamflow (from measurements 1-11-2007 to 31-10-2013)
Low streamflow	$Q_{obs} \leq Q_{75}$	$Q_{obs} \leq 2.76\ m^3/s$
Medium streamflow	$Q_{75} < Q_{obs} \leq Q_{25}$	$2.76\ m^3/s < Q_{obs} \leq 10.35\ m^3/s$
High streamflow	$Q_{25} < Q_{obs}$	$10.35\ m^3/s < Q_{obs}$

Table 2: Characterization of the high streamflow generating processes

Process	Characterization	Rules for classification
Snowmelt flood	Snowmelt floods and rain-on-snow floods (explained by Merz and Blöschl (2003)) are considered as one category. All high streamflow events where snow is involved are characterized as snowmelt floods, because the snowpack and/or frozen soil underneath play an important role in the runoff process.	• Snowpack (HBV) at ~~forecast~~ day-1
Short-rain flood	Short-rain floods and flash floods (characterized by Merz and Blöschl (2003)) are combined. Flash floods are classed in this category as well, because only daily measurements and forecasts are available.	• No snowpack (HBV) at ~~forecast~~ day-1 • Rainfall at ~~forecast~~ day-1 above 10 mm: With small initial storage in the catchment (HBV), precipitation of 10 mm day^{-1} at the day preceding the streamflow event causes a streamflow event above the high streamflow threshold.
Long-rain flood	Long-rain flood processes are explained by Merz and Blöschl (2003). This category applies when a streamflow event is not directly generated by snowmelt or high precipitation.	• No snowpack (HBV) at ~~forecast~~ day-1 • Rainfall at ~~forecast~~ day-1 below 10 mm

Table 3: Characterization of the low streamflow generating processes

Process	Characterization	Rules for classification
Snow accumulation	If precipitation is snow and does not melt directly, accumulation occurs.	• Snowpack (HBV) at ~~forecast~~ day-1
Precipitation deficit	When low rainfall and high evapotranspiration last over a prolonged period the catchment will dry out.	• No snowpack (HBV) at ~~forecast~~ day-1

Table 4: Calibration and validation performances of the model

Run	Calibration (1-11-1971 to 31-10-2000)			Validation (1-11-2000 to 31-10-2013, excluding 2007)		
	Y	NS	E_{RV}	Y	NS	E_{RV}
~~Calibration with uncorrected input data~~	~~0.78~~	~~0.78~~	~~0%~~	~~0.69~~	~~0.74~~	~~6.5%~~
Calibration run with input data corrected for elevation	0.81	0.81	0%	0.~~72~~75	0.~~77~~78	~~6.7~~4.8%
With updating at lead time 0 days	-	-	-	0.82	0.83	1.3%
With updating at lead time 10 days	-	-	-	0.75	0.79	4.4%

Table 5: Performance over the evaluation period 2008-2013, for low, medium and high streamflow simulations (~~perfect forecasts~~observed meteorological input forecasts). The initial states are updated at the lead time of 0 days.

Lead time	E_{RV} [%]			NS [-]			E_{RMA} [-]		
[days]	Low flows	Medium flows	High flows	Low flows	Medium flows	High flows	Low flows	Medium flows	High flows
No updating	43.3	7.29	1.81	-10.9	-2.36	0.82	0.71	0.43	0.33
0	3.23	4.69	2.16	0.34	-0.14	0.86	0.11	0.16	0.25
1	6.44	7.16	2.64	-0.64	-0.53	0.84	0.19	0.21	0.29
2	8.55	8.80	2.48	-1.12	-0.88	0.83	0.23	0.25	0.31
3	11.5	9.60	2.30	-2.09	-1.07	0.83	0.29	0.28	0.32
4	13.6	10.1	2.17	-2.76	-1.15	0.83	0.33	0.30	0.32
5	15.9	10.4	2.04	-3.50	-1.33	0.83	0.37	0.31	0.32
6	18.2	10.4	1.98	-4.36	-1.43	0.83	0.41	0.32	0.32
7	19.2	10.5	2.01	-4.56	-1.53	0.83	0.43	0.34	0.32
8	20.6	10.3	2.07	-4.88	-1.62	0.83	0.45	0.35	0.32
9	22.9	10.1	2.09	-5.73	-1.70	0.83	0.49	0.35	0.32
10	24.0	10.0	2.13	-6.09	-1.77	0.83	0.50	0.36	0.32