Performance and limitations of ensemble river flow forecasts

Master's thesis

H. F. Benninga
August 2015

UNIVERSITY OF TWENTE.

Frontpage pictures: Biała Tarnowska river, taken by Krzysztof Hankus (MGGP S.A.)

Performance and limitations of ensemble river flow forecasts

Master's thesis in Civil Engineering and Management

University of Twente

Faculty of Engineering Technology

Department of Water Engineering & Management

Author: **H.F. Benninga BSc**
h.f.benninga@alumnus.utwente.nl

Graduation Committee: **Dr. ir. M.J. Booij**
University of Twente, Faculty of Engineering Technology, Department of Water Engineering and Management

Dr. ing. T.H.M. Rientjes
University of Twente, Faculty of Geo-Information Science and Earth Observation, Department of Water Resources

Prof. dr. hab. R.J. Romanowicz
Institute of Geophysics Polish Academy of Sciences, Department of Hydrology and Hydrodynamics

Location and date: Enschede, August 2015

Summary

High and low flows may cause several problems to society. Flood forecasting, low flow forecasting and hydrological forecasting in general are important to mitigate the negative consequences of extreme flow events and for economic use of a river. Ensemble prediction systems are increasingly used for hydrological forecasting. These systems provide an ensemble of forecasts for each forecast period instead of a single, deterministic forecast. There have been various studies on ensemble flow forecasting, but these studies have mainly focused on large river catchments and exclusively on flood forecasting or low flow forecasting. The objective of this study is to develop an ensemble flow forecasting system for the Biała Tarnowska catchment (~1000 km^2) in Poland and to investigate the performance of this system for lead times from 1 to 10 days, for low, medium and high flows and for different hydrological circumstances.

The ensemble flow forecasting system consists of a deterministic lumped hydrological (HBV) model with input data in the form of ensemble precipitation and temperature forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). The meteorological ensemble forecast data from ECMWF consists of 1 control forecast and 50 perturbed ensembles. The deterministic calibration of the hydrological model has been based on observed precipitation, temperature and discharge data. It turned out that pre-processing of the precipitation forecasts with Quantile Mapping is best when it is applied separately for each lead time. For the temperature forecasts the best results are obtained if in addition seasonal distinction in a summer and winter season is applied. However, the best flow forecasts are obtained when no pre- or post-processing with Quantile Mapping is applied at all. Therefore no pre-processing of meteorological forecast and no post-processing of flow forecasts has been applied. To improve the representation of the current situation in the catchment at the forecast day, the initial conditions in the hydrological model are updated based on discharge observations at one day before the forecast day.

The performance of the flow forecasts deteriorates with lead time. The skill of the flow forecasts is determined with respect to the best reference forecast set, which are flow forecasts based on an ensemble of historical observations of precipitation and temperature on the same calendar day over the past 20 years. In general the skill of the flow forecasts is positive and maximum between lead times of 2 and 5 days, but this is very different for the low, medium and high flow forecasts. The low flow forecasts do not have skill until a lead of 2 days and after that they show a small positive skill. The medium flow forecasts do not provide skill for all lead times. The highest skills are obtained for the high flow forecasts. This has to do with the performance that historical observations of precipitation and temperature on the same calendar day provide for these flow categories and that the same initial conditions are used to generate the ensemble flow forecasts and the reference forecasts. Since in low flows initial conditions are more important, it is more difficult for the ensemble flow forecasting system to deviate from the reference forecasts and thus to be able to generate skilful flow forecasts.

The forecast skill is also very different for different high flow and low flow producing processes. Regarding high flow forecasts the highest skill is obtained for the short-rain floods, but the skill decreases considerably for lead times larger than 5 days. Long-rain floods and snowmelt floods are more dependent on the initial conditions in the catchment, which leads to small forecast skills at short lead times. From a lead time of respectively 3 days and 2 days also long-rain floods and

snowmelt floods are skilfully forecasted. The low skill of low flow forecasts is mainly caused by low rainfall/high evapotranspiration generated low flow forecasts, while the skill of snow accumulation generated low flows is relatively high. These results provide information about the system and in which situations it can be used to generate skilful flow forecasts.

The performance of the flow forecasting system has also been researched on different properties of forecast quality. The sharpness of the forecasts is good, because forecast probabilities of low and high flows are most often close to 0 or 1, instead of forecast probabilities close to the mean probability. The resolution is also good, with high hit rates compared to false alarm rates for high and low flow forecasts. However, the reliability of the system is not good, particularly for small lead times. To improve the reliability of the ensemble flow forecasts it is recommended to also include hydrological model parameter and initial condition uncertainty, or to improve post-processing of the flow forecasts. In addition it is recommended to do further research to improve the reliability of the precipitation and temperature forecasts.

The relative contribution of meteorological forecast errors and hydrological model errors (including initial conditions) has been researched to give recommendations about how the ensemble forecasting system can be improved effectively. In general the relative contribution of meteorological forecast errors increases with lead time and the relative contribution of hydrological model errors decreases with lead time. Regarding the different flow categories, when the objective of further research is to improve the high flow forecasts it is recommended to focus further research mainly on improving the meteorological forecasts, because in high flow forecasts errors from the meteorological forecasts are relatively more important. When the objective is to improve the low flow forecasts it is recommended to focus further research at first mainly on the hydrological model performance. The calibration was skewed to high discharges, so it is expected that an easy improvement of the forecasts can be achieved when the hydrological model would be calibrated on low flow situations. Besides improvement of the hydrological model, further research should be done to improve the meteorological forecasts.

After all, it is recommended to extend the research to other catchments and (if possible) with a longer period of data, to be able to draw more general conclusions and to test more extreme high and low flow thresholds before the system is potentially applied operationally. In addition, it is recommended to incorporate statistical tests for the evaluation scores to increase confidence in the conclusions.

Preface

This report presents the final project of my master study Water Engineering and Management, a specialization of Civil Engineering and Management at the University of Twente. In this research I have set up a system to forecast river discharges and I have investigated this system for different purposes and hydrological circumstances. During the Master's thesis project I had the opportunity to investigate this interesting topic in much detail, much deeper than this was possible during the regular courses. In addition I have learned a lot about carrying out a scientific research.

The idea of this research has been developed in cooperation with Renata Romanowicz from the Institute of Geophysics Polish Academy of Sciences (IGF PAN) and also the Master's thesis has been conducted in cooperation with IGF PAN. I had the opportunity to work 10 weeks on my Master's thesis at IGF PAN in Warsaw. I am very grateful for the welcome I got from all people of IGF PAN and for the very interesting and nice time that I had at the institute and in Poland. I would like to thank all people from IGF PAN for this great time and I especially would like to thank Renata and Marzena for all your ideas, the questions that I could ask, the detailed feedback from you and the time that you made free for me to help me with my research.

I would also like to thank my supervisors at the University of Twente. Martijn, thank you for the fact that I could always come along with some questions and for the thorough feedback on my report. Tom, thank you for your ideas, your feedback and for the suggestions to structure this research during the preparation phase of the project.

After all I would like to thank my roommates at the WEM graduation room, my friends and my family for their support during the last couple of months and during my complete study at the University of Twente.

Harm-Jan Benninga
Enschede, 20 August 2015

Table of contents

List of figures

List of tables

List of abbreviations

AUC	Area under the ROC curve
CDF	Cumulative distribution function
COSMO-LEPS	COnsortium for Small-scale MOdeling Limited Area Ensemble Prediction System
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
DBM	Data Based Mechanistic methodology
DE	Differential Evolution
DEGL	Differential Evolution with Global and Local neighbourhoods
ECMWF	European Centre for Medium-Range Weather Forecasts
EFAS	European Flood Alert System
GCM	General Circulation Model
GLUE	Generalized Likelihood Uncertainty Estimation
HBV	Hydrologiska Byråns Vattenbalansavdelning
IGF PAN	Institute of Geophysics Polish Academy of Sciences
MAE	Mean Absolute Error
MSC	Meteorological Service of Canada
NCEP	National Centers for Environmental Prediction
NS	Nash-Sutcliffe coefficient
QM	Quantile Mapping
RCI	Relative Confidence Interval
RCM	Regional Climate Model
RMAE	Relative Mean Absolute Error
ROC	Relative Operating Characteristic
RVE	Relative volume error
SCEM-UA	Shuffled Complex Evolution Metropolis algorithm
THORPEX	The Observing System Research and Predictability Experiment
TIGGE	THORPEX Interactive Grand Global Ensemble
WMO	World Meteorological Organization

1. Introduction

This chapter provides an introduction to the research. At first background information and relevance of the topic is described in section 1.1. Previous research is discussed in section 1.2. In section 1.3 and section 1.4 the research gap and research objective are defined. In section 1.5 follow the research questions and in section 1.6 the methodology and reading guide are presented.

1.1 Background and relevance

Floods and low flows cause several problems to society. Flood damage mitigation can be provided by structural and non-structural strategies (Carpenter et al., 1999; Yazdi et al., 2014). Yazdi et al. (Yazdi et al., 2014) recommend non-structural strategies instead or besides structural measures, because of insufficiency of structural methods for flood damage reduction and because of economic and environmental advantages of non-structural approaches. Yazdi et al. (Yazdi et al., 2014) mention watershed management techniques, flood forecasting and warning, flood insurance and flood management in reservoirs as frequently used non-structural measures for flood mitigation. Providing early flood warnings is an effective strategy in reducing flood damage, intangible impacts (stress) and loss of life due to flooding (Penning-Rowsell et al., 2000; Werner et al., 2006), because it gives civil protection authorities and the public more preparation time and thus reduces the impacts of the flood (Cloke & Pappenberger, 2009; Werner et al., 2006). Flood protection have risen on the political agenda over the last decade, for example in the Netherlands (Ministerie van Verkeer en Waterstaat et al., 2009). According to the Ministerie van Verkeer en Waterstaat et al. (2009), especially along rivers flood warning can be an effective component in the policy of flood protection, because a flood can be predicted some days before the event.

Besides floods also low flows regularly appear in rivers and these may also negatively affect several river functions like water supply, power production, navigation, ecological protection and water quality control (Demirel et al., 2013a; Fundel et al., 2013; A. Ye et al., 2015). To anticipate to possible low flow events, low flow forecasts up to 10 days ahead are essential (Demirel et al., 2013a). Low flow forecasts are usually investigated on a seasonal time scale, but forecasts on a monthly or lower time scale are also relevant (Fundel et al., 2013). Forecasting medium discharges is relevant for water supply, generating hydropower and controlling procedures for reservoirs and dams.

It can be concluded that flood forecasting, low flow forecasting and flow forecasting in general are relevant to mitigate the negative consequences of extreme flow events and for economic use of a river. This is increasingly important because as a result of global climate change it is expected that the frequency and intensity of low flow and high flow events will increase in many areas in the world (Intergovernmental Panel on Climate Change, 2014) and due to economic development also the economic consequences increase (Ministerie van Verkeer en Waterstaat et al., 2009). A warning must be provided at such a time before the event that it is possible to take responsive actions (Werner et al., 2006). Medium-range forecasting, with lead times up to 10 days (European Centre for Medium-Range Weather Forecasts (ECMWF), 2012; Werner et al., 2006), provides appropriate lead times for this purpose.

Hydrological forecasting systems are more and more implemented as ensemble prediction systems (Cloke & Pappenberger, 2009). These systems provide an ensemble of river flow forecasts for each forecast period, which can be understood as a probabilistic forecast of future river flows, instead of relying on one deterministic forecast (Cloke et al., 2013). Advantages of ensemble flow forecasts are:

- *Extended lead times:* Meteorological forecasts are required to obtain hydrological forecasts with lead times longer than the concentration time of the catchment (Cloke & Pappenberger, 2009; Werner et al., 2006). Meteorological forecasts are nowadays often provided in the form of ensemble forecast data, from meteorological services like ECMWF, the Meteorological Service of Canada (MSC) and the National Centers for Environmental Prediction (NCEP) (Buizza et al., 2005).
- *Account for uncertain information in flow forecasts:* Traditionally low flow (Demirel et al., 2013a) and high flow forecasts (Verbunt et al., 2007) are given as one deterministic value, even though there is uncertainty in the forecast. There is an increasing interest to account for uncertain information in decision support systems (Cloke & Pappenberger, 2009; Demirel et al., 2013a).
- *Earlier information about the possibility of an event:* With an ensemble flow forecast there is earlier knowledge about the *possibility* of a flood event and there can be anticipated to the possible flood event by the development of different flood scenarios, responses and actions and reinforced monitoring of the meteorological and hydrological conditions over the coming days (Thielen et al., 2009a). If one or a few ensemble members forecast a high flow event this indicates the possibility of a flood, while the ensemble mean forecast or deterministic forecast might still be below the warning threshold.

1.2 Previous research

Various previous studies have investigated hydrological ensemble forecasting systems. An important flood forecasting system is the European Flood Alert System (EFAS), which is an operational ensemble flood prediction system at a large-scale European level. The objective of this system is to predict high flows in large European rivers basins for lead times between three and ten days, so that preventive measures can be taken to reduce the consequences (Thielen et al., 2009a). On average, skilful predictions are found over the whole 10-day forecast range, increasing with increasing upstream area (Alfieri et al., 2014). Alfieri et al. (2014) attribute this to the more gradually varying discharge in large river catchments and the influence of initial discharge compared to forecasted precipitation input is larger in large catchments than in smaller catchments (more water is already in the system). There is a clear deterioration of scores for catchments smaller than 300 km^2, although in practice EFAS uses a minimum upstream area of 4000 km^2 to issue flood alerts to partner institutes (Alfieri et al., 2014). Thielen et al. (2009a) describe that there can be large discrepancies between model results and discharge observations, so critical flood warning thresholds are based on simulated discharges instead of thresholds directly extracted from observations. EFAS is aimed at providing early warnings of possible flooding, not to provide specific river flow forecasts (Demeritt et al., 2013). De Jong et al. (2012) argue that EFAS is not meant to provide precise estimations of discharge, because this is something that countries can better do by themselves. It can be concluded that the system should be used at the level at which it has been built, namely at large-scale European level. To provide detailed forecasts of discharge, smaller scale ensemble flow predictions systems are needed.

There have been several studies on smaller scale ensemble flow prediction systems (most of them not operational). Roulin and Vannitsem (2005) developed an ensemble flow prediction system for two Belgian catchments, using a water balance model with meteorological forecasts from ECMWF as input, and evaluated this system for high discharges. They found that the skill (compared to

reference forecasts based on historical precipitation observations) of ensemble streamflow forecasts is greater than the skill of precipitation forecasts and that the flow forecasts remain skilful beyond a lead time of 9 days, although skill is decreasing with increasing lead time. After all they found that during the winter period high discharge forecasts are more skilful than during the summer period. Precipitation forecasts were also more skilful during winter than during summer (Buizza et al., 1999; Roulin & Vannitsem, 2005). Also other studies investigated the performance of ensemble forecasts for different lead times, like Ye et al. (2014) for ECMWF's medium-range ensemble precipitation forecasts, Thielen et al. (2009b) and Alfieri et al. (2014) for the EFAS system, Renner et al. (2009) for flow forecasts for the River Rhine, Olsson and Lindström (2008) for flow forecasts for various catchments in Sweden and Bennett et al. (2014) for various catchments in Australia, which all found a deterioration of performance with increasing lead time. Demirel et al. (2013a) investigated the effect of different uncertainty sources on the skill of medium-range (until 10 days) ensemble low flow forecasts. They concluded that hydrological model parameter uncertainty has the largest effect and input uncertainty has the smallest effect on the uncertainty in low flow forecasts. On the other hand Werner et al. (2006) argue that if meteorological forecasts are used to extend the forecast lead time, the uncertainty in the precipitation forecasts will dominate uncertainties in the flow forecasts.

In the field of seasonal hydrological ensemble forecasting the limitations and uncertainties of forecasts has been researched extensively (e.g. (H. Li et al., 2009; Paiva et al., 2012; Yossef et al., 2013)). The underlying idea of these studies was that it is important to know the role of initial condition errors and meteorological forcing errors for the practical design of a forecasting system and where to focus on with further research to improve these systems.

Some studies investigated the contribution of errors in the meteorological forecast data and the hydrological model in errors of the medium-range hydrological forecasts. From the study of Demargne et al. (2010) follows that hydrological model uncertainty (initial conditions, model parameters and model structure) is more significant for short lead times and more important in low flow forecasts than in high flow forecasts. Renner et al. (2009) found an underprediction of high flows, while Olsson and Lindström (2008) found an overprediction of high flows. Both studies conclude that this bias mainly originates from the meteorological forecasts. Renner et al. (2009) also found an overprediction of low discharge, which they contribute to both the hydrological model and the meteorological input data. After all, Olsson and Lindström (2008) found an underestimation of the spread of hydrological ensemble forecasts and they concluded that the meteorological forecasts and the hydrological model have a comparable contribution to this.

1.3 Research gap

The studies that are mentioned in the previous paragraph are a selection of studies that has been done in this research field. One of the research directions is the EFAS system. However, this system is not suitable to generate detailed forecasts for the entire hydrograph (low, medium and high flows), as described in section 1.1, and for small river catchments.

Cloke and Pappenberger (2009) and Cloke et al. (2013) mention that more quantitative studies on hydrological ensemble prediction systems are required. Previous studies on medium-range ensemble flow forecasting have mainly focused on flood forecasting (i.a. Alfieri et al., 2014; Bürger et al., 2009; Olsson & Lindström, 2008; Roulin & Vannitsem, 2005; Thielen et al., 2009a, 2009b) or low flows (Demirel et al., 2013a; Fundel et al., 2013). There are some studies (Demargne et al., 2010; Renner et

al., 2009) that include all flow categories, but they do not explicitly address performance under different hydrological circumstances. In addition, previous research has mainly focused on flow forecasts at a large spatial scale. Generating flow forecasts at a smaller spatial scale is relevant for regional problems, like local floods and the operation of dams, and potentially as building blocks for larger hydrological systems. In a relatively small river catchment short and local meteorological processes play a more important role and it is interesting to investigate whether these processes can be captured in medium-range flow forecasting systems.

In studies on seasonal hydrological ensemble forecasting by Li et al. (2009), Paiva et al. (2012) and Yossef et al. (2013) the sources of errors in hydrological forecasts are investigated to give a recommendation about how to improve these forecasts. Previous studies on medium-range ensemble flow forecasting have linked the contribution of errors from meteorological forecast data and the hydrological model to different flows (high and low flows), but this has not been linked to different underlying processes for high and low flows.

1.4 Research objective

The objective is to investigate the performance and limitations of ensemble flow forecasts in a relatively small river catchment, by setting-up an ensemble flow forecasting system for the Biała Tarnowska catchment (~1000 km^2) with meteorological ensemble input data and investigating the performance of this system for different purposes and hydrological circumstances.

1.5 Research questions

The research objective is broken down into the following research questions:

1. *What is the most appropriate set-up of input data, the hydrological model and the calibration procedure to obtain an ensemble flow forecasting system for the Biała Tarnowska catchment?*
   This research question is aimed at developing the ensemble flow forecasting system to generate good flow forecasts.

2. *How does the ensemble flow forecasting system perform for different purposes and how does this relate to errors from meteorological input data and the hydrological model?*
   In this research question is focused on the use of the system for different purposes. Purposes that will be investigated are lead times from 1 day to 10 days ahead and different flow categories (low, medium and high flows).

3. *How does the ensemble flow forecasting system perform for different hydrological circumstances and how does this relate to errors from meteorological input data and the hydrological model?*
   There are different processes underlying low flows and high flows and the question is how the system performs for these processes. While research question 2 focuses on the use of the system, this is a scientific question that should provide more insight into the performance of the system under different hydrological circumstances. This provides valuable information about the system and in which situations it can be used.

1.6 Research methodology and report outline

In Figure 1 the research scheme is presented, showing the activities that will be carried out in this study. In research question 1 an ensemble flow forecasting system will be developed which should be able to reliably forecast low flows, medium flows and high flows based on meteorological forecasts. After the ensemble flow forecasting system has been developed the results will be investigated for different purposes (research question 2) and hydrological circumstances (research question 3).

