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 that gives insight into 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 different hydrometeorological conditions. Precipitation and temperature forecasts from the European Centre for Medium-Range Weather Forecasts served as inputs to a deterministic lumped hydrological (HBV) model. Due to a non-homogeneous bias in time, pre- and post-processing of the meteorological and streamflow forecasts are not effective. The best forecast skill, relative to alternative forecasts based on meteorological climatology, is shown for high streamflow and snow accumulation low streamflow events. Forecasts of medium streamflow events and low streamflow events under precipitation deficit conditions show less skill. To improve performance of the forecasting system for high streamflow events, the meteorological forecasts are most important. Besides, it is recommended that the hydrological model be calibrated specifically on low streamflow conditions and high streamflow conditions. Further, it is recommended that the dispersion (reliability) of the ensemble streamflow forecasts is enlarged by including the uncertainties in the hydrological model parameters and the initial conditions, and by 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 in mitigating the negative effects of extreme events, by enabling early warning. Accurate forecasting is becoming increasingly more important, because the 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). In addition, due to socio-economic development the impacts of extreme events increase further (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 on the possibility that an event will occur (Krzysztofowicz, 2001; Thielen et al., 2009), and allow a quantification of the forecast uncertainty (Krzysztofowicz, 2001; Zappa et al., 2011). Uncertainties in streamflow forecasts originate from the meteorological inputs, and the 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 have investigated the performance of ensemble forecasting systems, e.g. Alfieri et al. (2014) for the European Flood Awareness System, 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 demonstrated a deterioration of performance with increasing lead time. However, most studies 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). Studies on 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 the effects of runoff processes, such as snowmelt and extreme rainfall events, on the performance of ensemble forecasts. The only study we found that bears on this is the study by Roulin and Vannitsem (2005), who concluded that their high streamflow forecasting system is more skilful for the winter period than for the summer period. Next to an assessment of performance, information on the relative importance of uncertainty sources in the forecasts is essential in improving the forecasts effectively (Yossef et al., 2013). A number of studies have reported on how errors in the meteorological forecasts and the hydrological model contribute to errors in medium-range hydrological forecasts. Demargne et al. (2010) showed that hydrological model uncertainties (model parameters, initial conditions, and model structure) are most significant at short lead times. The extent depends on the streamflow category: hydrological model uncertainties significantly degrade the evaluation score up to a lead time of 7 days for all flows, whereas this is only up to a lead time of 2 days for 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 attributed to the meteorological forecasts having insufficient variability. In contrast, the high forecast probabilities (low threshold) are overpredicted, which Renner et al. (2009) attributed to both the hydrological model and the meteorological input data. Olsson and Lindström (2008) found an underdispersion of ensemble flood forecasts, which decreases with lead time. The meteorological forecasts and the hydrological model have a comparable contribution to this. In addition, Olsson and Lindström (2008) showed overprediction of forecast probabilities over high thresholds, which they primarily attributed to the meteorological forecasts. Demirel et al. (2013a) concluded that the uncertainty of the hydrological model parameters has the largest effect and meteorological input uncertainty has the smallest effect on low streamflow forecasts. Based on those studies, we can say that for high streamflow forecasts uncertainties in the meteorological forecasts are dominant, whereas for low streamflow forecasts the uncertainties in the hydrological model are more important.

The objective of this study is to investigate the performance and limitations of ECMWF's 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 the runoff generating processes. We aim to evaluate whether the performance of the forecasting system relates to runoff generating processes, based on hydrometeorological conditions. Further, we assess whether the main source of forecast error is the meteorological inputs or deficiencies in 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 mountainous Biała Tarnowska catchment in southern Poland serves as study area (Fig. 1). Napiorkowski et al. (2014) describe the catchment. The Biała Tarnowska River discharges into the Dunajec River, which is a tributary of the Vistula River. The length of the river is 101.8 km, with a catchment area of 956.9 km^2. We selected this catchment because of its large variation in streamflow and seasonal variation in runoff generating processes.The mean streamflow is 9.4 m^3 s^{-1} (1972–2013). The highest measured streamflow is 611 m^3 s^{-1}. During winter and spring, snow(melt) plays an important role. A comparison of the time series of precipitation and streamflow shows that the lag time between intense precipitation events and related peaks in streamflow varies between 1 and 3 days.

Precipitation and temperature 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. The measurement series were provided by the Polish Institute of Meteorology and Water Management. Given that meteorological stations are mostly located in valleys and precipitation and temperature vary with elevation, the catchment averages may be biased (Panagoulia, 1995; Sevruk, 1997). Following Akhtar et al. (2009), we corrected the precipitation measurements using relative correction factors (in %), whereas we corrected the temperature measurements using absolute correction factors (in °C). The precipitation 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 small number of stations limits the accuracy of the precipitation and temperature gradients, we used the calculated precipitation gradients because of the apparent difference between the two periods. The temperature gradient is rather constant over the year and therefore we applied the global standard temperature lapse rate of 0.65 °C 100 m^{-1}. The measurements from each station were corrected for the difference between the elevation of the station and the mean elevation of its respective Thiessen polygon. Subsequently, to represent the catchment averages, the corrected measurements were weighted based on the relative coverage of their Thiessen polygon (Fig. 1). With 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 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 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. They consist of one control forecast (no perturbation) and 50 ensemble members. The ensemble members should represent the 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 the maximum lead time is 10 days, 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) who used 9 or 10 days as maximum lead times for hydrological forecasting. Because we use a lumped hydrological model with a daily time step (Sect. 3.1.1), we averaged the 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 may lead to biases. We ignored correction for such elevation errors, 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 weighted the temperature forecasts at 00:00 hr., 12:00 hr. and 24:00 hr. by 25%, 50% and 25% respectively.

3 Methods

3.1 The ensemble streamflow forecasting system

The ensemble streamflow forecasting system consists of multiple components, shown in Fig. 2. Uncertainties in the meteorological forecasts, the model parameters, the model initial conditions and the model structure affect 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 these sources must be incorporated (Bourdin and Stull, 2013; Krzysztofowicz, 2001; Zappa et al., 2011). Bennett et al. (2014) and Cloke and Pappenberger (2009) stated that uncertainties in the 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 the uncertainty in the meteorological forecasts to focus on the effect of ensemble meteorological forecasts on streamflow forecasts. Consequently, an underdispersion of the streamflow forecasts may be expected.

3.1.1 Hydrological model

The hydrological model we use is a lumped Hydrologiska Byråns Vattenbalansavdelning (HBV) model that we run at a daily time step. 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 were based on air temperature 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 the result of the spatial and temporal resolution of the available data. The measurements of precipitation and temperature available from five meteorological stations and streamflow from one discharge gauging station do not justify the application of a spatially distributed hydrological model. The River Rhine forecasting suite also adopts the HBV model at a daily time step as a semi-distributed model to 134 sub catchments (Renner et al., 2009). The catchment area of Biała Tarnowska is comparable to the area of the sub catchments in the River Rhine forecasting suite.