The same approach that is used in the mentioned studies on seasonal hydrological forecasting will be applied, so in the first place it is investigated to which purposes and circumstances the ensemble flow forecasting system is limited for skilful forecasts and in the second place it is investigated what the dominant error source is (input or model) to give recommendations about how to improve the system effectively.

By using meteorological ensemble forecasts the uncertainty of meteorological input is incorporated. This is one of the most popular ways to generate ensemble flow forecasts (Cloke et al., 2013). Other sources of uncertainty are hydrological model parameters, initial conditions and model structure (Zappa et al., 2011). It is often assumed that the uncertainty of meteorological forecasts is the largest source of uncertainty beyond 2-3 days (Bennett et al., 2014; Cloke & Pappenberger, 2009), and therefore only meteorological forecast uncertainty is incorporated in many studies (Bennett et al., 2014). This is however also disputed in literature, for example by Cloke and Pappenberger (2009), Demirel et al. (2013a) and Zappa et al. (2011). In this study only uncertainty from the meteorological forecasts will be incorporated to focus the research on the effect of ensemble meteorological forecasts on flow forecasts.

The study will be applied to the Biała Tarnowska catchment, located in a mountainous part of southern Poland (Napiorkowski et al., 2014). The catchment has an area of about 1000 km^2. This is a suitable catchment for this research, because there is a large variation in flows and different processes are taking place in this catchment (rainfall and snowmelt related processes).

In Figure 1 it is also mentioned where the methods and results are described in the report. In chapter 2 the study area and data are described. Chapter 3 explains the methods that are used in this project, including the choices for several techniques. The results are analysed in chapter 4, in chapter 5 follows the discussion and in chapter 6 the conclusions and recommendations are given.

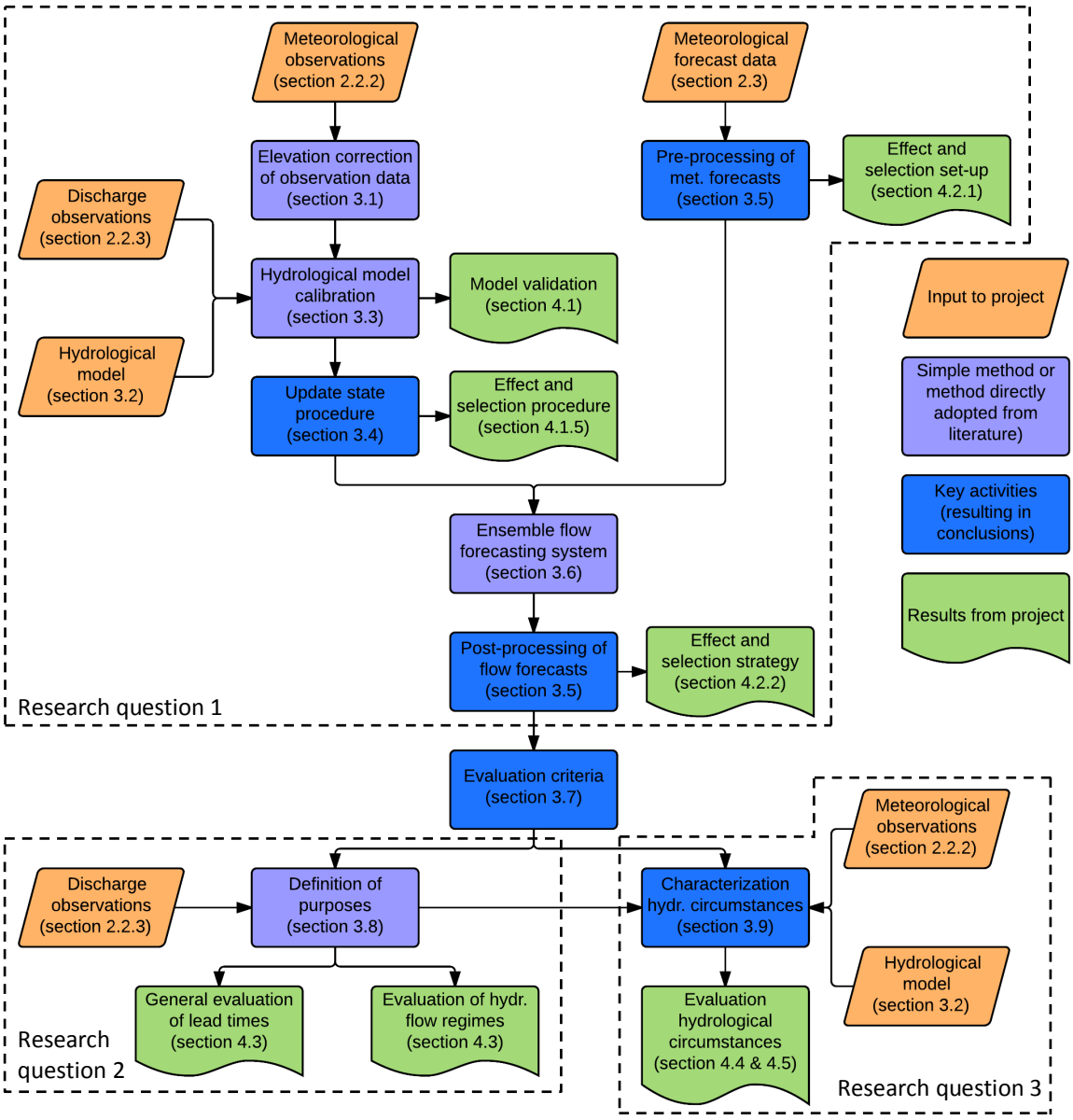


Figure 1: Research scheme and structure of the report

2. Study area and data

In section 2.1 the study area is characterized and in section 2.2 the observation data are described. The meteorological forecast data are introduced in section 2.3.

2.1 Study area

The Biała Tarnowska catchment has been selected as study area. This is a suitable study catchment, because different hydrological processes are taking place in this catchment during the year. Both snowmelt and precipitation are important driving mechanisms for discharge. In addition, the Institute of Geophysics Polish Academy of Sciences (IGF PAN) has used this catchment in earlier studies, precipitation, temperature and discharge measurement data are available and IGF PAN already has a hydrological model and ensemble forecast data available for this catchment (Kiczko et al., 2015). Figure 2 presents an overview of the Biała Tarnowska catchment. The Biała Tarnowska river catchment is located in a mountainous part of southern Poland (Napiorkowski et al., 2014). It is a sub-catchment of the River Dunajec, which is a tributary of the River Vistula. The total length of the river is 101.8 km and the catchment area is 956.9 km^2 (Napiorkowski et al., 2014). The average elevation is 376 m.a.s.l., but this varies considerably over the catchment. Napiorkowski and Piotrowski (2014) describe the river and the catchment. The majority of the river has unregulated banks and is in a natural state. The southern area of the catchment (about 25% of the catchment area) is a wooded mountainous part with an average river slope of 10‰. The northern area has deep river valleys and in general this area is deforested. The river slope in the northern part is in the range of 0.9-5‰. The Biała Tarnowska catchment is characterized by high precipitation amounts and a large variation of runoff during summer (Kiczko et al., 2015). During winter and spring snow plays an important role.

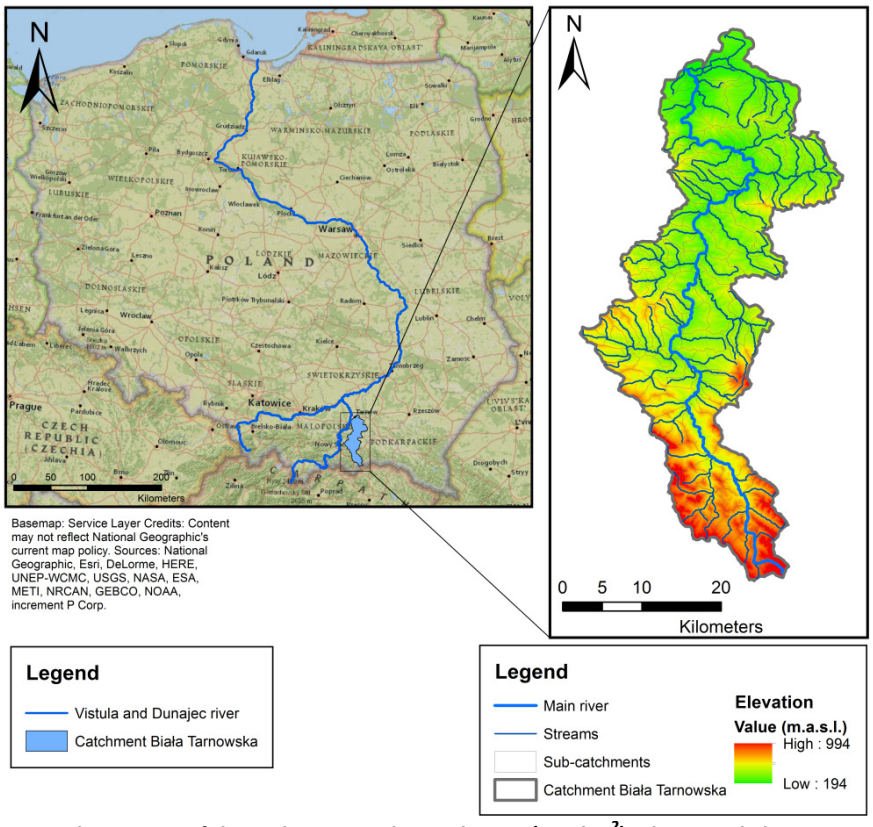


Figure 2: Location and overview of the Biała Tarnowska catchment (957 km^2). The Digital Elevation Model has been constructed by the Geographic Survey of the Polish Army

2.2 Observation data

The meteorological and discharge observation data are provided by the Institute of Meteorology and Water Management in Poland. A description of the measurement stations and how the observation data are aggregated to catchment-average data is presented in section 2.2.1. The meteorological observation data are characterized in section 2.2.2 and the discharge observations are characterized in section 2.2.3.

2.2.1 Measurement stations

The meteorological input data to the hydrological model, in the form of precipitation and temperature, originate from 5 measurement stations (Figure 3). These measurement stations have been selected because of coverage and data completeness. However, none of them is situated in the river catchment itself, which might degrade a reliable application of this input data. Since the catchment is relatively small it is expected that there are no large differences due to location. There can be differences due to elevation, for which a correction will be introduced in section 3.1. In Figure 3 other measurement stations are indicated in or close to the catchment, but they have shorter time series and/or contain gaps (Osuch, personal communication, 2015). The measurement data from the 5 measurement stations are aggregated to catchment-average precipitation and temperature by weighting factors based on Thiessen polygon surface areas (also indicated in Figure 3). Table 1 and Table 2 include the weightings of the measurement stations. The discharge measurements are provided by a water level measurement station at the downstream end of the river and water levels are translated into discharges using rating curves. Measurements are available for the period 1-1-1971 to 31-10-2013 at a daily time interval.

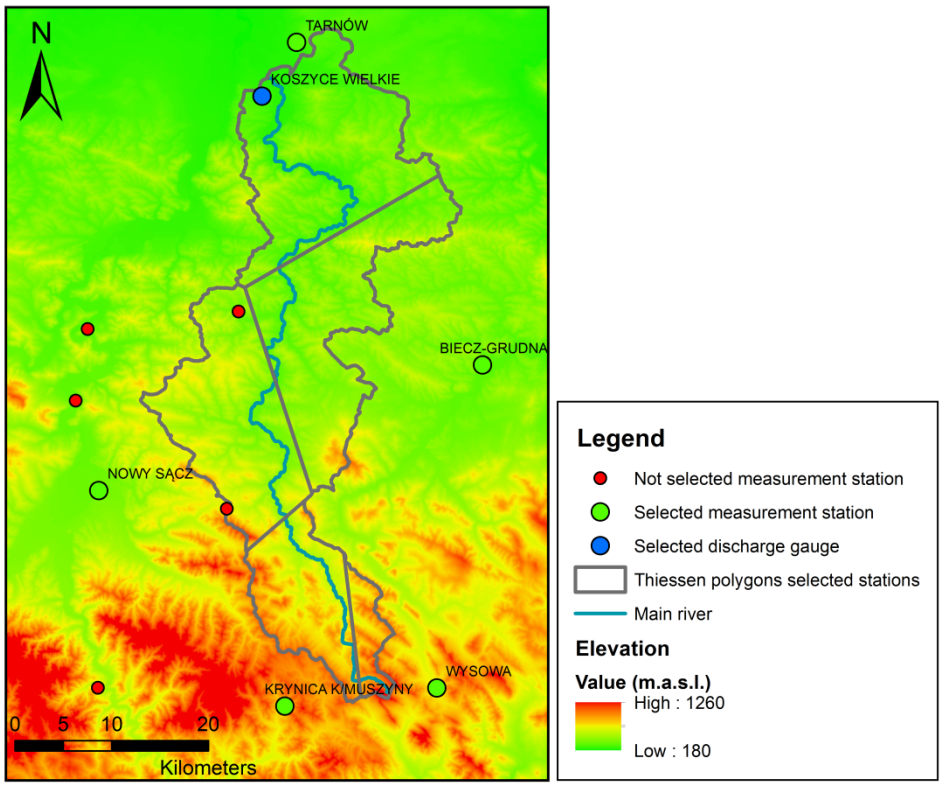


Figure 3: Locations of the measurement stations and Thiessen polygons of the selected meteorological stations

2.2.2 Meteorological observation data

The meteorological input data consists of precipitation and temperature. In this section the precipitation (section 2.2.2.1) and temperature (section 2.2.2.2) observation data are characterized.

2.2.2.1 Precipitation

Precipitation measurements are provided in millimetres per day with one decimal accuracy. Table 1 presents precipitation characteristics per measurement station. It can be seen that there are quite large differences in average precipitation per year at the different measurement stations. This indicates a relationship between precipitation and elevation, although other factors like on which side of the mountain a measurement station is located can also play a role. This is further elaborated in section 3.1.1. The uncorrected catchment-average precipitation per year over the catchment is 741.2 mm/year, and the runoff fraction (runoff divided by precipitation) is 41.3%. The highest daily precipitation amounts and most extreme precipitation events occur during summer. All maxima per measurement station that are presented in Table 1 occurred during the summer months June and July.

Table 1: Characteristics of the precipitation observation data over the period 1-11-1971 to 31-10-2013

Measurement station	Areal weight [%]	Elevation station [m.a.s.l.]	Average precipitation [mm/year]	Maximum in period [mm/day] (date)
Nowy Sącz	24.40	292	724.3	82.6 (30-6-1973)
Tarnów	30.67	209	725.1	90.7 (29-7-2000)
Biecz-Grudna	27.22	285	704.6	113.0 (3-6-2010)
Krynica K/Muszyny	13.86	585	843.7	94.5 (7-6-1999)
Wysowa	3.85	517	867.0	84.5 (3-6-2010)
Weighted areal average (uncorrected)			741.2	77.0 (3-6-2010)

2.2.2.2 Temperature

Temperature measurements are provided in Celsius degrees with one decimal accuracy, from the same measurement stations as precipitation observations. The temperature observation data per measurement station are summarized in Table 2. Also temperature will be corrected for elevation differences over the catchment (section 3.1.2). Temperature is used to calculate potential evapotranspiration and as input to the snow accumulation and snowmelt module in the hydrological model.

Table 2: Characteristics of the temperature observation data over the period 1-11-1971 to 31-10-2013

Measurement station	Areal weight [%]	Elevation station [m.a.s.l.]	Average temperature [^{o}C]
Nowy Sącz	24.40	292	8.5
Tarnów	30.67	209	8.8
Biecz-Grudna	27.22	285	7.8
Krynica K/Muszyny	13.86	585	6.1
Wysowa	3.85	517	6.1
Weighted areal average (uncorrected)			8.0

2.2.3 Discharge observation data

The average discharge over the period 1972-2013 is 9.4 m^3/s, with a standard deviation of 19.9 m^3/s. This means that there is a large variation in discharge, with very extreme peaks compared to the average discharge. The highest discharge that has been measured is 611 m^3/s. In Figure 4 the distributions of average discharge and extreme events over the year are presented. During spring the average discharge per day is largest, which is a result of snowmelt. Also during summer discharge

peaks occur frequently. These peaks are caused by high precipitation events, which mainly occur in summer. By laying the observation time series of precipitation and discharge side by side it appears that the lag time between precipitation peaks and discharge peaks is in general between 1 and 3 days. Hydrological years start on 1 November. The hydrological year that begins on 1-11-1971 will be called 1972.

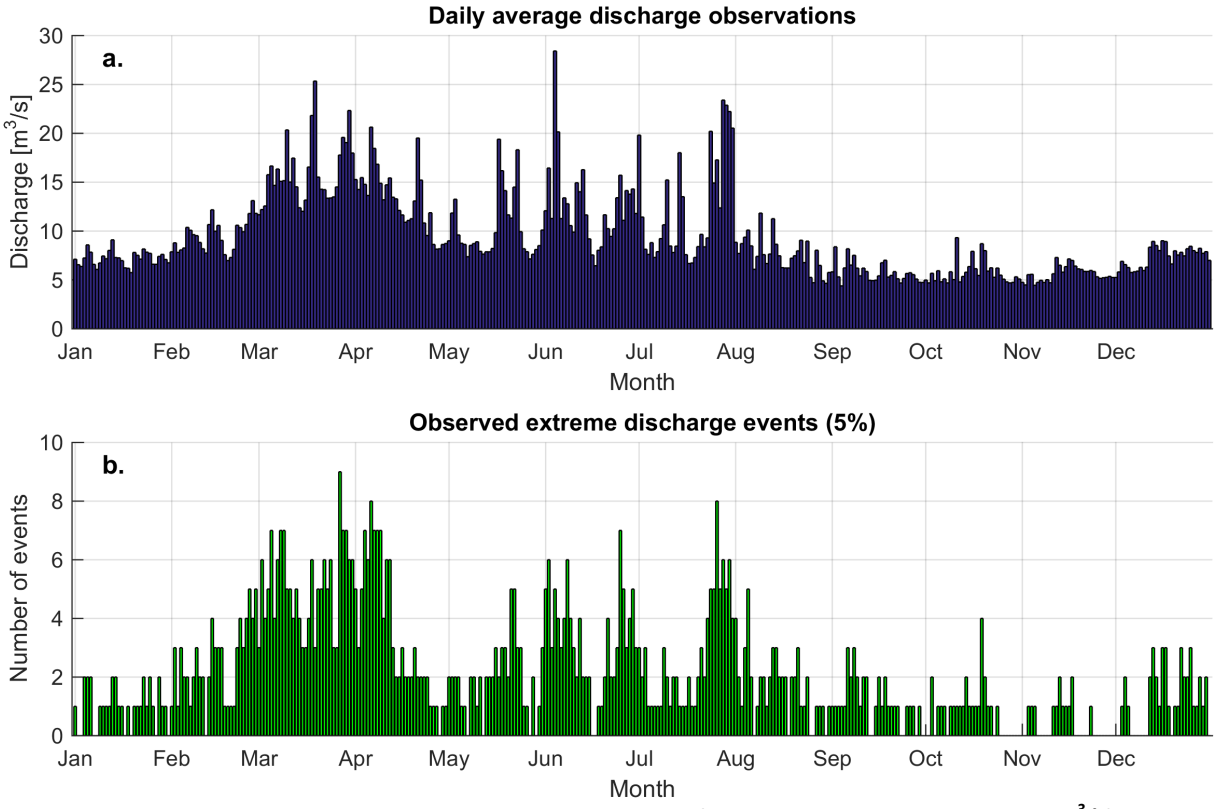


Figure 4: a. Average discharge per day b. Number of extreme events (exceedance probability of 5%, 28.3 m^3/s) per day, over 1972-2013

2.3 Meteorological forecast data

The THORPEX Interactive Grand Global Ensemble (TIGGE) project, developed by The Observing System Research and Predictability Experiment (THORPEX), provides historical ensemble forecast data from several forecast services (Bougeault et al., 2010). Buizza et al. (2005) have compared meteorological ensemble forecasts from three services, from ECMWF, MSC and NCEP. Their conclusion is that most verification measures indicate that the ECMWF ensemble forecast system provides the best overall performance, with the NCEP system being competitive during the first days and the MSC system during the last few days of the 10-day forecast period. Tao et al. (2014) conclude that among the evaluated numerical weather prediction models, ECMWF and the Japan Meteorological Agency have the best overall performance in both raw and processed forecasts. Of course the performance of the different services can be different for other locations. Because in previous literature it turned out that the performance of ECMWF forecasts is relatively good, these data have often been used in studies on hydrological ensemble forecasting (Cloke & Pappenberger, 2009) and IGF PAN already has experience with these data (Kiczko et al., 2015), the meteorological ensemble forecast data of ECMWF will be used in this project.