To calibrate the HBV model we used the Differential Evolution with Global and Local neighbourhoods (DEGL) algorithm, described by Das et al. (2009). The settings were adopted from the best performing variant of Das et al. (2009) and the maximum number of model runs is set at 50000. 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 a poor to satisfactory performance. The model was calibrated using the period 1 November 1971 to 31 October 2000, with the time series of precipitation and temperature as inputs and streamflow measurements as the reference output. The validation period was 1 November 2000 to 31 October 2013. Initialization periods of 10 months and 1 year, respectively, ensure realistic initial conditions on the first day of the calibration and the validation period.

3.1.2 Updating of initial states

To best represent the hydrological conditions in the catchment on the forecast issuing day, a hydrological forecasting system often relies on the updating of the hydrological model states, 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 applied the fairly simple and direct state updating procedure introduced by Demirel et al. (2013a), which relies on the autocorrelation of streamflow to update model states. The measured streamflow of the day preceding the forecast issuing day is divided into a fast and a slow runoff component to update the fast runoff reservoir and the slow runoff reservoir of the HBV model. To determine the ratio between these components, a relation between the total simulated streamflow and the fraction of fast runoff is established based on historical simulations.

3.1.3 Pre- and post-processing

Errors in the meteorological forecasts and in the hydrological models introduce biases in the mean and errors in the dispersion of ensemble streamflow forecasts (Cloke and Pappenberger, 2009; Khajehei and Moradkhani, 2017; Verkade et al., 2013). Several studies have suggested that post-processing of streamflow forecasts is more effective in improving the forecast 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. Zalachori et al. (2012) concluded 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 were tested.

Many studies have 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) have advised the use of QM, because of the good performance regarding sharpness and discrimination and the simplicity of the method. QM matches the cumulative distribution function (CDF) of the forecasts over a training period 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 explained QM. The empirical CDFs of the measurements and forecasts were established on the training period 1 November 2011 to 31 October 2013 (two hydrological years) and validated on the period 1 November 2007 to 31 October 2011.

Distributions may be different for different lead times and weather patterns or seasons (Boé et al., 2007; Wetterhall et al., 2012), so we tested three QM set-ups both with and without distinguishing lead times and seasons. Combining the options for pre-processing and post-processing results in four processing strategies. Strategy 0 applies no pre- and post-processing. Strategy 1 and 2 applies QM to pre-process the meteorological forecasts, without and with post-processing respectively. In strategy 2, the post-processing is performed on the basis of the difference between ‘observed meteorological input forecasts’ (streamflow simulations with inputs from the meteorological measurements) and streamflow measurements to account for hydrological model uncertainties (Verkade et al., 2013). Strategy 3 applies only post-processing, on the basis of the correction between measured streamflow and streamflow forecasts generated with uncorrected meteorological forecasts. This strategy treats meteorological and hydrological model uncertainties together (Verkade et al., 2013).

3.2 Evaluation scores of the ensemble forecasts

To measure the overall performance, we employed 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 used the Continuous Ranked Probability Skill Score (CRPSS), which is the CRPS of the forecasts relative to the CRPS of alternative forecasts (Sect. 3.2.1). According to Demargne et al. (2010) and Hamill et al. (2000) a single evaluation score is inadequate to evaluate the 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 evaluated 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 used 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). The low streamflow and high streamflow thresholds are defined in Sect. 3.4.

Sharpness is 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 a probability of occurrence close to the climatological probability, instead of close to 0 or close to 1, the forecasting system is not useful, although it might be well calibrated (Ranjan, 2009; Wilks, 2006). To evaluate sharpness, we employed histograms that show the sample size of the forecast probability bins of the reliability diagrams (Ranjan, 2009; Renner et al., 2009; WMO, 2015).

Resolution is the ability to correctly forecast the occurrence and nonoccurrence of events (Demirel et al., 2013a; Martina et al., 2006). We employed relative operating characteristics (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 the diagonal) (Fawcett, 2006; Velázquez et al., 2010; WMO, 2015).

3.2.1 Alternative forecast set

The CRPS converges to the average value of the evaluated variable (with the same unit), so 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 normalized the CRPS against the CRPS of a relevant alternative forecast, a principle which has also been 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. 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). Commonly, hydrological persistency, hydrological climatology or meteorological climatology are implemented as the alternative forecast set (Bennett et al., 2013, 2014; Pappenberger et al., 2015). For hydrological persistency the most recent streamflow measurement available (i.e., from the day preceding the forecast issuing day) serves as the forecast for all lead times. For 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 meteorological climatology, meteorological 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 serves as the alternative forecast set to evaluate skill (Bennett et al., 2013, 2014; Pappenberger et al., 2015). We used a single alternative forecast set for all streamflow categories. The forecasts based on meteorological climatology result in the best CRPS scores (Fig. 3) and thus are implied to be the most appropriate alternative streamflow forecasts, as also found by Bennett et al. (2013), Bennett et al. (2014) and Pappenberger et al. (2015).

3.2.2 Rank histogram

The consistency condition states that the reference streamflow (the measurement) is just one more member of the ensemble and should be statistically indistinguishable from the ensemble forecast (Wilks, 2006). In an ensemble forecast set with a perfect dispersion 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 more background on the rank histogram, readers are referred to Hamill (2001), Wilks (2006), Velázquez et al. (2010), WMO (2015) and Zalachori et al. (2012). We used the Mean Absolute Error as flatness coefficient ε of the rank histogram, with the uniform distribution as reference:

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

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

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

n = Number of ensemble members [-]

The rank histogram and flatness coefficient contain a random element if multiple ensemble members and the measurement have the same value, such as 0 mm precipitation (Hamill and Colucci, 1998). In this case, a random rank was assigned to the measurement from the pool of ensemble members and the measurement that have the same value.

3.3 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 the measurements that serve as the reference streamflow (Renner et al., 2009). By evaluation against observed meteorological input forecasts, the streamflow measurement error and the hydrological model errors 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 error, the evaluation against streamflow measurements ($\mathrm{CRPS}_{\mathrm{meas}}$) contains errors from the meteorological forecasts and the hydrological model and the evaluation against observed meteorological input forecasts ($\mathrm{CRPS}_{\mathrm{sim}}$) 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.

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

3.4 Evaluation of streamflow categories

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

3.5 Evaluation of runoff generating processes

The high streamflow forecasts and low streamflow forecasts were evaluated for the specific runoff processes that can generate these events, based on hydrometeorological conditions. Medium flows were not evaluated for different runoff generating processes, 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 generating processes can result in high flows. Table 2 defines the processes and rules for classification. 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 distribution of processes over the year (Fig. 4a) 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. These rules for classification result in a distribution of processes over the year (Fig. 4b) that is typical for this region.

4 Results

4.1 Ensemble streamflow forecasting system

4.1.1 Calibration and validation of the hydrological model

The calibration and validation performances of the hydrological model (Table 4) are satisfactory, which indicates that the lumped model approach is plausible. The updating of the 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.

Measurements and ECMWF forecasts are simultaneously available 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 a Data Based Mechanistic (DBM) model, the performance was worse for this year (Kiczko et al., 2015). This must be the result of measurement errors and/or human influence, because it is unlikely that in this period different hydrological processes were taking place that are not captured well by both the HBV and DBM models. Therefore, we excluded the period 1 November 2006 to 31 October 2007 from the evaluation period.

Table 5 lists 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, which means that the averages of streamflow measurements in these categories are a better approximation of the measurements than the simulations. The scores highlight that the calibration was skewed to high streamflow 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 performance of the hydrological model improves considerably as a result of the updating of initial states, especially for the low streamflow simulations. The effectiveness of the updating procedure depends on the autocorrelation of daily streamflow. 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 if QM is applied separately to each lead time, whereas the best temperature forecasts are obtained if, in addition, separate relations for the summer and winter seasons are applied. The CRPS and 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 the combined pre- and post-processing strategies, the results in Fig. 5 show that strategy 0 (no pre- and post-processing) results in the best CRPS. The slight improvement of the meteorological forecasts loses its effect after propagating through the hydrological model. This is the result of hydrological model deficiencies and was also shown by Verkade et al. (2013) and Zalachori et al. (2012).

4.2 Forecast performance

4.2.1 Forecast skill

The streamflow forecasts were 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 (Table 1). The CRPS increases with lead time for all streamflow categories (Fig. 6a), so the performance of the streamflow forecasting system deteriorates with lead time. For all streamflow categories aggregated, the CRPSS is positive for all lead times (Fig. 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 on the same calendar day.

Fig. 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 low streamflow situations, runoff is mainly generated by available water storage in the catchment instead of precipitation input. Since the same initial model conditions were used to produce the alternative forecasts, the low streamflow forecasts cannot skilfully be forecasted for small lead times (<3 days). In addition, the origin of the alternative forecasts plays a role. Low streamflow events normally occur in the same period of the year due to climatic seasonality, so historical meteorological measurements on the same calendar day provide 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 are even better forecasts than the meteorological forecasts by ECMWF for this category of flows. From a lead time of 3 days the accumulated meteorological forecasts are more skilful than the historical meteorological measurements.

The medium streamflow forecasts do not have clear positive skill for all lead times. Streamflow is most often close to the medium streamflow, so forecasts based on historical meteorological measurements will be a good approximation for this category of flows.

The system has a high positive skill in forecasting high streamflow. In general, initial conditions are less important 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 the alternative forecasts can more easily deviate. In addition, high streamflow events will be less well captured by historical meteorological measurements, and thus the alternative forecasts will have lower quality for these events.

4.2.2 Forecast quality

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

The rank histograms for the 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 increases with lead time. Low streamflow is generally overestimated. Both observations can be the result of a too coarse spatial and temporal model resolution. Using a lumped model, and aggregating the meteorological inputs spatially over the catchment and temporarily over one day flattens the extreme flow events. Also the reliability diagrams (Fig. S3) show the low reliability of the streamflow forecasts, especially for small lead times. It appears that for the low streamflow forecasts the observed relative frequencies are underestimated, whereas for the high streamflow forecasts the observed relative frequencies are overestimated. The latter observation does not contradict the rank histograms, because in the rank histogram the measurements and forecasts are compared directly, whereas in the reliability diagram the measurements and forecasts are compared to a streamflow threshold. The histograms containing the sample size in the probability bins of the reliability diagrams (Fig. S3) indicate that the sharpness of the forecasts is good, because forecast probabilities of low and high streamflow are mostly close to 0 or 1, instead of close to the mean probability. The sharpness decreases with lead time.

All AUC values are above 0.85 (Fig. S4), 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 of useful prediction systems and 0.8 of good prediction systems.

4.3 Dominant error contributors

Figure 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 (Table 5). Two effects contribute to this. First, the meteorological forecasts get worse with lead time (Fig. 5) and errors in the meteorological forecasts accumulate in the hydrological forecasting system. Second, the effect of the initial hydrological conditions at the forecast issuing day becomes smaller at larger lead times.

For high streamflow forecasts the contribution of the meteorological forecast errors is more important, whereas for low streamflow forecasts the contribution of the hydrological model errors is more important. Initial conditions have 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 conditions (Table 5), making the contribution of the meteorological forecast errors larger.

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. 9a), at small lead times (1-5 days). Two effects contribute to this. First, long-rain floods and snowmelt floods are essentially driven by the water storage conditions in the catchment whereas for short-rain floods the meteorological input has more influence. Figure 9b confirms the relative importance of meteorological forecasts for this category. This results in a higher potential to generate forecast skill, already at small lead times. The increasing contribution of meteorological forecast errors in long-rain floods and snowmelt floods demonstrates that at larger lead times the accumulation of rainfall during the forecast period becomes important. Second, the short and heavy rain events preceding short-rain floods will be less well captured in historical meteorological measurements than the longer term processes generating long-rain floods and snowmelt floods. Long-rain floods are skilfully forecast from a lead time of 3 days and snowmelt floods are skilfully forecast from a lead time of 2 days. The forecast skills of short-rain floods and snowmelt floods decrease 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.

4.4.2 Low streamflow generating processes

Figure 10a shows that the low forecast skill of low streamflow originates from the forecasts of the events under the precipitation deficit conditions, whereas the forecast skill of low streamflow events under snow accumulation conditions is rather high. The low forecast skill of the low streamflow events under precipitation deficit conditions can be explained by the fact that precipitation deficits often occur in the same period of the year, due to climatic seasonality, and are therefore well captured by historical meteorological measurements. In addition, 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. The fact that the contribution of initial hydrological conditions at the forecast issuing day decreases for larger lead times (reflected in Fig. 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.

5 Discussion

The methodology was applied to an ensemble streamflow forecasting system of the Biała Tarnowska catchment, for a 6 year period. In this, findings of this study do not allow a direct generalisation but they contribute to ongoing discussions on improving streamflow forecasting. Also, a longer evaluation period would allow an evaluation of more extreme definitions of high and low streamflow.

The effectiveness of QM in pre- and post-processing depends on whether during the validation period the same bias exists between the CDF of the measurements and the CDF of the forecasts as exists during the training period. Figure 11 shows the large differences in the biases between the different years and between the training period and the validation period, which suggests that the bias is affected by randomness. The relatively short time series of the forecasts constrains the effectiveness of the pre- and post-processing, because different weather patterns cannot be well identified and with a longer period a more consistent bias distribution could be obtained. A problem in the pre- and post-processing of forecasts is that the joint distribution of measurements and forecasts is often non-homogeneous in time 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). In addition, the 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.

The rank histogram results show that ignoring uncertainties in the hydrological model and the model initial conditions affects the reliability of streamflow forecasts for short lead times and low streamflow in particular. Regarding the effect on short lead times, Bennett et al. (2014) and Pagano et al. (2013) reported similar findings. The lower flatness coefficients of high streamflow forecasts compared to low streamflow forecasts reflect the fact that for high streamflow forecasts the meteorological inputs are more important.