The forecasting system of ECMWF consists of several components, including an atmospheric General Circulation Model (GCM), an ocean wave model, a land surface model, an ocean GCM and

perturbation models (used for the ensemble forecasts) (Persson & Andersson, 2013). ECMWF provides global medium-range weather forecasts consisting of a high-resolution forecast (HRES) and an ensemble of lower-resolution forecasts (ENS) (ECMWF, 2012). The high-resolution forecast is the most accurate prediction of future weather ECMWF can provide, up to 10 days ahead (ECMWF, 2012). The meteorological forecasting system is run at a coarser resolution to generate 50 ensemble forecasts and a control forecast (Buizza et al., 2005). By providing ensemble results it is recognized that forecasts should be considered as stochastic forecasts (Tracton & Kalnay, 1993). Errors in a single, deterministic forecast arise as a result of initial condition errors and model errors, and the effect of these causes is inseparable (Leutbecher & Palmer, 2008). Persson and Andersson (2013) explain that to estimate the effect of possible initial state errors and model uncertainty and the related uncertainty of the forecasts, an ensemble of 50 different perturbed initial states (like surface pressure) is generated. In fact a set of 25 global perturbations is developed and reversed to obtain a mirrored set of 25 global perturbations (Persson & Andersson, 2013). Model uncertainty is incorporated by two stochastic perturbation techniques, which randomly perturb the tendencies in the physical parameterisation schemes and the vorticity tendencies (Persson & Andersson, 2013). It should be realized that the consequence of modifying the initial conditions around the most likely estimate of the truth is that the quality of the perturbed analysis is on average slightly lower than the quality of the control forecast (Palmer et al., 2006; Persson & Andersson, 2013). However Palmer et al. (2006) argue that trying to make each member of the ensemble forecast relatively more skilful compared to the control forecast by reducing the initial perturbations of ensemble members will not make a better ensemble prediction system, because this will not give a realistic indication of the unpredictability of the weather. The individual ensemble forecasts are on average less skilful than the control forecast, but by their large number they should form a good ensemble mean value and a reliable estimation of uncertainty in meteorological variables (Persson & Andersson, 2013). Persson and Andersson (2013) state: “The information should be used in its totality, i.e. from all the members in the ensemble. The low proportion of perturbed members “better” than the control forecast in the short range makes the task of trying to select the “member of the day” very difficult and, perhaps, impossible. There are no known methods to a priori identify the “best” ensemble member beyond the first day or so.” (p. 32) It can be concluded that the ensemble forecasts should always be used as a set.

Raw (unprocessed) historical forecast data of ECMWF are available via the TIGGE data portal from 1-10-2006 until recently. Because observation data are available until 31-10-2013, ECMWF data are downloaded for the period 1-11-2006 until 31-10-2013. The original resolution of the ensemble and control forecasts is 32*32 km (ECMWF, 2012), but via the TIGGE data portal this is interpolated to a regular grid (Bougeault et al., 2010) with a cell size of 0.25^{o} x 0.25^{o} (~17.9 km x 27.8 km). ECMWF uses a bi-linear interpolation technique (Persson & Andersson, 2013). Each grid cell that covers a part of the catchment is selected and they are weighted according to the relative area of the catchment that they cover (Figure 5 and Table 3). To run the hydrological model, the meteorological variables precipitation and temperature are required. The ECMWF ensemble data consists of 50 members and also the control forecast is used, following Renner et al. (2009), Demirel et al. (2013a), Olsson and Lindström (2008) and the EFAS system (Thielen et al., 2009a). It is considered that no perturbation is also a possible state of the atmosphere. ECMWF forecasts are available with a time interval of 12 hours or 24 hours, until a maximum lead time of 360 hours. Observation data are available at a daily time interval and the hydrological model works on a daily basis, so also for the forecasts a time

interval of 1 day is used. A maximum lead time of 10 days is used, following the World Meteorological Organization (WMO) that defines medium-range forecasts as lead times from 3 days to 10 days (ECMWF, 2012) and many other studies on ensemble flow forecasting that also used 9 days or 10 days as maximum lead time (e.g. (Olsson & Lindström, 2008; Renner et al., 2009; Roulin & Vannitsem, 2005; Verkade et al., 2013)). Precipitation forecasts are provided as values accumulated over a time interval (Bougeault et al., 2010; Persson & Andersson, 2013). Precipitation forecasts per day are extracted by calculating the increase in accumulated precipitation between two days. As with observation data, it is considered that a forecast is representative for a calendar day. In Figure 6 an example precipitation forecast is presented. Temperature forecasts are not provided as mean temperatures over a period, but as values at a certain time (forecast at 00:00 applies to 00:00). This means that temperature could not directly be considered as representative for the whole day. To obtain a representative temperature it needs to be averaged over the day. Therefore temperature forecasts are downloaded with origin times 00:00 and 12:00, both with an interval time of 24 hours. To calculate an average temperature for the whole day, the temperature at 00:00 is weighted with 25%, the temperature at 12:00 with 50% and the temperature at 24:00 (00:00 of the next day) with 25%. It should be realized that this is a very simple method, which of course will introduce errors. But in general this should provide suitable daily average temperatures and it is a better approximation of daily average temperature than the temperature at 00:00 or 12:00.

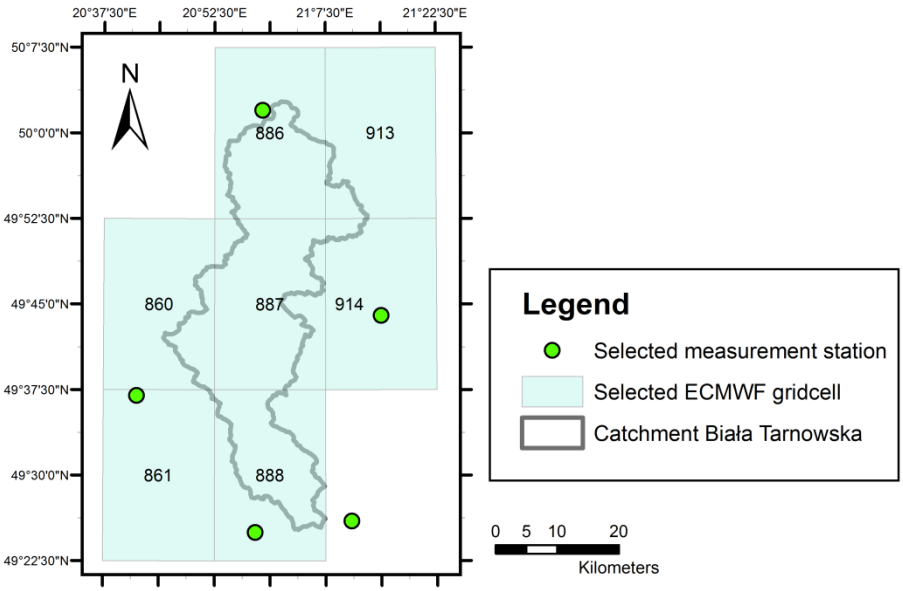


Figure 5: Locations of the measurement stations and ECMWF grids that cover the Biała Tarnowska catchment

Table 3: Grid cell weights

Gridcell	Weight [%]
860	8.3
861	0.1
886	24.3
887	39.9
888	21.9
913	3.6
914	1.9

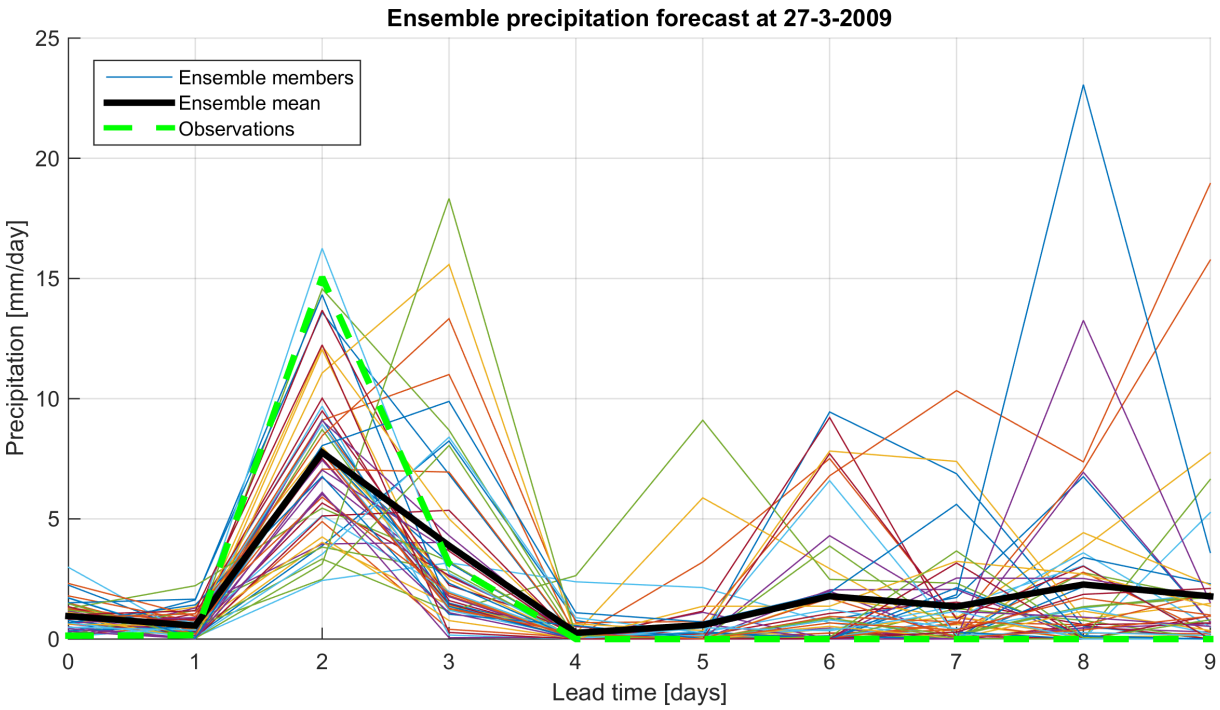


Figure 6: Example of an ensemble precipitation forecast from ECMWF

3. Methodology

This chapter describes the steps that are taken to fulfil the objective of this study. Figure 1 presents an overview of the methods that are applied. In section 3.1 the preparation of observation data is described. In section 3.2 the hydrological model is explained and in section 3.3 the calibration method is described. In section 3.4 an update state procedure is developed and in section 3.5 different pre- and post-processing strategies are implemented. In section 3.6 the ensemble flow forecasting system is described. These methods lead to an answer to research question 1. In section 3.7 to section 3.9 the evaluation procedure is described to answer research questions 2 and 3.

3.1 Data preparation observation data

In this section the preparation of the precipitation and temperature observation data is described. Due to elevation differences over the catchment, it is needed to correct the precipitation and temperature measurements.

3.1.1 Elevation correction of precipitation

There are quite large differences in elevation over the catchment (also see Figure 3). In general precipitation increases with elevation, particularly on windward slopes (Sevruk, 1997). In basins with large elevation differences the catchment-average precipitation may be underestimated if only measurement stations at low elevations are available (Martinec et al., 2008; Zhang et al., 2014), and in general most measurement stations are located in the valleys (Panagoulia, 1995; Sevruk, 1997). The precipitation gradient (in mm/100 m or %/100 m), which is the increase in precipitation proportionate to the increase in elevation, is often used for precipitation extrapolation (Panagoulia, 1995; Zhang et al., 2014). For the Biała Tarnowska catchment this linear relationship is presented in Figure 7, based on the elevation (from metadata) and annual average precipitation of the 5 measurement stations during the period 1971-2009 (full period of data that was available at this moment in the study). An increase in annual precipitation of 44.84 mm/year per 100 m is found, which is an increase of 6.1% per 100 m (relative to uncorrected catchment-average precipitation).

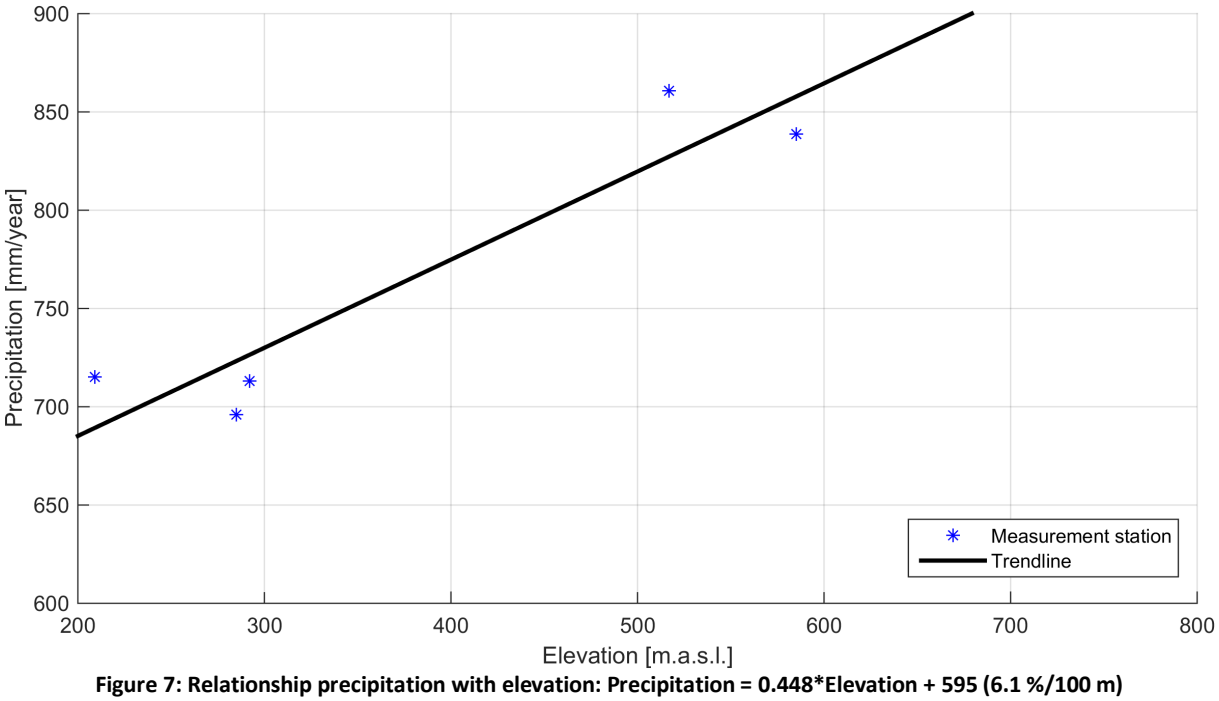


Figure 7: Relationship precipitation with elevation: Precipitation = 0.448*Elevation + 595 (6.1 %/100 m)

Zhang et al. (2014) and several studies that are mentioned by Zhang et al. (2014) show variability of the precipitation gradient over the year. Figure 8 is based on relationships as in Figure 7, but for each month separately. In absolute sense the precipitation gradient is larger during the summer months, because precipitation is larger in these months and the absolute differences between the stations are larger. The relative precipitation gradient is larger during the winter months, especially from December until February, so in these months precipitation increases relatively more with elevation than during the other months. A possible explanation for this is that during the months December until February the air is colder at higher elevations and there is snowfall at these locations, while at the lower elevations there is no precipitation. Possibly this is mixed in November, which shows a slightly higher precipitation gradient compared to the other months between March and November. From Figure 8 two periods are distinguished, namely December – February with a high relative precipitation gradient and March – November with a low relative precipitation gradient. This is not done per month, because within these periods the differences are not large and using separate months would decrease the amount of data on which the relationships are based. The precipitation gradients for these periods (10.5 %/100 m and 5.4 %/100 m) are used to calculate the correction factors. It is assumed that the linear precipitation gradients can be extrapolated to higher elevations than the elevations of the measurement stations on which they have been based.

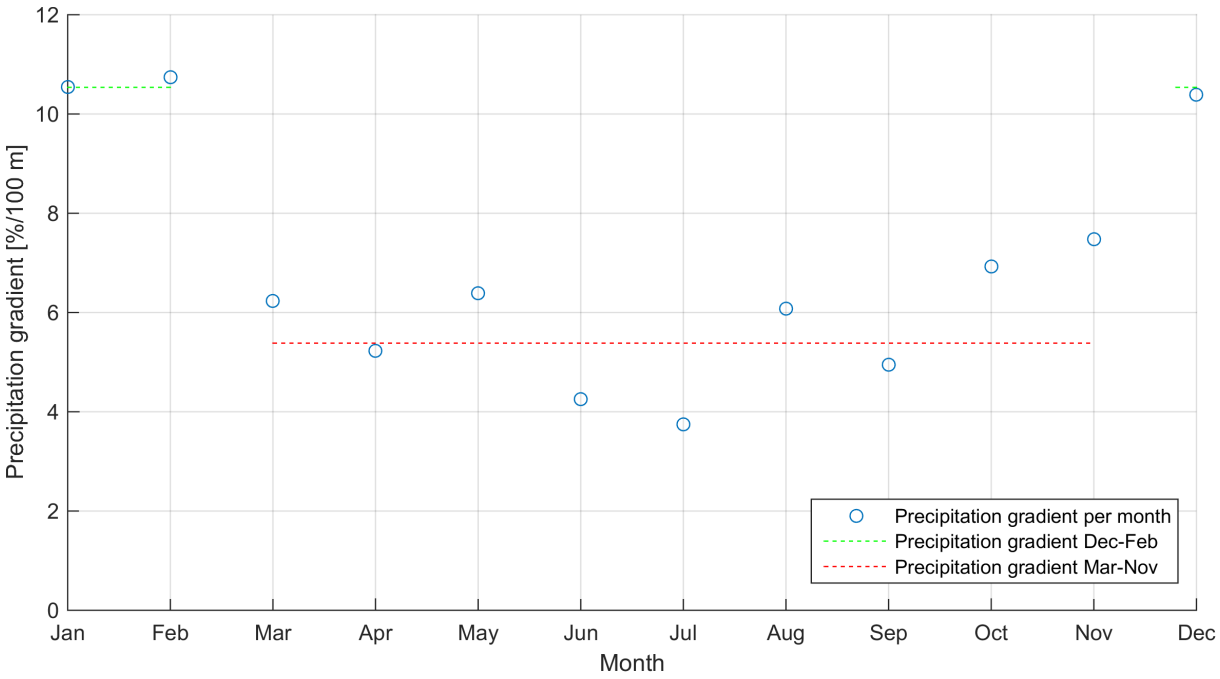


Figure 8: Precipitation gradient variation over the year in mm/month/100 m relative to uncorrected catchment-average precipitation

It is difficult to compare these values with other studies, because many studies do not provide the precipitation gradient as a percentage but in mm/100 m (like Sevruk (1997) and Zhang et al. (2014)). In the study of Sevruk (1997) the precipitation gradient varies strongly over Switzerland, and for this study it has been estimated that it varies roughly between 0%/100 m and 10%/100 m. For the study of Zhang et al. (2014) precipitation gradients of about 6%/100 m has been estimated, which in relative sense do not vary much between months. Panagoulia (1995) found a precipitation gradient of 12.6%/100 m for a catchment in Greece, based on only 4 measurement stations. The default value that is applied in Hydrologiska Byråns Vattenbalansavdelning (HBV) models is 10%/100 m. In general it can be said that the precipitation gradients that are found in this study are in the same order of

magnitude as the values in literature. Although the number of measurement stations is only 5, which is few, it has been chosen to use the calculated values instead of a default value from literature. There is a clear difference between two periods of the year, so one default value over the whole year is not appropriate. In addition the points are close to the fitted lines, so this provides confidence in the obtained precipitation gradients.