The classification of low and high streamflow generating processes is based on hydrometeorological information that is available from the measurement series and the HBV model (Table 2). Using this information provides more insight into the performance of the forecasting system than a seasonal characterisation. However, some assumptions must be kept in mind when interpreting the results. The assumption that snow accumulation before an event is embedded in the snowpack storage of the lumped HBV model neglects the fact that only part of the catchment may be covered by snow. If a snowpack is present, the event was classified as snowmelt flood or snow accumulation low streamflow. If no snowpack is present, it was assumed that the low streamflow event or high streamflow event is caused by low or high rainfall. The threshold of 10 mm day^{-1} is a simple rule 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 were caused by a combination of processes, which often occur in practice and result in the most extreme low and high streamflow events. Another point is that only information from the day preceding the forecast issuing day was used to classify the processes. The lag time between the precipitation events and the streamflow events does not always match the HBV model calculation time step and the classification rules used. Consequently, the streamflow on the day following a high rainfall event was 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 event was set at 1 day. However, the timing of a rainfall event on a day is important, particularly in a small catchment. The lag time is a critical aspect in the study's forecasting system, especially for short-rain floods. The ratio between the CRPS against observed meteorological input forecasts and the CRPS against streamflow measurements is above 100% for high streamflows, and short-rain floods in particular (Fig. 9b). This means that forecasts are closer to the measurements than to the observed meteorological input forecasts. On 28% of the high streamflow days at a lead time of 1 day to 48% of the high streamflow days at a lead time of 10 days, the ensemble forecasts are closer to the measurements than to the observed meteorological input forecasts. On 50% to 66% of these days, the ensemble forecasts are closer to the measurements than the observed meteorological input forecasts are. This indicates a hydrological model deficiency in high streamflow conditions, either from simulating the rainfall-runoff relation or the flood peak timing. The precipitation peak in the measurements and the precipitation peak in the meteorological forecasts may be shifted one day with respect to each other and this may cause that the timing of the peak of the streamflow forecasts better corresponds to the streamflow measurements. 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 measurements, while the peak day of the mean of the ensemble forecasts does match to the peak day of the measurements. This illustrates that the hydrological model deficiency regarding flood peak timing has a considerable effect on the observed meteorological input forecasts and the ensemble forecasts.

It is not trivial to compare the CRPS results to 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. Because Bennett et al. (2014) used the same alternative forecast set, the CRPSS results can be compared. Although Bennett et al. (2014) used a different forecasting system and applied it to different conditions, the forecast skills are comparable to the forecast skills obtained in this study.

6 Conclusions

We have developed a methodology that gives insight into the performance of 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 meteorological climatology, for different runoff generating processes. The system skilfully forecasts high streamflow events, although the skill depends on the runoff generating process and the 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. The sharpness of the forecasts is also good, although it decreases with lead time. Medium streamflow events and low streamflow events under precipitation deficit conditions are not skilfully forecasted.

- When this or any other forecasting system is (further) developed with the objective of generating more accurate high streamflow forecasts, it is recommended that the focus is on improving the meteorological forecast inputs because errors from the meteorological forecasts are dominant in high streamflow forecasts. This can be achieved by better meteorological forecasts (e.g. using the higher resolution forecasts from COSMO-LEPS (Renner et al., 2009)) or by improved pre-processing . The hydrological model performance on high streamflow conditions can be improved by specific calibration on flood peak timing and high streamflow conditions. To improve the low streamflow forecasts, it is recommended to focus on the hydrological model performance first. In this study, the calibration of the hydrological model was skewed to high streamflow conditions. An improvement of the low streamflow forecasts can be achieved by calibrating the hydrological model specifically on low streamflow conditions. Besides improvement of the hydrological model, further research should be done to improve the meteorological forecasts as input to low streamflow forecasts, especially to 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 must be reconsidered.
- The ensemble streamflow forecasting system shows good resolution and sharpness, but the reliability must be improved, particularly for the small lead times and the low streamflow forecasts. It is recommended to include the uncertainties in the hydrological model parameters and the initial conditions in the forecasting system. Because the precipitation and temperature forecasts are also underdispersed, we recommend an investigation into how the reliability of the precipitation and temperature forecasts can be improved, potentially by adding meteorological forecasts from other forecasting systems (i.e. creating ‘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 is not effective either. A longer time series of forecasts would promote the success of pre- and post-processing. ECMWF provides a homogeneous retrospective forecast set, consisting of twice-weekly forecasts with one control and 10 ensemble members over a period of 20 years, that is generated by the current operational system (Hagedorn, 2008; Vannitsem and Hagedorn, 2011; Vitart, 2017). 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, may improve the effectiveness of pre- and post-processing procedures.
- It is recommended that the study be extended 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 apply to the study catchment and the developed system set-up only, but the methodology of analysing an ensemble streamflow forecasting system is generally applicable. The methodology provides valuable information about the forecasting system; in which conditions it can be used and how the system can be improved effectively.

Acknowledgements

We thank Marzena Osuch (Institute of Geophysics Polish Academy of Sciences) and Adam Kiczko (Warsaw University of Life Sciences) for valuable discussions and support on methods. We thank Raymond MacDonald for his thorough check of the manuscript on English. We also acknowledge the editor Dimitri Solomatine and two anonymous reviewers whose comments helped improve this manuscript.

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.

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.

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.

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.

Hagedorn, R.: Using the ECMWF reforecast dataset to calibrate EPS forecasts, edited by B. Riddaway, ECMWF Newsl., 117, 8–13 [online] Available from: https://www.ecmwf.int/sites/default/files/elibrary/2008/14608-newsletter-no117-autumn-2008.pdf, 2008.

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.

Vannitsem, S. and Hagedorn, R.: Ensemble forecast post-processing over Belgium: comparison of deterministic-like and ensemble regression methods, Meteorol. Appl., 18(1), 94–104, doi:10.1002/met.217, 2011.

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.

Vitart, F.: ECMWF Model, S2S [online] Available from: https://software.ecmwf.int/wiki/display/S2S/ECMWF+Model (Accessed 3 July 2017), 2017.

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.