Column 3 and 4 in Table 4 and Table 5 present the elevation of the measurement stations (from metadata) and the mean elevation of the Thiessen polygon they represent. The elevation of most Thiessen polygons is under-represented by the elevation of the measurement station, which means that also precipitation over this part of the catchment is underestimated. Panagoulia (1995) proposes the following equation to calculate the correction factor:

$$\lambda_{P,i} = 1 + \frac{e_{c,i} - e_i}{P_i}\gamma_P \qquad [3.1]$$

$\lambda_{P,i}$ = Relative correction factor of precipitation for elevation per measurement station *i* [-]
$e_{c,i}$ = Thiessen polygon catchment elevation per measurement station *i* [m.a.s.l.]
e_i = Elevation of measurement station *i* [m.a.s.l.]
P_i = Uncorrected precipitation at measurement station *i* [mm]
γ_P = Rate of precipitation variation with elevation [mm/m]

Panagoulia (1995) applied this formula to the whole catchment after weighting of the measurement stations. It is more suitable to apply this to the individual measurement station-Thiessen polygon combinations, since these are the basic units that need to be corrected. For $e_{c,i}$ in equation 3.1 the elevation distribution in a Thiessen polygon (from the elevation distribution histograms) is used, instead of the mean elevation in a Thiessen polygon. This gives the correction factors per measurement station in column 6 in Table 4 and Table 5. Subsequently, to correct the precipitation observations from the measurement stations and average them to catchment-average precipitation, the original Thiessen polygon weighting factors are multiplied by the correction factors (column 7 in Table 4 and Table 5). On average, the elevation correction results in an increase of precipitation over the entire catchment of 7.0% in the period December – February and an increase of 3.4% in the period March – November, but on a daily basis this can be much different depending on distribution of precipitation over the measurement stations.

By using relative correction factors, instead of absolute differences, large precipitation amounts are in absolute sense increased more than low precipitation amounts. Using a fixed absolute correction would mean that even on days with zero or very low precipitation there would be precipitation after correction. As a result of precipitation correction the average annual precipitation increases from 741.2 mm to 768.4 mm.

Topographic effects are not included in this simple correction method. According to Sevruk (1997) larger precipitation gradients are found for measurement stations in valleys compared to precipitation gradients that are based on measurement stations on the slopes. Since in general most measurement stations are located in valleys (Panagoulia, 1995; Sevruk, 1997), this will result in an overestimation of the precipitation gradient. The effect of wind is also not included. This might be present in the observations, but the number of measurement stations is too few to establish different gradients for the windward and leeward slope of mountains.

Table 4: Calculation of precipitation correction factors per measurement station for the period December-February

Measurement station	Weighting [%]	Elevation station [m.a.s.l.]	Mean elevation Thiessen polygon [m.a.s.l.]	Average precipitation Dec-Feb [mm/day]	Weighted average correction factor Dec-Feb [-]	New weighting Dec-Feb [%]
Nowy Sącz	24.40	292	403.5	1.075	1.127	27.50
Tarnów	30.67	209	274.2	1.146	1.070	32.81
Biecz-Grudna	27.22	285	339.2	1.070	1.063	28.93
Krynica K/Muszyny	13.86	585	553.5	1.473	0.973	13.49
Wysowa	3.85	517	642.1	1.506	1.103	4.25
Sum	100					106.98

Table 5: Calculation of precipitation correction factors per measurement station for the period March-November

Measurement station	Weighting [%]	Elevation station [m.a.s.l.]	Mean elevation Thiessen polygon [m.a.s.l.]	Average precipitation Dec-Feb [mm/day]	Weighted average correction factor Dec-Feb [-]	New weighting Dec-Feb [%]
Nowy Sącz	24.40	292	403.5	2.241	1.061	25.89
Tarnów	30.67	209	274.2	2.225	1.036	31.77
Biecz-Grudna	27.22	285	339.2	2.180	1.031	28.06
Krynica K/Muszyny	13.86	585	553.5	2.567	0.985	13.64
Wysowa	3.85	517	642.1	2.636	1.059	4.08
Sum	100					103.44

3.1.2 Elevation correction of temperature

Temperature also varies with elevation. The temperature lapse rate is defined as the decrease of temperature with elevation in ^{o}C/100 m. According to Martinec et al. (2008) the temperature lapse rate can be determined from historical temperature observations from different measurement stations or it must be adopted from other basins or with regard to climatic conditions. The temperature lapse rate can vary between the dry adiabatic lapse rate (0.98 ^{o}C/100 m) and the isothermal lapse rate (0 ^{o}C/100 m) (Dobrowski et al., 2009; Zhang et al., 2014). Dobrowski et al. (2009) explain that the actual temperature lapse rate will be close to the dry adiabatic lapse rate during summer and close to the isothermal lapse rate for low temperatures. The global standard atmospheric lapse rate is 0.65 ^{o}C/100 m (Dobrowski et al., 2009; Zhang et al., 2014). Li et al. (2013) mention that commonly a fixed temperature lapse rate in the range of 0.60 ^{o}C/100 m - 0.65 ^{o}C/100 m is used for hydrological models. The default value that is applied in HBV models is 0.60 ^{o}C/100 m.

However, the temperature lapse rate can differ substantially in space and time and several studies, like Li et al. (2013) and Zhang et al. (2014), did not use a fixed temperature lapse rate but calculated how the lapse rate differs in time and in space. Based on 533 meteorological stations over the mainland of China, Li et al. (2013) found substantial spatial and seasonal differences in the temperature lapse rate. Also Zhang et al. (2014) found for a high mountainous area in the Himalayan Mountains, based on 4 meteorological stations, that the temperature lapse rate varies over the year, with larger gradients in spring, summer and fall and smaller or even temperature inversion during winter. Li et al. (2013) found that temperature lapse rates were weaker than 0.65 ^{o}C/100 m over most regions, but the temperature gradients are steeper in some regions. According to Li et al. (2013) this shows that the usually assumed fixed value is not appropriate for China. There have been many more studies on temperature lapse rates or where temperature lapse rates are applied, but many of them focus on high mountainous areas and/or glaciers. This is because elevation differences are larger in these areas and thus the temperature gradient is more important. In this study there are not enough measurement points available to establish temperature lapse rates that vary spatially over the catchment and in addition the catchment is relatively small so it is expected that there are

no large differences in temperature lapse rate over the catchment. In Figure 9 it can be seen that the temperature lapse rate does not vary much over the year, but it is higher than the global standard atmospheric lapse rate and default HBV value. One fixed temperature lapse rate is appropriate in this case. The temperature lapse rate based on complete years would be 0.77 °C/100 m. Li et al. (2013) found that when the number of meteorological stations is less than 16, the resulting temperature lapse rate shows a great variation, which indicates a larger uncertainty. Therefore it has been chosen to apply the global standard temperature lapse rate of 0.65 °C/100 m. In chapter 5 the used precipitation gradients and temperature lapse rate are further discussed.

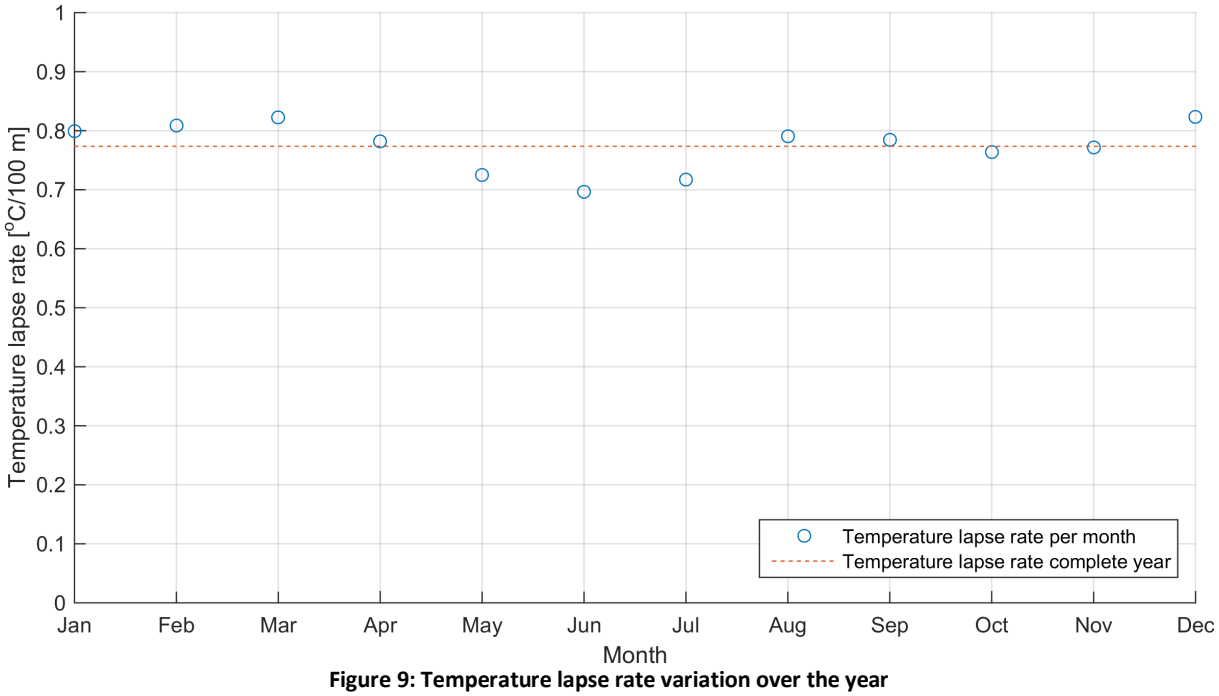


Figure 9: Temperature lapse rate variation over the year

Martinec et al. (2008) propose the following equation to calculate the elevation adjustment of temperature:

$$\Delta T_i = \gamma_T (e_i - e_{c,i}) \left(\frac{1}{100}\right) \qquad [3.2]$$

ΔT_i = Absolute temperature correction for elevation per measurement station *i* [°C]
γ_T = Temperature lapse rate = $0.65\ °C/100\ m$
e_i = Elevation of measurement station *i* [m.a.s.l.]
$e_{c,i}$ = Thiessen polygon catchment elevation per measurement station *i* [m.a.s.l.]

With equation 3.2 a temperature correction per measurement station is calculated, constant over the year. As with precipitation correction the elevation distributions in the Thiessen polygons are used to calculate an average correction factor per measurement station. In this case an absolute correction is used, because temperature is not bounded by zero like precipitation. So precipitation is increased with a relative amount and temperature with an absolute amount. This is the same approach as Akhtar et al. (2009) used. The resulting temperature adjustment values are presented in Table 6. After temperature correction the annual average potential evapotranspiration decreases from 695.3 mm to 674.4 mm.

Table 6: Correction factors of temperature per measurement station

Measurement station	Elevation station [m.a.s.l.]	Mean elevation Thiessen polygon [m.a.s.l.]	Average temperature [°C]	Temperature adjustment [°C]
Nowy Sącz	292	403.5	8.5	-0.7
Tarnów	209	274.2	8.8	-0.4
Biecz-Grudna	285	339.2	7.9	-0.4
Krynica K/Muszyny	585	553.5	6.1	0.2
Wysowa	517	642.1	6.0	-0.8

3.2 Hydrological model description

Hydrological models are an essential part of ensemble flow forecasting systems. In a hydrological model the input data, in the form of precipitation, temperature and potential evapotranspiration over a catchment are translated into runoff from the catchment by simulating the relevant hydrological processes. Precipitation and temperature are provided by observation data or forecasts and potential evapotranspiration is based on temperature data. In section 3.2.1 the hydrological model is further explained and in section 3.2.2 the choice for a method to calculate potential evapotranspiration is explained.

3.2.1 Hydrological model

A variety of hydrological models exists, inter alia presented by Singh (1995). In this study the Hydrologiska Byråns Vattenbalansavdelning (HBV) model will be used. The HBV model was already developed in the early 1970s, but since then it has been further developed (Lindström et al., 1997). The model has been applied in more than 90 countries and for several applications, including hydrological forecasting (Bergström & Lindström, 2015). According to Bergström and Lindström (2015) the model is suitable to generate ensemble forecasts, because of the limited demand for computer power. It has also been applied in several previous studies on ensemble flow/flood forecasting (Cloke & Pappenberger, 2009; Demirel et al., 2013a; Olsson & Lindström, 2008; Renner et al., 2009; Verkade et al., 2013), it has been used earlier at IGF PAN (Kiczko et al., 2015; Osuch et al., 2015; Romanowicz et al., 2013) and at the University of Twente (Booij, 2005; Demirel et al., 2013a; Knoben, 2013; Maat, 2015; Van den Tillaart, 2010). In addition the HBV model contains a snow accumulation routine, which is an important process in the study catchment.

The HBV model is a conceptual rainfall-runoff model (Knoben, 2013; Werner et al., 2006). A conceptual hydrological model includes the relevant hydrological processes to determine the relationship between input (meteorological variables) and output (runoff) (Knoben, 2013). Not all parameters have a direct physical meaning, so they have to be calibrated against observation data (Pechlivanidis et al., 2011). The HBV model is based on several storage boxes, fluxes between these storage boxes and fluxes in (precipitation) and out (evapotranspiration and runoff) of the hydrological system. Nowadays the HBV-96 version, developed by Lindström et al. (1997), is often used. The HBV model that is currently used by IGF PAN, and that will be used in this project, contains 14 parameters. In Figure 10 the HBV model structure is presented. In appendix 1 the HBV model structure and equations are further described.

In addition to the discharge observations at the downstream end of the river (see section 2.2.1), also discharge observations are available halfway the river. Nevertheless the HBV model will be applied as a lumped model over the Biała Tarnowska catchment, because the study catchment is relatively small and it is not expected that separating the catchment in different sub-catchments will result in

an improvement of the model performance. In a lumped model the whole catchment is treated as one hydrologic entity within which different hydrological processes are taking place.

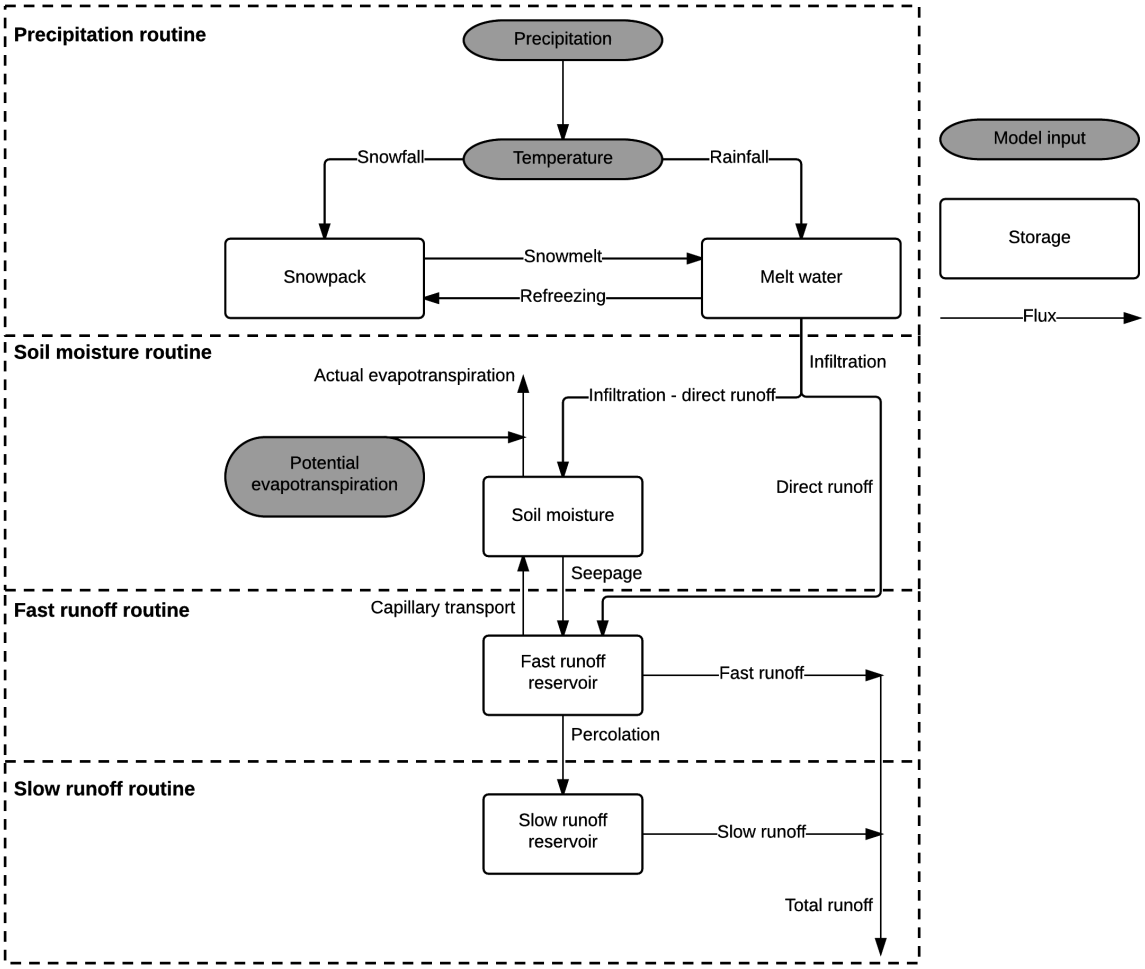


Figure 10: Structure of the applied HBV model (Knoben, 2013), further explained in appendix 1

3.2.2 Estimation method for potential evapotranspiration

Potential evapotranspiration is one of the required inputs to the hydrological model. Potential evapotranspiration is defined as the amount of water that could evaporate and transpire from a vegetated landscape without restrictions other than atmospheric demand (Lu et al., 2005). Different methods to estimate the potential evapotranspiration will give significantly different results (Rao et al., 2011) and in this way this has influence on how the hydrological model performs. Previously, IGF PAN used the method of Hamon to calculate the potential evapotranspiration because it is a simple method and it has provided good results in several impact studies (Osuch, personal communication, 2015). Oudin et al. (2005) recommend to use a temperature-based potential evapotranspiration model in a daily rainfall-runoff model, among which the method of Hamon is mentioned. Lu et al. (2005) only compares actual evapotranspiration and potential evapotranspiration rates from different methods on annual basis, for which the actual evapotranspiration is estimated by the water balance of the catchments. According to Lu et al. (2005) care must be paid to selecting the appropriate method for a particular catchment, because different methods produce inconsistent results for some catchments and years. Lu et al. (2005) recommend to use the Priestley-Taylor method if radiation data are available, and otherwise the method of Hamon could be used. ECMWF provides radiation forecasts but there are no radiation measurements available, so the method of Hamon is the better choice. The method of Hamon will be used, because in general this method was

able to provide reasonable results in previous studies and it is simple to use with the available meteorological data (only temperature is needed as input). In appendix 2 it is further elaborated how the method of Hamon has been applied.

3.3 Calibration

The calibration procedure consists of several elements. At first in section 3.3.1 an objective function is chosen and in section 3.3.2 the calibration and validation period are defined. The calibration procedure will be started with a sensitivity analysis (section 3.3.3) to identify the most sensitive parameters that are important in the calibration. Next to that the calibration algorithm to find the optimum parameter set is explained (section 3.3.4). Two calibration rounds will be applied. In the first calibration round all 14 parameters in the HBV model are calibrated and in the second calibration round only the most sensitive parameters are calibrated. After all an uncertainty analysis will be executed to assess the uncertainty of the calibrated parameters (section 3.3.5).