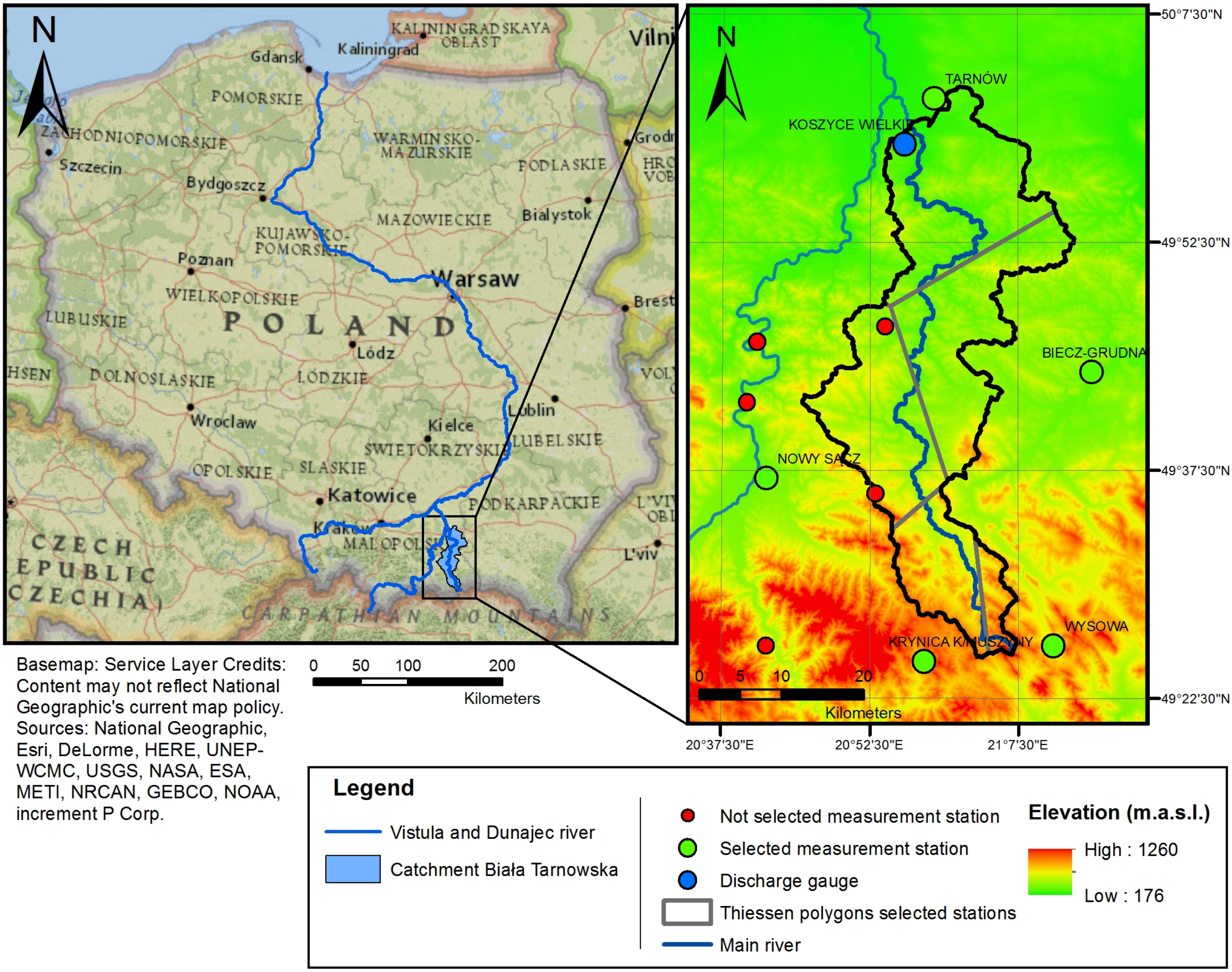
Ye, J., He, Y., Pappenberger, F., Cloke, H. L., Manful, D. Y. and Li, Z.: Evaluation of ECMWF medium-range ensemble forecasts of precipitation for river basins, Q. J. R. Meteorol. Soc., 140(682), 1615–1628, doi:10.1002/qj.2243, 2014.

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



Basemap: Service Layer Credits: Content may not reflect National Geographic's current map policy. Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp.

Figure 1: Location and overview of the Biała Tarnowska catchment

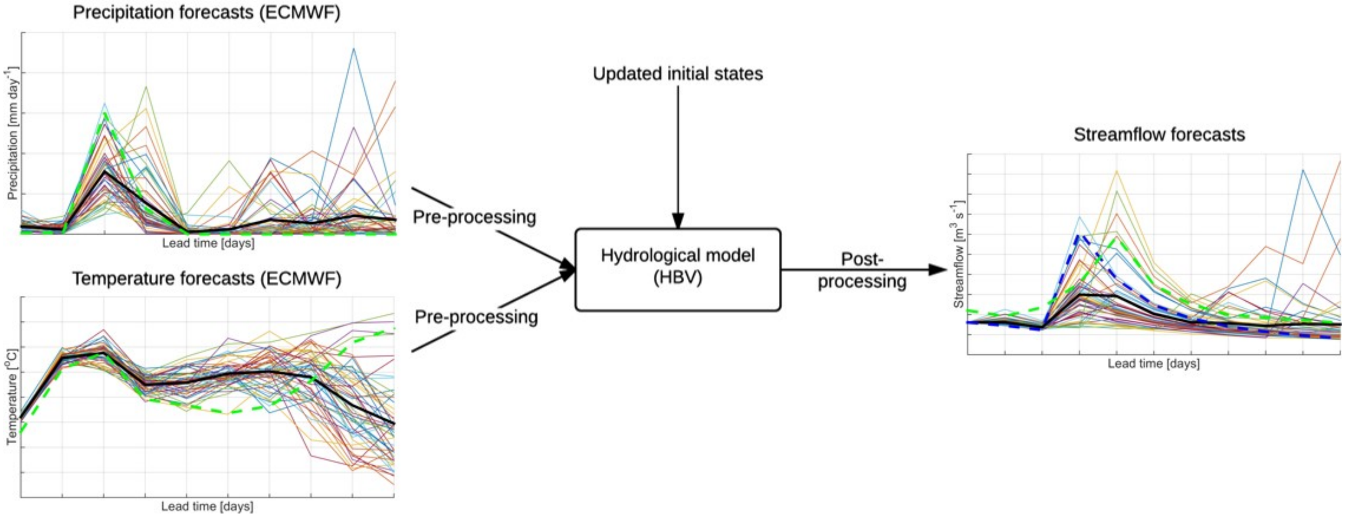
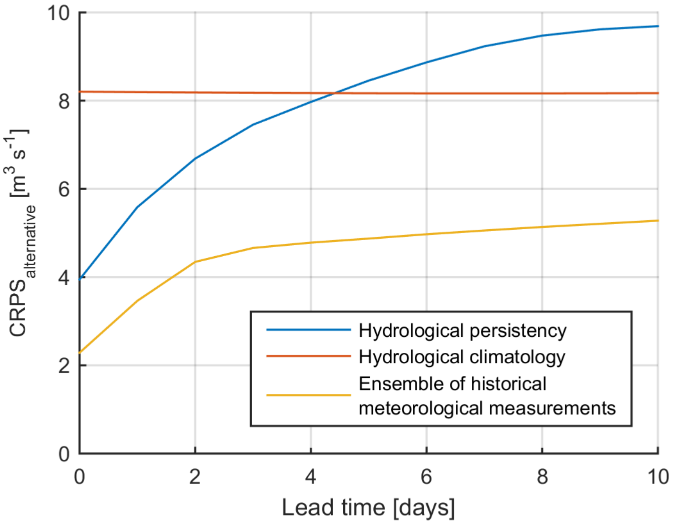