3.3.1 Objective function

An objective function quantifies the quality of the model with a certain parameter set. Different objective functions measure different properties of the quality of the simulations. They can for example be conditioned on peak flows or low discharges. Since this study focuses on flows in general, the model should perform well during low, medium and high flows. An objective function that is often used to calibrate hydrological models is the Nash-Sutcliffe coefficient (*NS*) (Nash & Sutcliffe, 1970; Romanowicz et al., 2013). With *NS* the shape of the hydrograph is evaluated (Van den Tillaart, 2010). However, *NS* is more sensitive to peak flows, which can be at the expense of model performance during low flows. The relative volume error (*RVE*) quantifies volume errors, so that consistent over- or underprediction of discharge is prevented. Akhtar et al. (2009) incorporated *NS* and *RVE* into one objective function, called *Y*. *Y* is calculated as follows:

$$Y = \frac{NS}{1 + |RVE|} \qquad [3.3]$$

$$NS = 1 - \frac{\sum_{t=1}^{t=N}[Q_s(t) - Q_0(t)]^2}{\sum_{t=1}^{t=N}[Q_0(t) - \overline{Q_0}]^2} \qquad [3.4]$$

$$RVE = \frac{\sum_{t=1}^{t=N}[Q_s(t) - Q_0(t)]}{\sum_{t=1}^{t=N} Q_0(t)} \qquad [3.5]$$

Q_s = Simulated discharge [m^3/s]
Q_0 = Observed discharge [m^3/s]
$\overline{Q_o}$ = Average observed discharge [m^3/s]
t = Time step [day]
N = Total number of time steps [days]

NS can range between -∞ and 1. Higher values indicate a better agreement between observed and modelled discharge, with 1 being a perfect fit (Q_s equal to Q_o) (Knoben, 2013). For a *NS* value of 0 the model does not perform any better than the mean value of observed discharges (WMO, 2015). The *RVE* can be positive and negative. Positive values indicate that over the whole time period discharge is oversimulated by the model and negative values indicate that discharge is undersimulated. The optimum value of the *RVE* is 0%, which means that there is no difference between observed and

simulated volumes (Knoben, 2013). Thus in the optimum situation (perfect simulation) *NS* has a value of 1 and *RVE* has a value of 0%, which gives an optimum value of 1 for *Y*.

3.3.2 Calibration and validation period

To calibrate and validate the hydrological model a split-sample test will be applied, which means that the calibration is based on one time period and the validation is performed on another period of data of the same catchment (Klemes, 1986). This has been chosen because the calibrated model will only be applied to one study catchment. It is assumed that catchment characteristics stay the same over the calibration period and the period of application. Observation data are available from 1971, while the model will be applied to ensemble forecast data over the period 2006-2013. 1971-2000 is chosen as calibration period and 2001-2013 as validation period. To ensure that the calibration is done on full hydrological years the model is calibrated over the period 1-11-1971 to 31-10-2000 and validated over the period 1-11-2000 to 31-10-2013. This provides 10 months of initialization before the calibration period to set initial conditions (because observational data are available from 1-1-1971) and before the validation period an initialization period of 1 year will be used. With two simple executions of the calibrated HBV model it has been proven that after 250 days the difference in discharge output between the model running from 1-11-1971 and the model running from 1-11-2004 is consistently below 0.001 m^3/s, so it is considered that an initialization period of 10 months is sufficient.

3.3.3 Sensitivity analysis

The HBV model contains 14 parameters, which are many parameters to calibrate a model. Therefore, in a second calibration round and to establish a more efficient parameter uncertainty analysis only the parameters that have a considerable influence on the model output are varied. A sensitivity analysis will be used to identify the model parameters with the highest influence on the model results (Hamby, 1994; Song et al., 2015).

A number of sensitivity analysis techniques exists (Hamby, 1994; Song et al., 2015). One form of sensitivity analysis is screening. The purpose of a screening method is to identify which parameters contribute significantly to the output uncertainty, rather than to quantify sensitivity exactly (Song et al., 2015). This fits the purpose of the sensitivity analysis here, because it is not the purpose to quantify sensitivity exactly but only to classify parameters in sensitive and insensitive parameters. When the objective is to obtain insights into the characteristics of sensitivity indices, quantitative methods are more appropriate (Song et al., 2015). One of the most commonly used global screening methods in hydrological modelling is the method of Morris (1991) (Song et al., 2015), for example also used by Osuch et al. (2015). Campolongo et al. (2007) state that the method of Morris has proven to be a very good compromise between accuracy and efficiency, especially for the analysis of large models. The method of Morris is especially suitable for models where the number of uncertain factors is high and/or the model is expensive to compute (Campolongo et al., 2007). It has lower computational costs compared to other global sensitivity analysis methods and it is easy to implement and interpret (Song et al., 2015). Song et al. (2015) mention several studies on hydrological modelling in which the method of Morris has been used. For distributed hydrological models Herman et al. (2013) conclude that the method of Morris provides results sufficiently similar to the Sobol' method, at a greatly reduced computational expense. The results suggest that the method of Morris is not able to reliably reproduce the precise ranking of high-sensitive parameters provided by the Sobol' method, but the method successfully distinguishes sensitive from insensitive

parameters (Herman et al., 2013). Also Shin et al. (2013) found that the results of the method of Morris are similar to the results of the Sobol' method, although there are slight differences in the ordering of parameter sensitivities. Since the purpose of this sensitivity analysis is to identify the sensitive parameters the method of Morris is a suitable method to use. The choice for the method of Morris has been made in view of efficiency and easiness to understand and apply the method. In the next section the method of Morris is further explained.

3.3.3.1 Method of Morris

The method of Morris is based on repeating individually randomized one-factor-at-a-time experiments (Campolongo et al., 2007; Morris, 1991). Campolongo et al. (2011) explain that to each factor a so-called elementary effect is assigned, which is defined as the ratio between the variation in the model output (represented by an objective function) and the variation in the input factor itself. *r* different elementary effects are estimated for *r* different trajectories of parameter values (Campolongo et al., 2011). One trajectory contains a sequence of *M* perturbations, one for each parameter (Herman et al., 2013). It is possible to calculate the elementary effects of each parameter with only $M + 1$ model evaluations. However, this would make the result highly dependent on the location of the initial parameter set and it does not account for interactions between parameters (Herman et al., 2013). Therefore this is repeated over multiple trajectories (Campolongo et al., 2011; Herman et al., 2013). Basic statistics of the resulting elementary effects are calculated, which reflect the sensitivity of the model on each parameter (Campolongo et al., 2007; Morris, 1991). The sensitivity measure μ (mean) and σ (standard deviation) are calculated from the elementary effects for each parameter (Campolongo et al., 2007). μ indicates the overall influence of the parameter on the output and σ estimates the higher order effects, which are non-linear and/or interaction effects (Campolongo et al., 2007; Song et al., 2015). The problem of μ as sensitivity measure is that if the distribution of Elementary Effects contains positive and negative elements, some effects may cancel out others (Campolongo et al., 2007). Campolongo et al. (2007) introduced an additional sensitivity measure μ^*. μ^* is the mean of absolute values of the Elementary Effects. This measure solves the problem of opposite signs in the Elementary Effects, but it does not provide information about the sign of the effect. The sensitivity measures allow to classify the inputs in three groups: inputs having negligible effects, inputs having linear and additive effects and inputs having non-linear and/or interaction effects (Morris, 1991).

Regarding the sampling strategy of parameter values, Campolongo et al. (2007) propose an improvement compared to the original method of Morris (1991) which should provide a better scanning of the input domain without increasing the number of model executions. The idea is to select the *r* trajectories that maximise dispersion in the input space (Campolongo et al., 2007). The procedure starts with a high number of Morris trajectories (*C*), and then *r* trajectories with the highest spread are selected. It is assumed that parameters are uniformly distributed between their predefined lower and upper bounds (Campolongo et al., 2007). $r * (M + 1)$ parameter sets are generated and for each of these parameter sets the hydrological model is run and the objective function is calculated, to see what happens if one of the parameter values changes.

There are also limitations of the method of Morris. At first, this screening method cannot quantify the effects of different parameters on outputs (Song et al., 2015). Since the objective of this sensitivity analysis is not to quantify the effect but just to distinguish sensitive and insensitive parameters this screening method is sufficient here. Another drawback is that type II errors might

occur with this method, which means that an important parameter might be classified as non-influential (Song et al., 2015). Both the Sobol' method and the method of Morris are vulnerable to produce these errors (Shin et al., 2013). This should be kept in mind when parameters are classified as insensitive. Song et al. (2015) conclude that the best choice for a sensitivity analysis is to use a combination of more than one sensitivity analysis method.

3.3.3.2 Application method of Morris

A Matlab script to execute the method of Morris with the extensions of Campolongo et al. (2007) has been provided by IGF PAN. This has been applied to the calibration period, using *Y* as objective function. It is expected that the parameter values from the first calibration round are already in the good direction of the true global optimum. This is confirmed by running a second calibration with 14 parameters, which resulted in a different parameter set but the parameter values were close to each other. The problem with using the full parameter range as defined in appendix 1 is that the sensitivity of parameters is tested in situations far away from the optimum. Therefore it has been chosen to estimate the sensitivity in a range of 20% of the original parameter range around the calibrated parameter value and bounded by original parameter range boundaries. The other settings that are used are given in Table 7. In total this yields 300 runs of the hydrological model ($r * (M + 1) = 300$).

Table 7: Settings of the sensitivity analysis with the method of Morris

Setting	Description	Value
M	Number of model parameters	14
C	Number of different Morris trajectories. Typically between about 500 and 1000 (Campolongo et al., 2007)	1000
r	Final number of trajectories. Typically 4 – 8 (Campolongo et al., 2011), but this is increased to 20 trajectories. The number of varying parameters is relatively small (reduces number of model runs) and it turned out that the results can change per application of the method of Morris with 8 trajectories. With 20 trajectories the results are more reliable.	20
p	Number of levels. Each variable can vary across *p* levels over the parameter range.	10

3.3.4 Deterministic calibration

The objective of the calibration procedure in this study is to obtain an optimum deterministic parameter set. As calibration algorithm Differential Evolution with Global and Local neighbourhoods (DEGL) will be applied, because it is efficient in finding the global optimum deterministic parameter set (Osuch, personal communication, 2015). The hydrological model will be calibrated with precipitation and temperature observation data as input data and evaluated against the discharge observations. The optimum deterministic parameter set will be applied to generate ensemble flow forecasts with meteorological forecast data as input. One parameter set is used for all lead times, because the catchment characteristics do not change with lead time.

Differential Evolution (DE) and DEGL have been described extensively by Das et al. (2009). DE is a simple and efficient scheme for global optimization over continuous spaces. However, DE faces problems of stagnation (stop proceeding to the global optimum) and premature convergence (converges to a local optimum, losing its diversity). Das et al. (2009) mention that the performance of DE deteriorates with the growth of the dimensionality of the search space. With DE an initial population of parameter vectors is generated, assuming uniform distributions between predefined lower and upper parameter boundaries. By mutation and crossover algorithms a next generation is created and to keep the population size constant it is determined whether the target or trial vector

survives to the next generation, based on the objective function. Most population-based search algorithms try to balance between two contradictory aspects of their performance, namely exploration and exploitation. The ability to explore means that the algorithm is able to search every region of the feasible search space, while exploitation means the ability to efficiently converge to near-optimal solutions. These abilities depend on the mutation scheme that is applied. In the case of the DE variant 'DE/target-to-best/1' the best vector of the population is used to generate donor vectors. This promotes exploitation, since all vectors are attracted towards the same best position. As a result of this, the population loses its global exploration ability within a relatively small number of generations. So it will converge to some locally optimal point in the search space. The DEGL algorithm is based on the DE principle. With DEGL Das et al. (2009) has improved the trade-off between exploration and exploitation by using two kinds of neighbourhood models, a local neighbourhood model and a global mutation model. Within the local neighbourhood model each vector is mutated using the best position found in a small neighbourhood of it and in the global neighbourhood model the globally best vector of the entire population is used for mutating a population member. In the end the local and global model are combined using a weight factor. Das et al. (2009) provide different schemes for selection of the weight factor, one of them is a self-adaptive weight factor. In the study of Das et al. (2009) at first a low weight factor (favouring exploration) and then an increasing of the weight factor is observed (favouring exploitation).

Das et al. (2009) show that compared to other DE variants and other evolutionary optimization techniques DEGL is clearly and consistently superior for most benchmark functions. DEGL with the self-adaptive weight factor gave the best performance (Das et al., 2009). This variant has been used earlier by IGF PAN to calibrate hydrological models and it will also be used in this study. A Matlab script to apply DEGL has been provided by IGF PAN and in Table 8 the settings that are used are mentioned. In appendix 1 the parameter ranges in the calibration are presented. In the first calibration round all 14 parameters will be calibrated, while in the second calibration round the insensitive parameters (after a sensitivity analysis, see section 3.3.3) are fixed. The second calibration round, with only the most sensitive parameters, will be done to ensure that indeed the global optimum has been found in the first calibration round.

Table 8: Settings of the DEGL calibration procedure. *F*, C_r, r_k and w_r are adopted from best-performing variant that Das et al. (2009) report and these settings are also used earlier by IGF PAN to calibrate hydrological models.

Setting	Description	Calibration 1	Calibration 2 (on the most sensitive parameters)
M	Number of model parameters to calibrate	14	8
MaxFunEvals	Maximum number of model runs. Initially set by IGF PAN.	50000	50000
TolFun	Calibration is stopped if the objective function is very close to the optimum value of the objective function (Y = 1).	$1.0*10^{-16}$	$1.0*10^{-16}$
N_{pop}	Number of elements in the population. For DE Das et al. (2009) indicate that a reasonable size of the population is $5 * M$ to $10 * M$. $5 * D$ has been used earlier by IGF PAN and this is retained.	70	40
randchild	Number of individuals needed to generate random offspring.	2	2
F	Scaling factor. In literature values of F are in the range 0.4 – 1 (Das et al., 2009). 0.8 is adopted	0.8	0.8

	from the best performing set-up of Das et al. (2009).		
C_r	Crossover rate. Values that are mentioned in literature are in the range of 0.3 to 1, with higher values if the parameters are dependent. 0.9 is adopted from Das et al. (2009).	0.9	0.9
r_k	Radius size of neighbourhood for local model. Das et al. (2009) used a neighbourhood size of 10% of the population size (N_{pop}). Neighbourhood size is $2r_k + 1$, so 10% of population size. Adopted from Das et al. (2009).	3	2
w_r	Initial weight factor balancing the global and local neighbourhood models (Das et al., 2009). Randomly set between 0.05 and 0.95. After the initial value the weight factor is determined with a self-adaptive scheme, following the best performing set-up of Das et al. (2009).	$0.05 + rand(N,1) * 0.9$	$0.05 + rand(N,1) * 0.9$

In section 3.3.5 it will be explained how parameter uncertainty is estimated with the Generalized Likelihood Uncertainty Estimation (GLUE) approach. Another option to calibrate the parameters and provide parameter uncertainty is by the calibration scheme Shuffled Complex Evolution Metropolis algorithm (SCEM-UA), described by Vrugt et al. (2003). However, the main objective of the calibration is to obtain the optimum deterministic parameter set and it is considered that DEGL is a robust and efficient scheme for this purpose. It is also an option to calibrate the model with a Monte Carlo analysis, but this would require more model runs to obtain a reliable parameter set.

3.3.5 Parameter uncertainty analysis

To estimate parameter uncertainty and to provide more insight in the calibrated parameter values GLUE will be used. The idea of GLUE is that many different sets of parameters and model structures can fit the data acceptably well and therefore it may be impossible to distinguish between them as useful predictors of the system (equifinality concept) (Beven, 1993). GLUE is based on a Monte Carlo analysis, it is simple to understand and easy to implement and it allows to use the same objective function as before (Demirel et al., 2013a; Jin et al., 2010; Shen et al., 2012). The main limitations of the GLUE method, that have been discussed frequently in literature, are the subjective choices that it contains and not using a formal representation of model error (Stedinger et al., 2008; Thiemann et al., 2001). Demirel et al. (2013a) and Shen et al. (2012) distinguish three steps in the GLUE method:

1. *Definition of the likelihood function:* The objective function (*Y*) that is used in the calibration will also be used as likelihood function in the GLUE analysis. The model will be run over the calibration period.
2. *Sampling parameter sets:* Only the parameters for which it appeared that the output is sensitive will be varied (see sensitivity analysis in section 3.3.3). The GLUE analysis would result in very uncertain parameters for insensitive parameters, because a very wide range of parameter values will result in more or less the same output. These parameters are not interesting to include in the GLUE analysis because their influence on the output is small. Due to model simplification, i.e. setting fixed values to certain parameters, the uncertainty analysis is focused on parameters that are important for the simulations. Most studies use a uniform parameter distribution, because of a lack of knowledge about the true distributions (Demirel et al., 2013a). Another option is to use for example a normal distribution. This would mean that in fact an a priori estimation of

parameter uncertainty is defined and in addition it would be very subjective to choose the shape of such a distribution. A uniform parameter distribution will be applied with the earlier defined parameter ranges (see appendix 1). It has been chosen to apply 50000 model runs (equal to the number of model runs in the calibration) and parameter values are sampled with Latin hypercube sampling. Other studies used more model runs, for example Demirel et al. (2013a) used 120000 model runs, but they also used GLUE to calibrate the model. In this project the optimal parameter set has already been found by DEGL and GLUE is only applied to investigate uncertainty, so it is considered that 50000 model runs is sufficient.

3. *Threshold definition for behavioural model parameter selection:* The 50000 tested parameter sets are divided into behavioural (accepted) parameter sets and nonbehavioural (rejected) parameter sets. The threshold for behavioural parameter sets significantly affects uncertainty estimation and it should be determined very carefully (Zheng & Keller, 2007). However, in literature no suggestions have been found to determine an appropriate threshold. Obviously a lower threshold leads to wider parameter uncertainty (Jin et al., 2010). Different thresholds and percentages of the sample have been used in previous studies. Zak and Beven (1999) used 20% of the parameters sets as behavioural. However, absolute thresholds are of course dependent on the used model and application, while percentages are also dependent on the applied parameter ranges. So it is difficult to adopt an appropriate threshold from previous literature, and in addition no previous studies are found that applied GLUE with likelihood function *Y*. It has been chosen to apply a threshold value of 0.5, because it is considered that this gives acceptable model performance. It must be realized that choosing an appropriate threshold for behavioural model selection is a very subjective element in the GLUE method (Blasone et al., 2008).

The behavioural parameter sets provide an estimation of the parameter uncertainty, but this will not be used to generate ensemble flow forecasts. In section 1.6 it has been explained that parameter uncertainty will not be included in the flow forecasts, but in section 5.4 and in appendix 9 a simple implementation of parameter uncertainty is tested to investigate the effect of including hydrological model parameter uncertainty.

3.4 Update state procedure

In this section it is described how model states/storages in the hydrological model are updated. Hydrological models are far from perfect and in order to better represent the current situation in the catchment the model inputs, states or output need to be updated continuously (Wöhling et al., 2006). To obtain a more accurate forecast model, simulations can be combined with real-time data (Moradkhani et al., 2005; Werner et al., 2006). Werner et al. (2006) explain that both the model simulations and the available real-time data must be seen as sources of information about the behaviour of the catchment. Both will also contain some degree of uncertainty, resulting in differences between the observed data and the simulated data and these differences can be taken into account by a feedback mechanism (Refsgaard as cited in Werner et al. (2006)), with the objective to reduce uncertainty in the forecasts (Werner et al., 2006). Reliable operation of a hydrological forecasting model requires continuous correction of the forecast as observational data become available (Moradkhani et al., 2005). In section 3.4.1 different approaches to establish an updating procedure are explained. Next to that in section 3.4.2 an appropriate updating procedure is chosen and in section 3.4.3 it is explained how this procedure is incorporated in the hydrological model.

3.4.1 Updating approaches

Refsgaard (as cited in Werner et al. (2006)) distinguishes four main approaches of data assimilation, depicted in Figure 11 and further explained below the figure.