Figure 2: Structure of the ensemble streamflow forecasting system



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

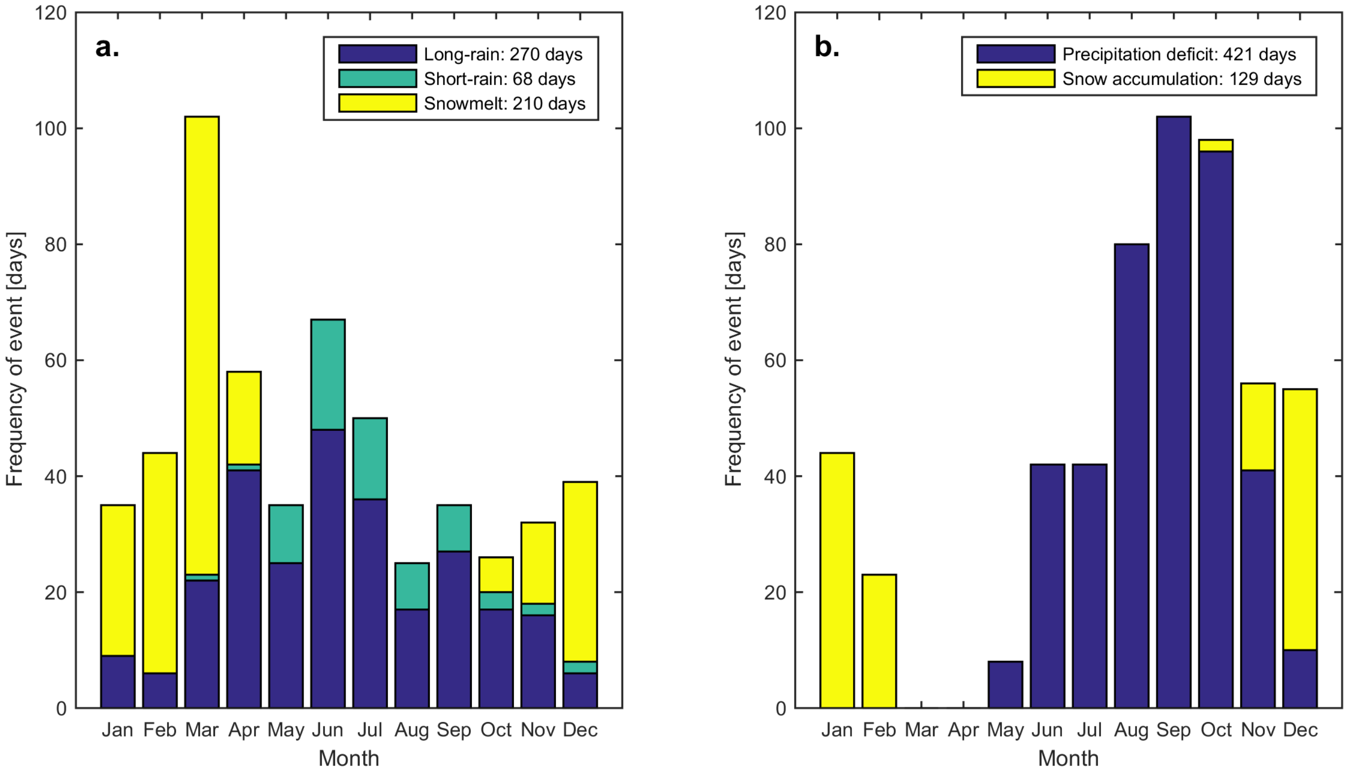
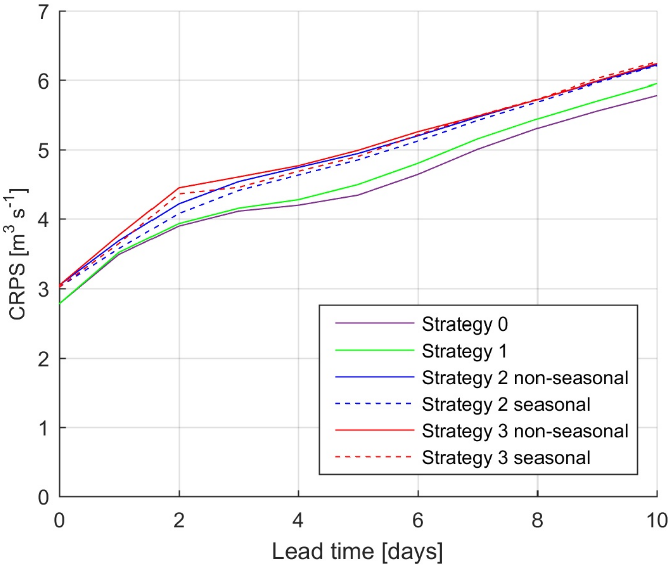


Figure 4: a. High streamflow generating processes over the year b. Low streamflow generating processes over the year, 1-11-2007 to 31-10-2013



5

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.

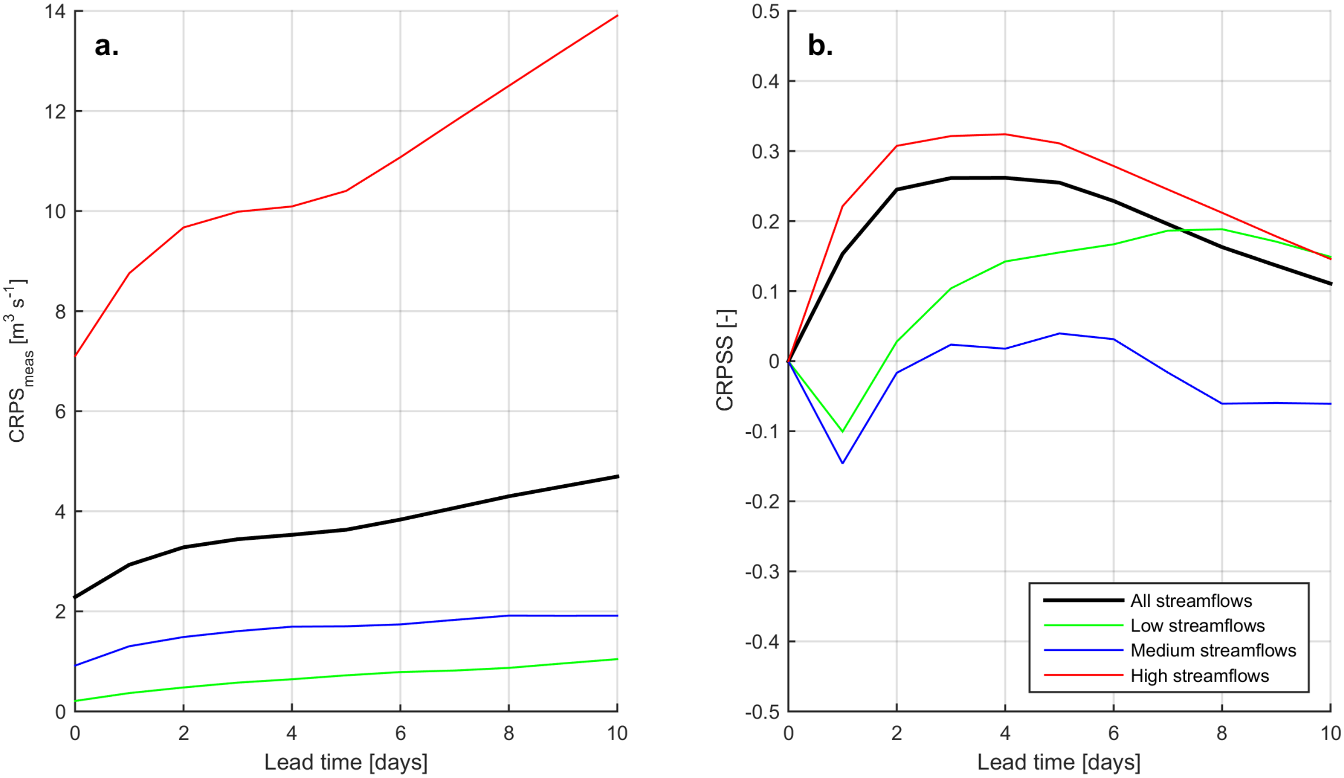


Figure 6: a. Streamflow forecasts evaluated against streamflow measurements b. Skill of the streamflow forecasts, defined in Eq. (1)

5

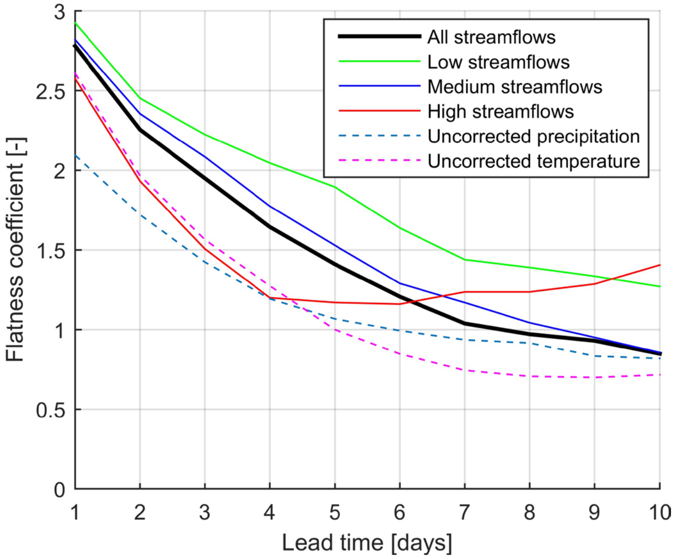


Figure 7: Rank histogram flatness coefficients. The flatness coefficients of the precipitation and temperature forecasts refer to the preceding day.

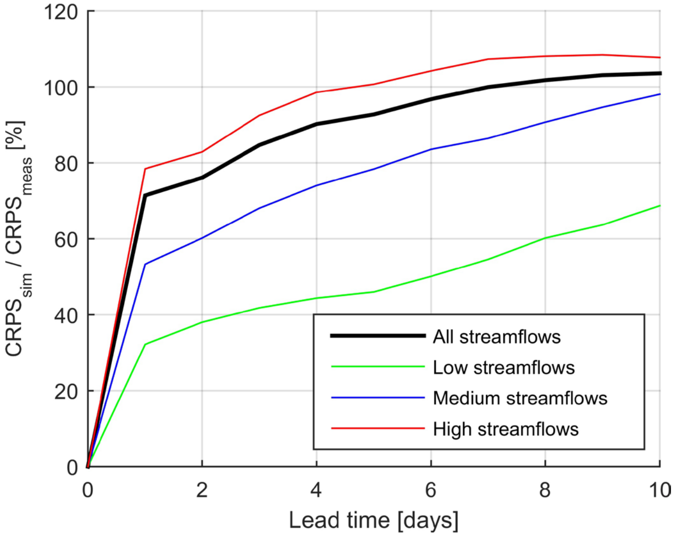


Figure 8: Ratio of errors in meteorological forecasts (CRPS$_{sim}$) to meteorological forecast + model errors (CRPS$_{meas}$)

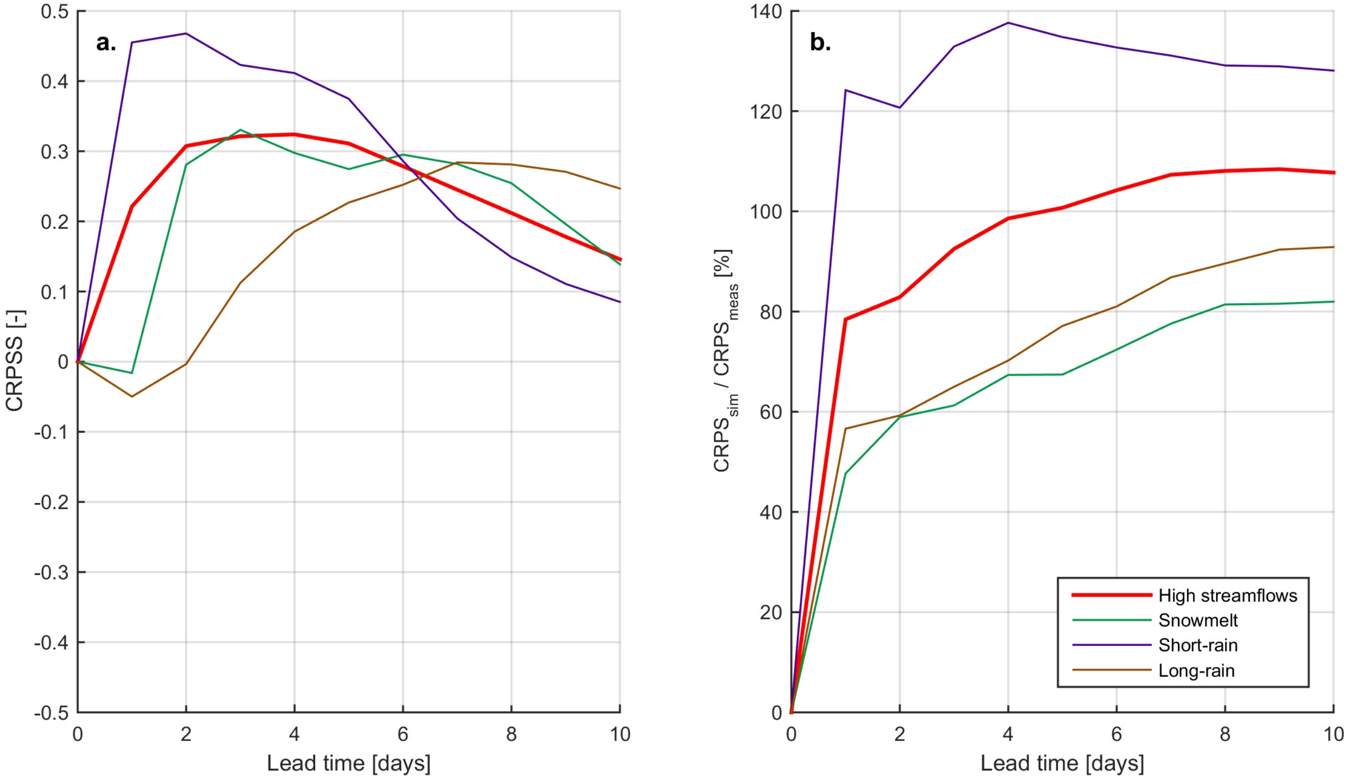


Figure 9: a. Forecast skill of high streamflow generating processes b. Ratio of errors in meteorological forecasts (CRPS$_{sim}$) to meteorological forecast + model errors (CRPS$_{meas}$).