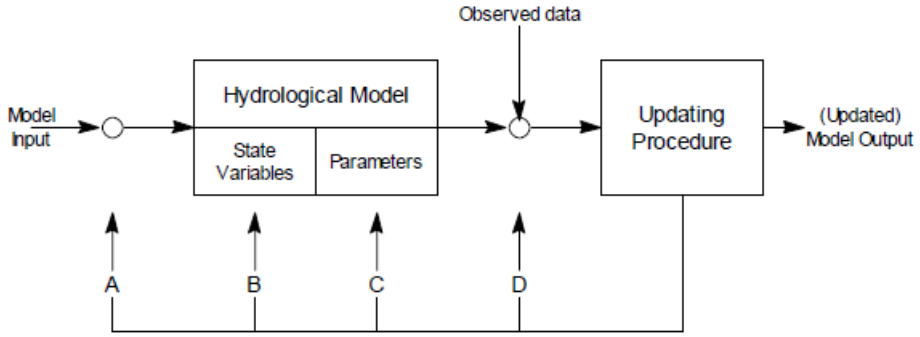


Figure 11: Schematization of 4 approaches of data assimilation (Refsgaard, as presented in Werner et al. (2006)):

- **A. Updating of input variables: Input variables are adjusted to minimize the differences between model output and observed variables (Werner et al., 2006).**
- **B. Updating of state variables (storages in HBV model): State variables in the hydrological model are updated (Werner et al., 2006; Wöhling et al., 2006). Measurements of actual states in the catchment are not available in this project, but actual discharge observations also provide information about actual states.**
- **C. Updating of model parameters: It is considered that flow is a non-stationary process, so a completely fixed model with constant parameters will not be able to characterize the catchment behaviour for all circumstances in the future (Young, 2002). This can be seen as a constant recalibration of the model (Werner et al., 2006).**
- **D. Updating of output variables or error prediction: Establish a statistical structure of the model error and based on this a forecast of the error is incorporated in the simulated value to obtain an updated forecast (Werner et al., 2006).**

The approaches differ in what is adapted to minimize the model error. Disadvantages of approach A are that the model itself is in the optimization loop and that the number of degrees of freedom results in a badly posed optimization problem (Werner et al., 2006). Werner et al. (2006) suggest that this approach is most applicable to situations where the water is already in the catchment (precipitation has already fallen in the catchment). Approach C is mainly applied to data-driven models, since in a more physically based model the parameters have a physical meaning and it is unlikely that these parameters will vary with the same dynamics as model errors (Werner et al., 2006). In addition, Serban and Askew (1991) state that updating of model parameters is not recommended because in most models the parameters are not independent and the modification of one parameter would require the modification of other parameters. Approach D is also not suitable in this project because in forecasting systems based on meteorological forecasts the statistical correlation of model errors is often not very persistent for longer lead times (Werner et al., 2006). In addition, for flood forecasting the performance is worse because in the vicinity of floods the error persistence between the measured and computed hydrographs is the smallest and errors show a tendency to oscillate rapidly and widely (Serban & Askew, 1991).

It can be concluded that approach B, the updating of state/storage variables, is most appropriate here. Initial conditions are important for the generation of runoff. State variable updating is an often used updating approach in hydrological forecasting (e.g. (Demirel et al., 2013a; Werner et al., 2006; Wöhling et al., 2006)). Otherwise initial conditions are taken from continuous deterministic near-real-time model runs (Verkade et al., 2013; Zappa et al., 2008). Demirel et al. (2013a) explain that the estimation of model states based on observed discharge is superior to deterministic near-real-time model runs, because there can be errors between simulated and observed discharge affecting the model states. Werner et al. (2006) mention that there is number of methods to bring this updating

approach into practice, ranging from a simple method like direct insertion, where a state variable is substituted by an observed variable to more advanced methods like Kalman Filter approaches. In the next section an appropriate updating procedure will be selected.

3.4.2 Selection of an updating procedure

In real-time flow forecasting often the Kalman Filter method and its derivatives are used (i.a. (Chen et al., 2013; Komma et al., 2008; Moradkhani et al., 2005; Serban & Askew, 1991; Weerts & El Serafy, 2006)). In addition a simpler model state updating procedure will be examined that is used by Demirel et al. (2013a). There are many more updating methods and some of them are mentioned in the discussion in section 5.1.3.

3.4.2.1 (Ensemble) Kalman Filter

The underlying idea of the Kalman Filter is to provide an estimate of a state vector based on model information and measurements and balancing out the errors of the two by minimizing the state error variance (Komma et al., 2008). Originally the Kalman Filter was developed for linear systems, but with the Extended Kalman Filter this can also be applied to nonlinear models (Serban & Askew, 1991). If the nonlinearities in the model are strong the linearization of the Extended Kalman Filter becomes very inaccurate, which led to the development of the Ensemble Kalman Filter (Burgers et al., 1998).

3.4.2.2 Direct model storage updating

Although the Kalman Filter is an often used updating procedure it is a complicated method to understand and apply. The method that Demirel et al. (2013a) used is simpler and it is a more directly physically based method. At first the hydrological model is run with the calibrated parameter set. The next step is to establish an empirical relationship between simulated discharge and the fraction of fast runoff to be able to divide the observed discharge in a fast and a slow runoff component (see Figure 10 for the HBV model structure). These components are used to update the fast runoff reservoir and slow runoff reservoir to what the storages should have been to generate the observed discharge. Because the model storages are updated directly with a simple relationship with observed discharge, this method will be called ‘direct model storage updating’ in this report.

3.4.2.3 Procedure choice

According to Chen et al. (2013) several studies showed that the Ensemble Kalman Filter method is easy to implement, offers flexibility in covariance modelling, is robust and computationally efficient. Weerts and El Serafy (2006) conclude that the Ensemble Kalman Filter is more robust in flow forecasting with real data than particle filters methods and they recommended to use the Ensemble Kalman Filter. However Xiong and O’Connor (2002) report that, although there are high expectations among hydrologists about the Kalman filter as an updating tool, also reservations are made. Xiong and O’Connor (2002) mention that in an earlier study updated forecasts produced by a simple standard autoregressive model were practically identical to results obtained with a Kalman Filter. Wöhling et al. (2006) have based their study on the idea that Kalman Filters are mathematically too complex to be easily implemented in the highly non-linear hydrological models, and they therefore developed a simple and effective updating procedure. Also Serban and Askew (1991) indicated that the efficiency of Kalman filters is still open to discussion with respect to phase and shape errors of floods. A disadvantage of the Ensemble Kalman Filter that is proposed by Weerts and El Serafy (2006) is that you need a priori uncertainty in states, precipitation input and discharge. There are simpler

methods, like autoregressive approaches and artificial neural networks (Xiong & O'Connor, 2002) and direct model storage updating (Demirel et al., 2013a). The effectiveness of direct model storage updating as proposed by Demirel et al. (2013a) and the Kalman Filter are not compared in an earlier study. It seems that the main reservation about the Kalman Filter is that simpler techniques produce nearly as good results. Because of this, because direct model storage updating is a simpler approach and because in the limited available time in this project it is not possible to compare different updating procedures, the direct model storage updating approach of Demirel et al. (2013a) will be applied. An updating procedure with an iterative process, like Wöhling et al. (2006) applied, will not be used to keep the procedure simple and computationally easy.

3.4.3 Implementation of direct model storage updating

With the approach of Demirel et al. (2013a) an empirical relationship between simulated discharge and the fraction of fast runoff is established to divide the observed discharge in a fast and a slow runoff component. These components are used to update the surface water (fast runoff reservoir) and groundwater (slow runoff reservoir) storages in the HBV model. For the storages that cannot be updated directly with observed discharge, because there is no direct link with discharge (see Figure 10), the deterministic near-real-time model runs are used with a minimum initialization period of 1 year. The formulas to update the surface water and groundwater storages are:

$$S_{sw}(t) = \left(\frac{k * q_0(t)}{K_f}\right)^{\frac{1}{1+\alpha}} \qquad [3.6]$$

$S_{sw}(t)$ = Surface water storage [mm]
$k = f(Q_0)$ = Fraction of fast runoff (Q_f/Q_t), based on an empirical relationship with observed discharge (in m^3/s) [-]
$q_0(t)$ = Observed discharge [mm/day]
K_f = Fast runoff parameter [d^{-1}]
α = Non-linearity parameter of fast runoff [-]

$$S_{gw}(t) = \frac{(1-k) * q_0(t)}{K_s} \qquad [3.7]$$

$S_{gw}(t)$ = Groundwater storage [mm]
K_s = slow runoff parameter [d^{-1}]

The relationship between total discharge and the fraction of fast runoff is yet unknown. In Figure 12 it is clearly visible that the fraction of runoff that comes from the fast runoff reservoir (*k*) increases when total discharge increases. This relationship could be expected, because runoff from the fast runoff reservoir (or surface water storage) reacts faster and to a larger extent to precipitation and snowmelt while outflow from the slow runoff reservoir represents the slower response from the groundwater (base flow). Total discharge will be used as basic variable to estimate *k*, analogously to Demirel et al. (2013a). However *k* as a function of total discharge shows a large uncertainty (see Figure 12). For example for a total simulated discharge of 10 m^3/s, *k* varies between 0 and about 0.6 and for a total discharge of 20 m^3/s *k* varies between about 0.3 and 0.7. This spread can be explained by the non-updated initial conditions. Different combinations of slow and fast runoff can result in the same total discharge and thus in a spread of *k* as a function of total discharge. Therefore it will be investigated whether using the non-updated storages from the HBV model can provide an improvement to the method of Demirel et al. (2013a). This gives the updating scheme in Figure 13.

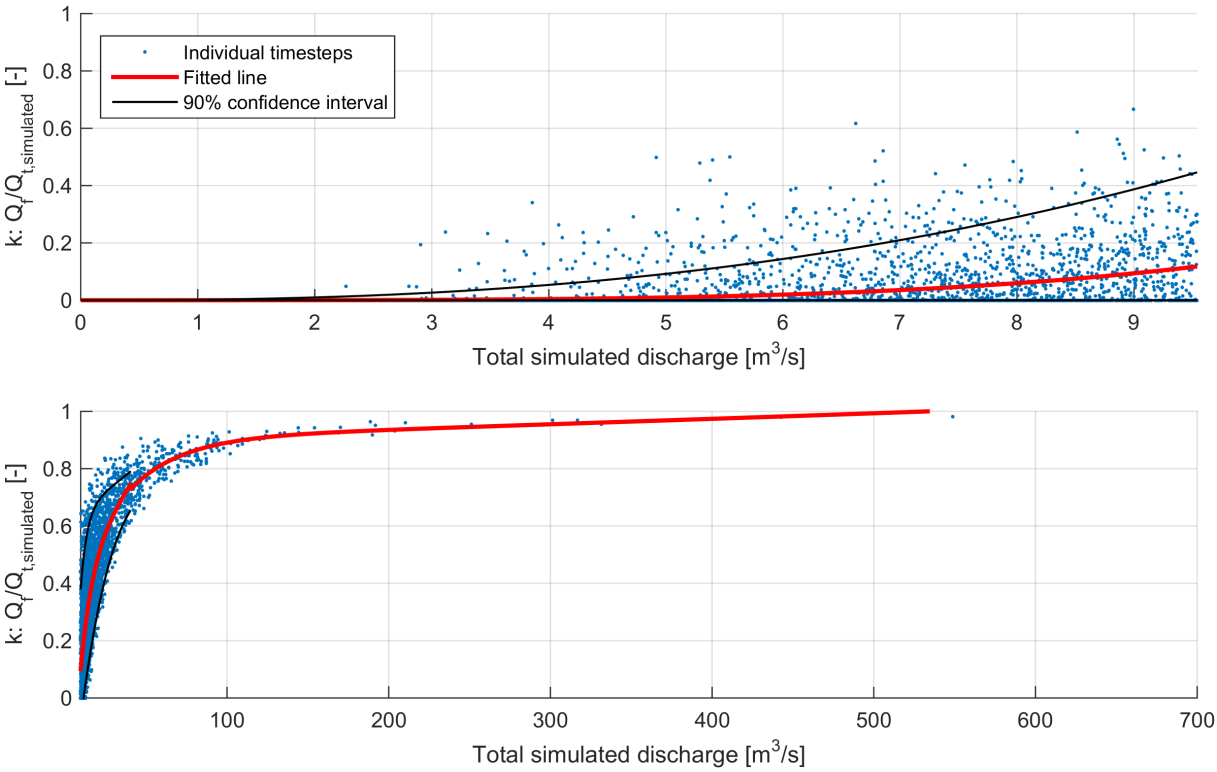


Figure 12: Fraction of fast runoff to total simulated discharge as a function of the total simulated discharge, established over the calibration period. See text below for an explanation about the fitted lines and the 90% confidence interval.

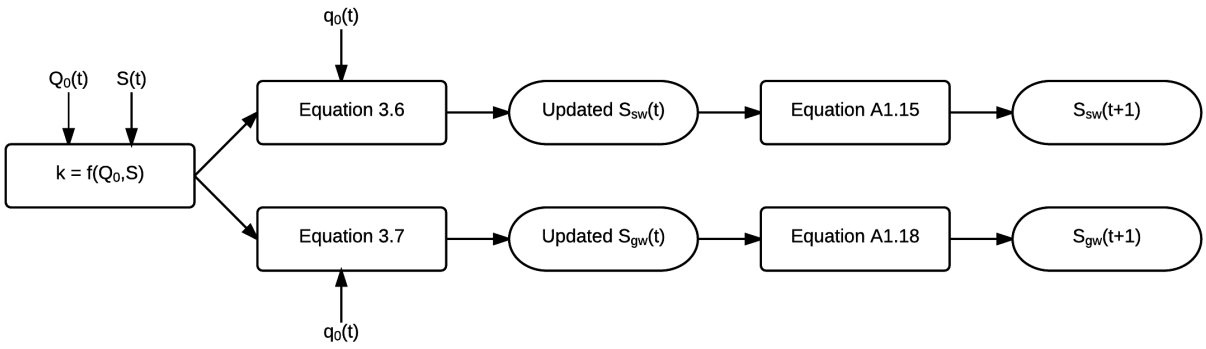


Figure 13: Updating scheme of direct model storage updating including the use of current storages from the HBV model

The aim is to relate the spread in *k* as a function of discharge to non-updated initial storages from the HBV model. A storage that could be useful is the soil moisture storage. However this gives problems when there is snow accumulation, because the storage can be high (wet conditions in the catchment) but due to snow accumulation precipitation will not form direct runoff and there is also no seepage then. So there can be high soil moisture storage, but no generation of fast runoff. Therefore it will be investigated whether the sum of direct runoff and seepage minus capillary transport on day *t-1* (this can form discharge at day *t*) can be used to explain the spread in *k*. The sum of these fluxes forms the net inflow into the fast runoff reservoir storage and is dependent on the soil moisture state and whether there is snow accumulation or not. Figure 14 provides insight into the relationship of the net inflow with the spread in *k*. Below Figure 14 it is explained how the figure has been constructed. It can be seen that the different lines in Figure 14 are close to each other, so the net inflow does not really explain the spread in *k*. This can be explained by the fact that also the initial storages in the fast runoff reservoir and slow runoff reservoir are important.

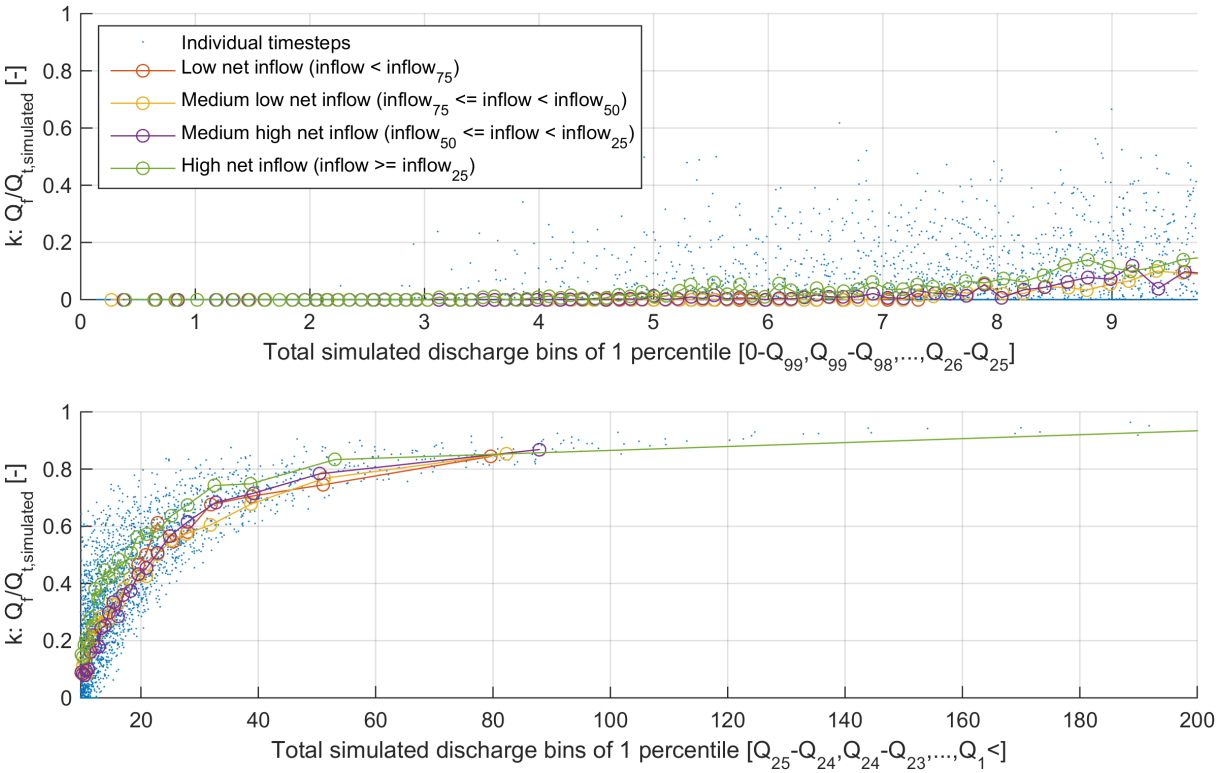


Figure 14: Fraction of fast runoff to total simulated discharge as a function of the total simulated discharge for different percentiles of net inflow (e.g. below the 25% percentile) in percentile bins of simulated discharge. Discharges within one percentile bin of simulated discharge (on x-axis) are considered as one discharge, so different net inflows give different *k* for the same discharge.

A second option is to relate the spread in *k* to the non-updated storage of the fast runoff reservoir (S_{sw}) or slow runoff reservoir (S_{gw}). The fast runoff reservoir has the highest correlation with *k* for high and low discharges, so *k* will be related to discharge and non-updated initial fast runoff reservoir storage. Figure 15 provides insight into the relationship of these storages with the spread in k. With an high initial storage in the fast runoff reservoir storage the fraction of fast runoff is higher, which makes sense because of the definition of *k*. For discharges larger than 40 m^3/s the difference in *k* related to differences in initial fast runoff reservoir storage are much smaller and also the spread of *k* is less, so for this spectrum it is decided to relate *k* only to discharge.

It is possible to fit lines through the points in Figure 15, but the question would be which line should be applied to which situation. After all, the lines do not apply to a certain percentile of all initial fast runoff reservoir storages in the calibration period, but they apply to a percentile of storages within one discharge bin of 1 percentile. *k* should be related to both discharge and initial surface water storage, which can be established by surface fitting in the Curve Fitting Toolbox of Matlab. In Figure 12 the fitted lines if *k* is only related to discharge are plotted and in Table 9 all resulting equations are given. The simplest (least parameters) polynomial functions are used that fit the data acceptably well. This has been done in multiple parts, because there are many points in a certain part of the domain and as a result of this in other parts large deviations from one fitted line would occur. The thresholds Q_{25} (exceedance probability of 25%) and 40 m^3/s have been used for this. The threshold of 40 m^3/s has been chosen in line with what will be used when also initial storages are incorporated. In addition the 'confidence intervals' of *k* related to discharge and initial storage are presented in Figure 12, based on the 5% percentile and 95% percentile lines like in Figure 15. These confidence intervals will be used as upper and lower limit for the surface plots, to prevent that unrealistically large

deviations from the relationship in Figure 12 can occur. This was not possible for the highest flow category, because there are too few points here. For the low category of flows ($Q_0 \leq Q_{25}$) the maximum difference between the plotted surface function and the points is 0.04 and for the medium flow category ($Q_{25} < Q_{obs} \leq 40$ m^3/s) this is 0.02. As explained above, for the highest flow category k is only related to discharge. In appendix 3 the contour plots of the surface functions are shown, to give an indication about how k depends on both discharge and non-updated storages. Also the fitted surface plot for the highest flow category is added.