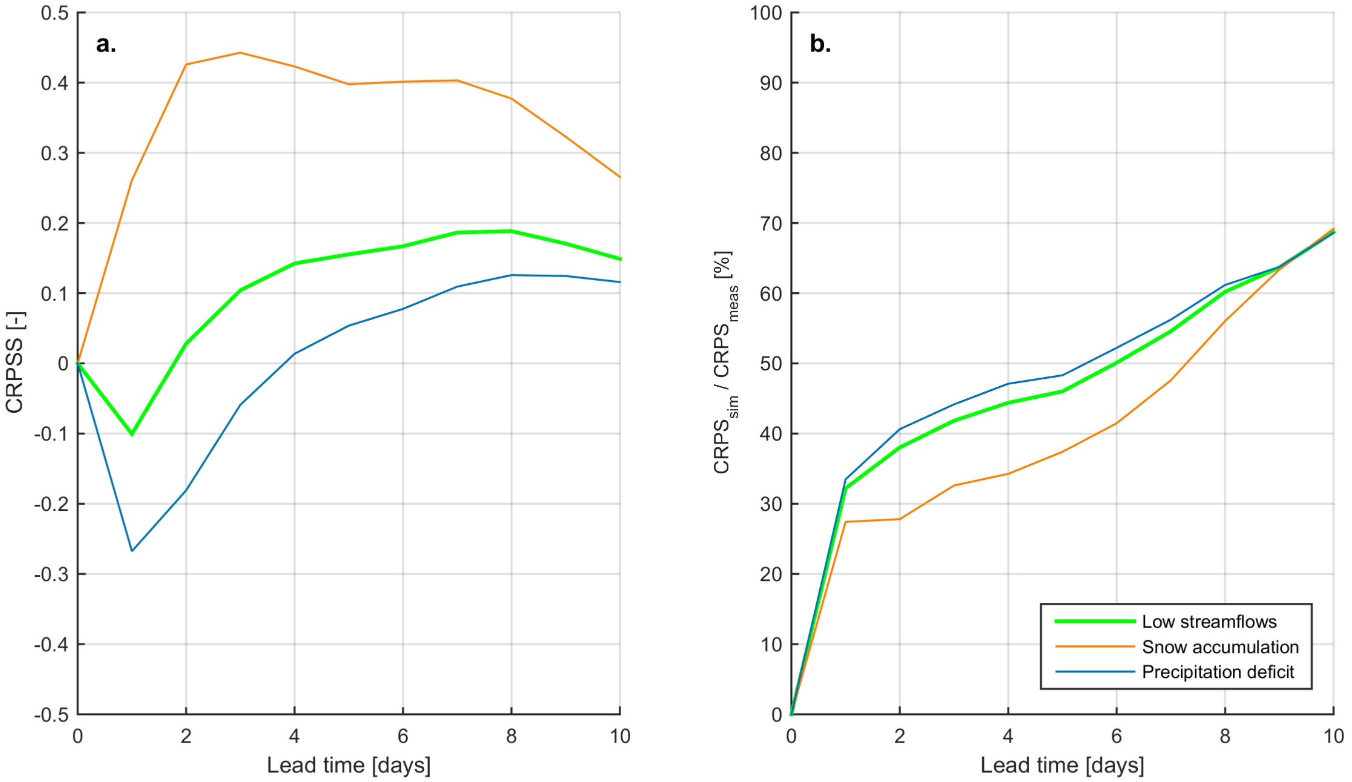


Figure 10: a. Forecast skill of low streamflow generating processes b. Ratio of errors in meteorological forecasts (CRPS$_{sim}$) to meteorological forecast + model errors (CRPS$_{meas}$).

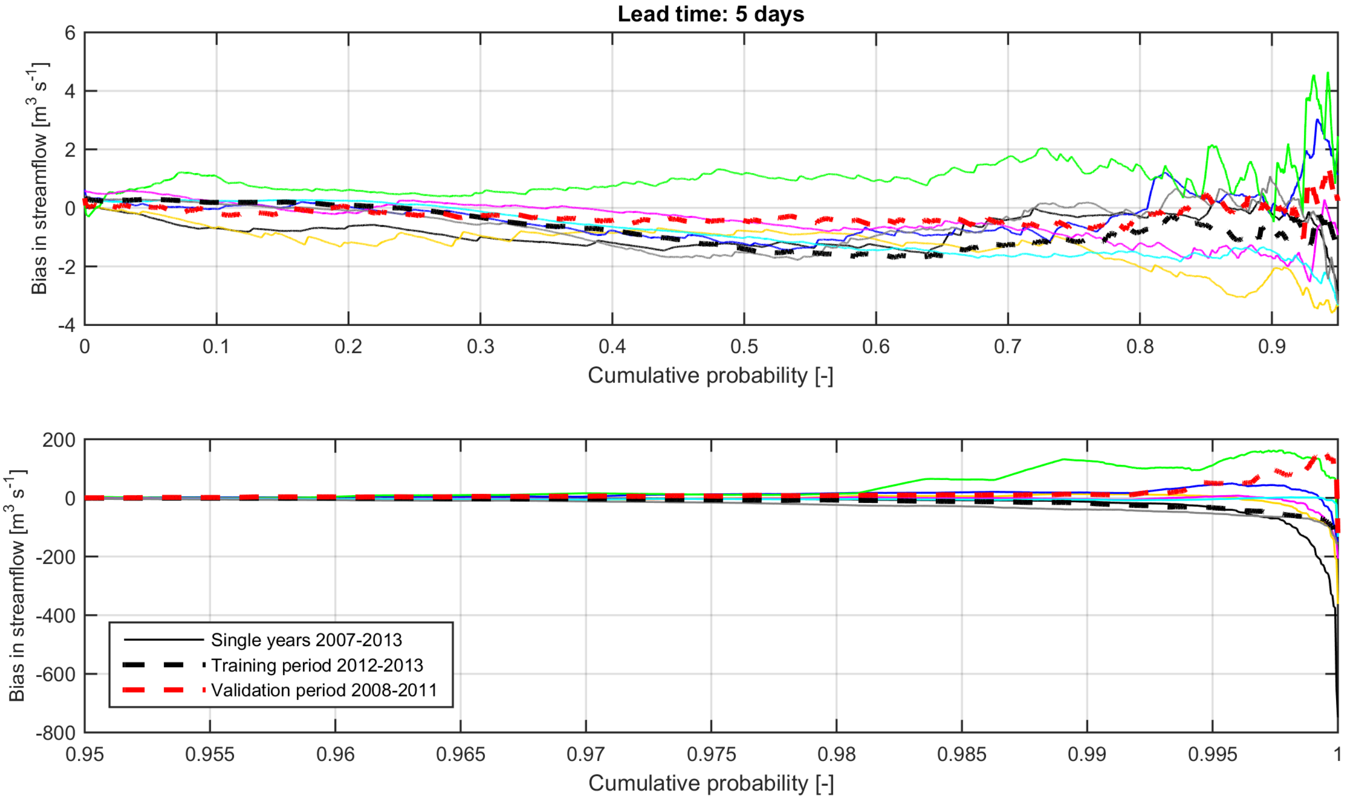


Figure 11: Difference between CDFs of the measurements and CDFs of the uncorrected streamflow forecasts per hydrological year (upper panel cumulative probability 0 – 0.95 and lower panel 0.95 – 1.0). Each thin line refers to a single year between 2007 and 2013. This figure is for a lead time of 5 days.

5

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 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 day-1 • Rainfall at 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 day-1 • Rainfall at 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 day-1
Precipitation deficit	When low rainfall and high evapotranspiration last over a prolonged period the catchment will dry out.	• No snowpack (HBV) at 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 run with input data corrected for elevation	0.81	0.81	0%	0.75	0.78	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 (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