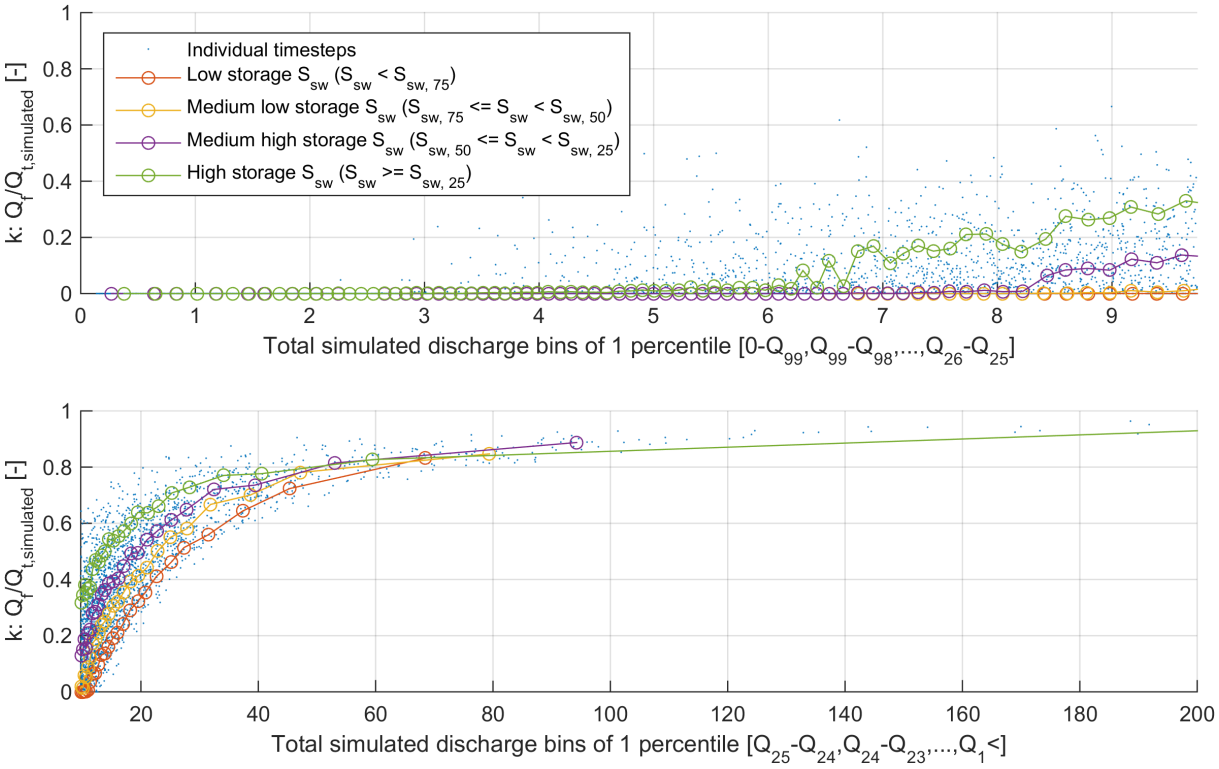


Figure 15: Fraction of fast runoff to total simulated discharge as a function of the total simulated discharge for different percentiles of initial fast runoff reservoir storages in percentile bins of simulated discharge.

Table 9: Fitted and implemented lines to relate the fraction of fast runoff to observed discharge and initial fast runoff reservoir storage

	$Q_0 \leq Q_{25}$	$Q_{25} < Q_0 \leq 40$ m^3/s	40 m^3/s $< Q_0$
k* related to *Q	$k = 1.9 * 10^{-5} * {Q_o}^{3.9} + 1.6 * 10^{-4}$	$k = 0.49 * \exp(0.011 * Q_o) - 1.4 * \exp(-0.12 * Q_o)$	$k = \min(1, 0.90 * \exp(2.0 * 10^{-4} * Q_o) - 0.64 * \exp(-0.031 * Q_o))$
***k* related to *Q* and S_{sw}**	$k = \max(0, 3.4 * 10^{-4} - 3.5 * 10^{-4} Q_o + 0.84 S_{sw} + 9.3 * 10^{-5} {Q_o}^2 - 0.14 Q_o S_{sw} + 0.34 S_{sw}^2 - 9.0 * 10^{-6} {Q_o}^3 + 0.0081 {Q_o}^2 S_{sw} - 0.030 Q_o S_{sw}^2)$	$k = \max(0, 0.30 - 0.070 Q_o + 0.56 S_{sw} + 0.0050 {Q_o}^2 - 0.044 Q_o S_{sw} + 0.066 S_{sw}^2 - 1.1 * 10^{-4} {Q_o}^3 + 0.0011 {Q_o}^2 S_{sw} - 0.0024 Q_o S_{sw}^2)$	For this spectrum the relationship between *k* and only Q_0 will be used (explained in the text).

There are three points in the HBV model where this updating of storages could be implemented, namely at the beginning of the time step, after calculation of the fluxes or after calculation of the initial storages for the next time step. Implementing the updating procedure at the beginning of the time step means that the observed discharge is always perfectly simulated, because the observed discharge of the same time step is divided over the surface water storage and groundwater storage. This is unrealistic, since in practice the discharge observations are not available at the beginning of a time step. Implementing the updating of storages at the end of the time step to calculate the initial storages for the next time step means that always the observed discharge of the previous time step is exactly simulated. Therefore the updating procedure will be implemented after calculation of the

fluxes and before the calculation of the new storages. In this way the fluxes depend on the initial storages as they were calculated in the previous time step and the fluxes also have influence on the initial storages for the next time step. Regarding the updating procedure in which *k* is related to discharge and the fast runoff reservoir storage, for the fast runoff reservoir storage previously updated or non-updated storages can be used. When updated storages are used, the fast runoff reservoir storage that is used in the formula to determine *k* has been updated in the previous time step. On the other hand, when non-updated storages are used the storage that is used is adopted from a deterministic model run without updating. Both variants will be tested, so in total 3 updating procedures are tested.

3.5 Pre-processing and post-processing of ensemble forecasts

In this report (from a hydrological point of view) the term pre-processing will be used for the correction of input variables (precipitation and temperature forecasts) and the term post-processing will be used for the correction of output (flow forecasts). Several studies mention pre-processing and post-processing as an important step to improve meteorological and hydrological forecasts (i.a. (Bürger et al., 2009; Cloke & Pappenberger, 2009; Olsson & Lindström, 2008; Roulin & Vannitsem, 2005; Verkade et al., 2013; Zalachori et al., 2012)). At first it will be outlined why processing of input and output might be necessary (section 3.5.1), then a suitable data processing technique is chosen (section 3.5.2) and different processing strategies are developed (section 3.5.3).

3.5.1 Data processing

The reasons for pre-processing are described in section 3.5.1.1 and the reasons for post-processing are explained in section 3.5.1.2.

3.5.1.1 Meteorological data pre-processing

Obviously the quality of the flow forecasts is highly dependent on the quality of the meteorological forecasts. Several studies indicate that forecasts from meteorological models need some form of prior correction before being used in hydrological models (Cloke et al., 2013). This is related to data assimilation approach A (Figure 11). In the study of Tao et al. (2014) (statistical) processing of meteorological forecasts from ECMWF substantially improved the raw precipitation forecasts. This is also found by Bürger et al. (2009), who downscaled ECMWF's ensemble forecasts to obtain precipitation and temperature for a small river catchment in Germany (50 km^2). Bürger et al. (2009) compared their results with the study of Roulin and Vannitsem (2005). This is a comparable ensemble flood forecasting study applied to the Demer (1775 km^2) and Ourthe (1616 km^2) in Belgium. However much lower skill scores has been found in this study. Roulin and Vannitsem (2005) did not apply a processing procedure at all and Bürger et al. (2009) contribute the difference in skill to the pre-processing step that they applied. Problems that were observed in precipitation output from GCMs are a low number of dry days, a bias in the mean and the inability to reproduce high precipitation events (Piani et al., 2010). Pre-processing of meteorological forecasts is needed because of the points mentioned below:

1. *Scale of weather forecasts:* The numerical weather forecasting products are often generated at a too coarse spatial scale compared to relevant catchment scales (Cloke et al., 2013; Schaake et al., 2010; Tao et al., 2014). According to Bürger et al. (2009) the spatial scale of typical flood producing weather systems is often much below the smallest scales of numerical weather

predictions. Downscaling is required to obtain probabilistic meteorological forecasts at an appropriate scale (Bürger et al., 2009; Cloke & Pappenberger, 2009).

2. *Grid values:* The grid values provided by ECMWF should not be considered as representing the weather conditions at the exact location of a grid point, but as a time-space average within a two- or three-dimensional grid box (Persson & Andersson, 2013). Numerical weather prediction systems are based on large computational cells, while hydrological models usually need data at a finer resolution or direct point measurements (Kiczko et al., 2015). When the meteorological forecasts are compared with point observations, additional systematic and non-systematic errors are introduced, due to the unrepresentativeness of the location and the sub-grid variability, partly appearing as under-dispersivity and bias (Persson & Andersson, 2013).
3. *Under-dispersivity:* In the case of under-dispersivity of meteorological forecasts they do not display the variability of the observations and underestimate uncertainty of the forecasts (Cloke & Pappenberger, 2009; Schaake et al., 2010). Renner et al. (2009) mention that in their study the global ECMWF ensemble forecasts do not produce enough variability, because their spatial resolution is very coarse compared to the hydrological scales.
4. *Systematic bias:* The meteorological forecasts often contain systematic biases (Cloke & Pappenberger, 2009; Schaake et al., 2010). This means that there is a difference between climatological statistics of ensemble predictions and corresponding statistics of observations (Cloke & Pappenberger, 2009). Wilks (as cited in Schaake et al., 2010) states that bias and dispersion errors in meteorological forecasts arise from limitations in the description of the physical processes in the meteorological models, errors in the subgrid-scale parameterisations and uncertainties in the initial state.

To correct for point 1 and 2, to obtain meteorological data at a relevant hydrological scale and area, dynamical or statistical downscaling should be applied. Dynamical downscaling is a model-based methodology, in which sub-grid scale meteorological features are obtained by an increase of the spatial resolution of a physical model (Boé et al., 2007). Statistical downscaling is characterized by a statistical relationship between a local variable (predictand) and large-scale variables (predictors), modelled by a global or regional model (Wetterhall et al., 2012). The ECMWF data has been downloaded at a resolution of 0.25^{o} x 0.25^{o} and 7 grid cells are used to cover the catchment (see Figure 5). The meteorological ensemble forecasts will not be downscaled further, because a lumped hydrological model is used and it is considered that the scale of the meteorological forecast grids is already comparable with the scale of the application.

In climate studies after dynamical downscaling the output of Regional Climate Models (RCM) still needs a bias and dispersion correction step before the meteorological forecasts can be applied to hydrological models (point 3 and 4) (Wetterhall et al., 2012; Wood et al., 2004). The results of the RCM still contain biases, especially concerning precipitation, and therefore raw RCM results need to be corrected (Wood et al., 2004). Also in the study of Renner et al. (2009) the forecasted precipitation from the COnsortium for Small-scale MOdeling Limited Area Ensemble Prediction System (COSMO-LEPS) still contain bias and dispersion errors. COSMO-LEPS is a dynamic downscaling approach of ECMWF ensemble forecasts (Renner et al., 2009). Bias and dispersion errors can be improved by bias and dispersion correction techniques. With a bias and dispersion correction technique a statistical relationship between simulated forecast variables and observations is established. Bias and dispersion correction is distinguished from statistical downscaling, because the

forecast variables are not downscaled to a higher resolution. Many different bias and dispersion correction techniques exist, this is further explained in section 3.5.2.

3.5.1.2 Post-processing of flow forecasts

By correction of meteorological forecasts and using the simulated flow forecasts without correction it is assumed that the hydrological model is perfect. However, hydrological models also often introduce simulation biases that degrade forecast quality (Hashino et al., 2007). Instead of correction of meteorological input data an alternative method is to correct for bias and dispersion errors in the outcomes of the hydrological model (Cloke & Pappenberger, 2009). This is related to data assimilation approach D (Figure 11). By post-processing of flow forecasts based on raw meteorological forecasts against observed discharges, meteorological and hydrological uncertainties are treated together (Verkade et al., 2013). It is also possible to post-process the flow forecasts based on a correction between observed discharge and simulated discharge generated with observed precipitation and temperature input ('perfect forecasts'). This correction accounts for hydrologic uncertainties, where it is assumed that meteorological bias and dispersion errors have been addressed adequately in a correction step of meteorological forecasts (Verkade et al., 2013; Zhao et al., 2011).

Most studies indicate that post-processing of flow forecasts is more effective to improve the forecast skill than pre-processing of meteorological input data. The results of Verkade et al. (2013) indicate that after pre-processing of meteorological forecasts the improvements in these forecasts were modest, especially regarding precipitation. Moreover the improvements in precipitation and temperature did not translate proportionally into improvements in the flow forecasts (Verkade et al., 2013). In addition, pre-processing of meteorological forecasts is generally complex, resource intensive and imperfect regarding space-time covariability (Verkade et al., 2013). Verkade et al. (2013) therefore propose other methods to improve the skill of flow predictions, such as data assimilation and post-processing techniques. According to Kang et al. (2010), to generate flow forecasts pre-processing should always be accompanied with post-processing to effectively reduce the hydrological model uncertainty. Zalachori et al. (2012) found that pre-processing of meteorological ensemble forecast data improves the skill of flow forecasts, but correction of flow forecasts results in a higher improvement. They conclude that the errors linked to hydrological modelling remain a key-component of the total predictive uncertainty of hydrological ensemble forecasts. Although correction of meteorological forecasts is of high importance to obtain reliable inputs to the hydrological model, corrections made to meteorological forecasts lose their effect when propagated through the hydrological model (Verkade et al., 2013; Zalachori et al., 2012). The results of Zalachori et al. (2012) indicate that implementation of pre- and post-processing techniques together gives the highest forecast quality. Zalachori et al. (2012) did not test exclusively post-processing of flow forecasts based on raw meteorological forecasts. Olsson and Lindström (2008) found that the spread in the ensemble flow forecasts (based on raw meteorological forecasts) was systematically underestimated and they improved this by post-processing of the flow forecasts. Olsson and Lindström (2008) conclude that a simple calibration on simulated ensemble forecasts can substantially improve the ensemble spread. According to Olsson and Lindström (2008) total adjustment of discharge output has advantages over applying both pre-processing and post-processing, because one combined adjustment requires less effort in design and management of the system and with separate adjustments it has to be verified whether integrating these techniques in one system is indeed able to generate accurate ensemble flow predictions.

3.5.2 Pre-processing and post-processing techniques for ensemble forecasts

The previous section explains why bias and dispersion correction are required in a pre-processing and post-processing step. There are many bias and dispersion correction techniques, partly similar to techniques that are used in statistical downscaling. Techniques that have been used in previous studies are the delta method for pre-processing (Déqué, 2007; Wetterhall et al., 2012), distribution-based scaling for pre-processing (Wetterhall et al., 2012), (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 (Kang et al., 2010; Madadgar et al., 2014; Shi et al., 2008), (conditional) regression for pre-processing (Verkade et al., 2013) and post-processing (Hashino et al., 2007; A. Ye et al., 2015), event bias correction for pre-processing (Hashino et al., 2007), Copula functions for post-processing (Madadgar et al., 2014), Bayesian joint probability modelling for post-processing (Bennett et al., 2014) and more, sometimes combined with the Schaake shuffle to ascribe realistic spatial and temporal patterns (e.g. (Bennett et al., 2014; Verkade et al., 2013)). With conditional correction additional predictors are used (Verkade et al., 2013).

Wetterhall et al. (2012) compared three simple correction techniques for precipitation forecasts from a RCM: the delta method, QM and distribution-based scaling. The more sophisticated methods QM and distribution-based scaling performed better than the delta method (Wetterhall et al., 2012). The distribution-based scaling that was conditioned on circulation patterns resulted in the best performance, although the scores are not statistically significant higher than for the other methods (Wetterhall et al., 2012). With the delta method all forecasts are corrected with one factor. Wetterhall et al. (2012) suggest that in a study where floods are also important the delta method is not appropriate and in addition the delta method has the major limitation that there is no correction for dispersion errors. So the delta method will not be used here. QM is easier to establish than distribution-based scaling, because with QM the probability density function is based on observations while with distribution-based scaling a theoretical density function needs to be defined first. QM is the most popular post-processing technique in hydrological forecasting (Madadgar et al., 2014). According to Kang et al. (2010) QM generally performs well in both pre- and post-processing. Hashino et al. (2007) advise to use QM, looking at the good performance in sharpness and discrimination and the simplicity of this method. This recommendation by Hashino et al. (2007) is also mentioned by Cloke and Pappenberger (2009). According to Wilks and Hamill (2007) there is no unambiguous answer to what is the most efficient correction technique for all applications. Finding the 'optimal' technique could be a complete research on its own, so it is chosen to apply QM to both pre-processing of meteorological forecasts and post-processing of flow forecasts. In an earlier study of IGF PAN Kiczko et al. (2015) focused on the same river catchment and applied QM to correct the ECMWF forecasts. In the study of Kiczko et al. (2015) QM significantly improved the prediction skills of the forecasting system.

3.5.2.1 Quantile Mapping

The principle of QM is shown in Figure 16. The cumulative distribution function (CDF) of forecasts over a control period is matched to the CDF of the observations over the same period, after which a correction function is generated (Boé et al., 2007). This is based on the idea that the meteorological prediction model is able to correctly simulate ranked categories of the variable, for example that 'very high precipitation' is correctly simulated by the model and by QM this is corrected to observed 'very high precipitation' (Déqué, 2007). This means that the correction is conditional on the value of the forecasted variable itself.

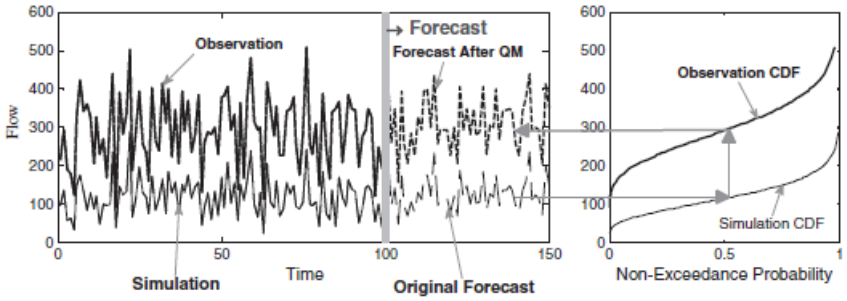


Figure 16: Principle of bias correction by QM (Madadgar et al., 2014). At first the CDFs of the forecasts and observations are made over the training period. To correct a forecast, the probability of non-exceedance is extracted from the CDF of the forecasts and in the CDF of the observations the corresponding observation value is found.

3.5.3 Application of Quantile Mapping in different strategies

At first QM is applied as pre-processing technique in different set-ups (section 3.5.3.1). Then different processing strategies (combinations of pre-processing and post-processing) are proposed (section 3.5.3.2).

3.5.3.1 Pre-processing set-up

In section 4.1.4 it will be explained that there are doubts about the quality of observation data during 2008 and especially during 2007. Therefore the empirical CDF of the observations and forecasts will be established on the training period 1-11-2011 to 31-10-2013 (two hydrological years) and validated over the period 1-11-2007 to 31-10-2011. This training period is very short compared to other studies, but also a sufficient period of data is needed for evaluation and this should not overlap with the training period.

Several set-ups of the QM procedure will be tested for pre-processing (see Table 10). The set-ups differ in (not) distinguishing different lead times and seasons. Establishing a different CDF per ensemble member does not make sense, because ensemble member 1 is not specifically related to ensemble member 1 of a new forecast. The meteorological forecasts are first aggregated to catchment-average forecasts (see section 2.3) and then corrected against catchment-average observations, so that the corrections are applied at the same spatial level as the level at which the forecasts will be applied.

Table 10: QM set-ups for pre-processing of precipitation and temperature forecasts

Set-up	Lead times	Seasons
Set-up 0	No correction	
Set-up 1	All lead times together	No seasonal distinction
Set-up 2	Separately for lead times from 0 days to 9 days, like Verkade et al. (2013) also applied. The distributions can be different for different lead times, so it is expected that this gives better performance than set-up 1.	No seasonal distinction
Set-up 3	Idem as set-up 2	Seasonal distinction in a summer and winter season. Bias can depend on weather type, so previous studies have applied independent QM distributions for different seasons (Boé et al., 2007) or weather patterns (Wetterhall et al., 2012). Only two seasons will be distinguished, because the available training period is very short. The winter period is defined as November until April and the summer period is defined as May until October.

After establishing the empirical CDFs of observations and forecasts during the training period, the forecasts are matched to the CDF of the observations. In accordance with Boé et al. (2007), if forecasts exceed the maximum or minimum forecast during the training period a simple extrapolation is applied based on the edges of the CDFs. Otherwise all forecasts would be bounded by the maximum observation during the training period. As with correction for elevation (section 3.1), to precipitation a relative correction is applied if the forecasts that will be corrected exceed the maximum forecast during the training period and to temperature an absolute correction is applied equal to the correction at the edge.

Boé et al. (2007) and Piani et al. (2010) mention the problem that meteorological models tend to forecast a very low precipitation instead of zero precipitation (tending to drizzle). The relative frequency of nonzero precipitation in the forecasts is thus greater than the relative frequency of nonzero precipitation in the observations. To ensure that the probability of nonzero precipitation in the forecasts is equal to the probability of observed nonzero precipitation, if the non-exceedance probability of a forecast is lower than the non-exceedance probability of 0 mm precipitation in the observations the precipitation forecast is set to 0 mm.

The ECMWF forecast data are not corrected for elevation like observation data. It is assumed that the grid cells of the ECMWF forecasts are representative for their part of the catchment, so it is not needed to correct for this. Persson and Andersson (2013) dispute this. However, consistent bias is also corrected by QM.

In Figure 17 and Figure 18 two example CDFs of precipitation and temperature are presented. In both figures over the largest part of the cumulative probability domain the CDFs of the forecasts and observations are quite close to each other, so no large corrections are made here. Regarding precipitation the correction (from CDF of forecasts to CDF of observations) is over the largest part of the cumulative probability domain in the same direction and of about the same magnitude for the summer and winter period. Based on this it is expected that QM with seasonal distinction will not result in an improvement of precipitation forecasts compared to QM without seasons. Important corrections are made near a precipitation 0 mm, where for the winter period all forecasted non-exceedance probabilities below 0.35, corresponding to a forecasted precipitation of 0.24 mm/day, are corrected to 0 mm/day and for the summer period all forecasted non-exceedance probabilities below 0.42 (forecasted precipitation of 0.27 mm/day) are corrected to 0 mm/day. For very extreme precipitations, above a non-exceedance probability of 0.99, also large corrections are made.

In Figure 18 it is clearly visible that the CDF of temperature is much different in winter than in summer and also the direction of corrections is different. For the winter period the CDF of the observations is quite close to the CDF of the forecasts, while for the summer period a bias is present over a large part of the cumulative probability domain. Since the correction is different in summer and winter it might be expected that QM with seasonal distinction will result in an improvement of temperature forecasts compared to QM without seasonal distinction. However the effectiveness of QM to improve the meteorological forecasts totally depends on whether during the validation period the same conditional bias is present as during the training period.

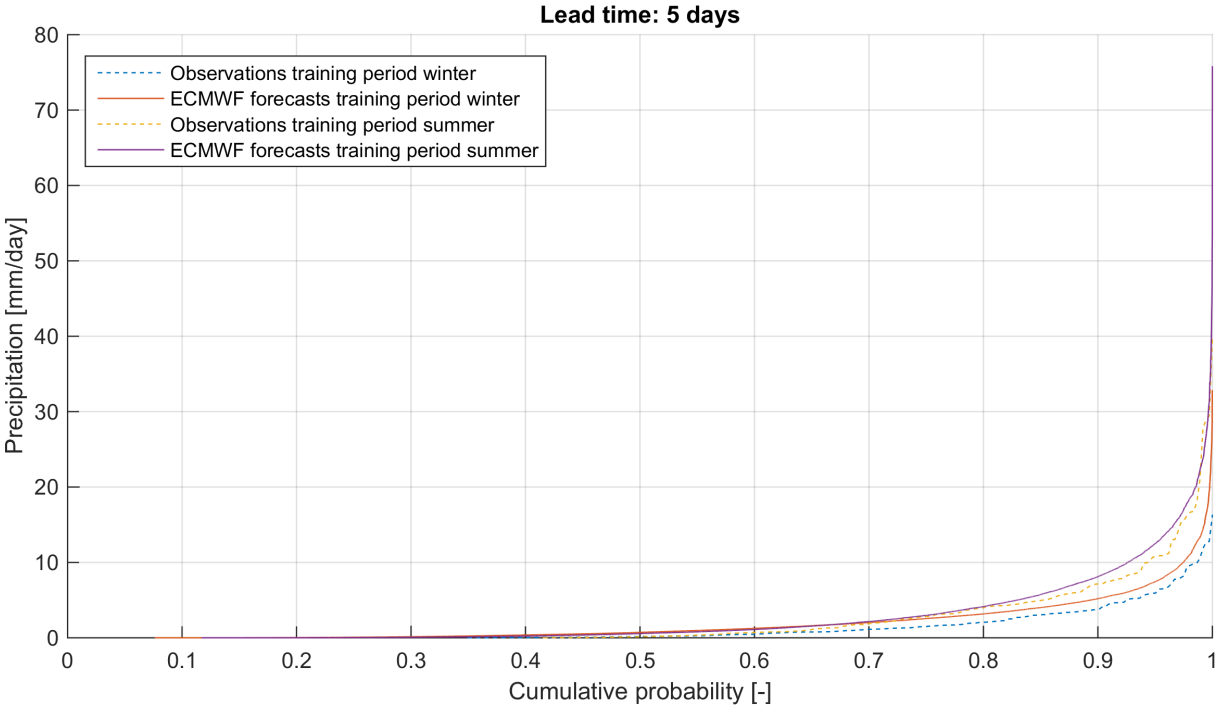


Figure 17: Example CDFs of precipitation observations and forecasts for two seasons, training period 2012-2013

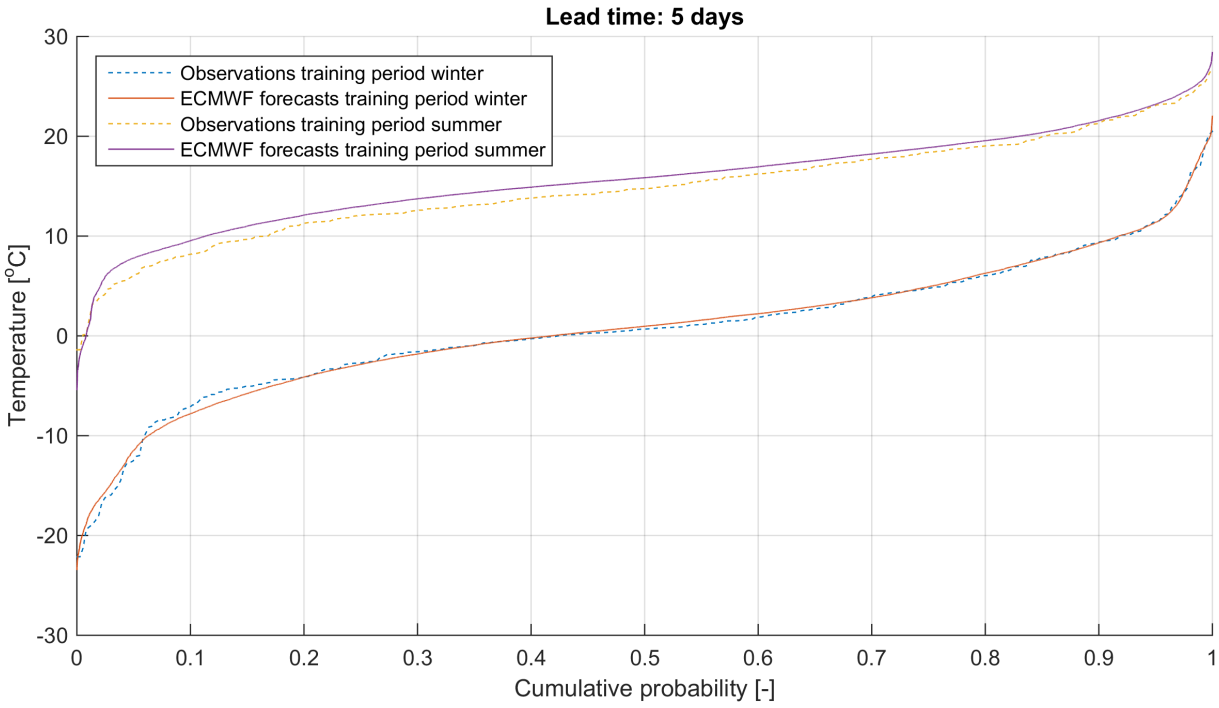


Figure 18: Example CDFs of temperature observations and forecasts for two seasons, training period 2012-2013

3.5.3.2 Processing strategies

In Table 11 different combinations of pre-processing and post-processing are presented. The best pre-processing set-up will be chosen as pre-processing approach (see the results in section 4.2.1). Post-processing will be applied separately for lead times from 1 day to 10 days and it is tested whether QM with seasonal distinction or QM without seasonal distinction gives better results. Forecasted flows that are beyond the maximum and minimum forecasted flows during the training period are corrected in the same way as precipitation, so the correction is equal to the relative correction of the edge. Discharge is the same kind of variable as precipitation (not bounded by zero).

When post-processing is applied (strategy 2 and 3) the deterministically simulated discharges based on observed precipitation and temperature data are also corrected, because these simulations will also contain hydrological model bias. This is important for the evaluation of flow forecasts (see section 3.7.1). For strategy 3 this is done like in strategy 2, so based on the difference between deterministically simulated discharges and observed discharges.

Table 11: Pre-processing and post-processing strategies

Strategy	Pre-processing	Post-processing
Strategy 0	No correction	No correction
Strategy 1	Precipitation: QM separately for lead times from 0 days to 9 days and no seasonal distinction (set-up 2). Temperature: QM separately for lead times from 0 days to 9 days and seasonal distinction (set-up 3). These are the best performing QM set-ups for pre-processing (see section 4.2.1).	No correction
Strategy 2	Idem as strategy 1	Correction of flow forecasts based on the correction between deterministically simulated discharges with observed temperature and precipitation data and observed discharges: A. Without seasonal distinction B. With seasonal distinction (summer and winter) It is expected that this will not result in a large improvement, because 'easy' consistent bias should already be captured by the calibration procedure.
Strategy 3	No correction	Correction of flow forecasts based on the correction between flow forecasts generated with raw meteorological forecasts and observed discharges. A. Without seasonal distinction B. With seasonal distinction (summer and winter)

The possible strategy of no pre-processing and post-processing like strategy 2 will not be tested, because this strategy does not make sense when there is looked to which errors can be present in the flow forecasts. Meteorological forecast errors would be intentionally not corrected.

3.6 Ensemble flow forecasting system

Meteorological ensemble forecasts from ECMWF are used as input data to the hydrological model to generate flow forecasts. Figure 19a presents the model set-up to generate ensemble flow forecasts. The forecast day refers to the first day for which no observations are available and meteorological forecasts should be used to generate flow forecasts. If for example the forecasts are made on 1-11-2006, this is the forecast day. It is considered that when the forecasts are made on this day no observations of precipitation, temperature and discharge available for 1-11-2006 and later, so meteorological forecasts should be used. A lead time of 0 days refers to the forecast day.

The update state procedure (section 3.4) provides initial model storages for the forecast day. In the updating procedure the surface water storage and groundwater storage are updated by using observed discharge, but for the soil moisture storage, melt water storage and snowpack the analogous deterministic model run is used with an initialization period of at least 1 year. At a lead

time of 0 days all ensemble flows are equal to the deterministic flow from the HBV model, because the flow on this day is fully determined by the initial states. The spread of meteorological forecasts at a lead time of 0 days has effect on the flow forecasts at a lead time of 1 day. Each ensemble member develops independently. So ensemble member 1 of the precipitation forecasts and ensemble member 1 of the temperature forecast are used as input to the hydrological model and the initial states following from this are used again with ensemble member 1 as input data. The initial states are not updated after day 0, because there are no observations available in future. This results in 'spaghetti hydrographs' for each day in the period from 1-11-2006 to 31-10-2013, consisting of 51 ensembles with lead times from 0 days to 10 days. Flow forecasts do not apply to a certain time (like 12:00), but they represent the flow on a calendar day. The hydrological model also runs at a daily time interval.

Besides ensemble flow forecasts also deterministically simulated flows are generated (see Figure 19b), based on the same updated initial states at day 0 and observed precipitation and temperature as input data. These flows are also called 'perfect forecasts' (Olsson & Lindström, 2008).

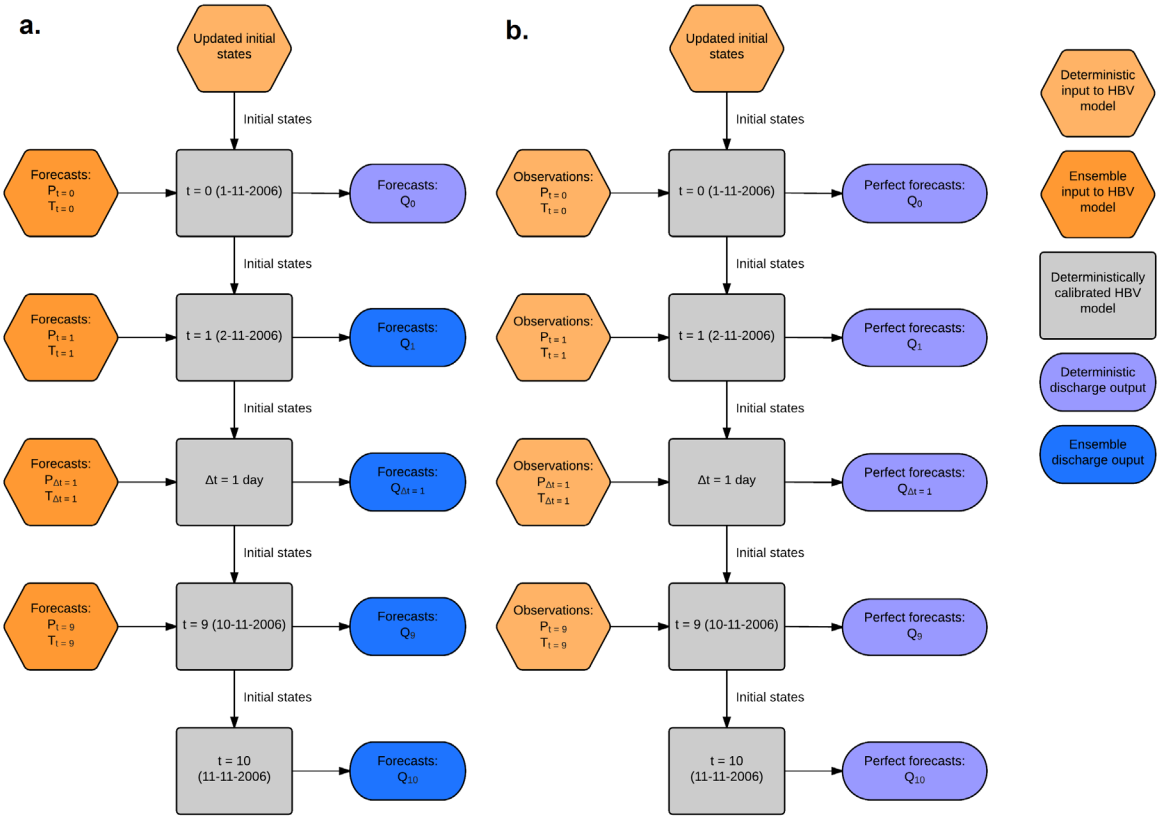


Figure 19: a. Model set-up to generate ensemble flow forecasts b. Model set-up to generate deterministically simulated forecasts ('perfect forecasts')

3.7 Evaluation criteria

In this section it is explained how the ensemble flow forecasting system will be evaluated. Forecast evaluation allows to monitor the forecasts, it provides an objective comparison between different forecast sets and it enables to improve the forecast quality by discovering strengths and deficiencies (Renner et al., 2009; Wilks, 2006). At first it is explained which flows form the reference flows to which the ensemble forecasts are compared (section 3.7.1) and after that appropriate evaluation scores are selected (section 3.7.2 to 3.7.7).

3.7.1 Reference flows

The ensemble flow forecasts are compared to reference flows. The reference flow determines what is actually evaluated. The evaluation of ensemble flow forecasts against observed discharge contains error components from the meteorological forecasts, the hydrological model and discharge observation measurement errors (Renner et al., 2009). Hydrological model errors also include errors in the (updated) initial conditions. By using perfect flow forecasts (see Figure 19b) as reference flows, the hydrological model error component and discharge observation measurement errors are eliminated (Demargne et al., 2010; Olsson & Lindström, 2008; Renner et al., 2009). If it is assumed that observation errors can be neglected, evaluation against observed discharge contains errors from the meteorological forecasts and the hydrological model and evaluation against perfect flow forecasts contains errors from the meteorological forecasts (Demargne et al., 2010; Olsson & Lindström, 2008; Renner et al., 2009). Observed discharge will be used as basic score to investigate the limitations of the forecasting system and the results will be compared with evaluation based on perfect flow forecasts to investigate whether hydrological model errors or meteorological forecast errors are dominant (see Figure 20). It is not possible to compare evaluation scores of flow forecasts and meteorological forecasts, because the general evaluation score (explained in section 3.7.3) depends on the magnitude of the investigated parameter.

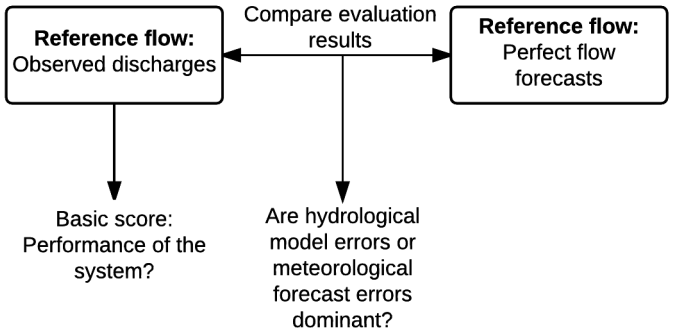


Figure 20: Evaluation approach

3.7.2 Evaluation scores

The evaluation scores for ensemble forecasts are different from common verification scores for deterministic forecasts, because it is not useful to compare individual ensemble members to the reference discharge (Renner et al., 2009). After all the ensemble forecasts should be used as a set. Forecasts from an ensemble flow forecasting system can be considered as probabilistic forecasts and the evaluation procedure should be arranged to that. There are many evaluation scores available. A single evaluation score is generally inadequate to evaluate the overall performance of a forecasting system (Demargne et al., 2010; Hamill et al., 2000). Three properties of forecast quality are reliability, sharpness and resolution (WMO, 2015). The different scores that will be used should evaluate the different properties of forecast quality. From the workshop on 'Ensemble Forecasting in the Short to Medium Range' it has been concluded that at least the following measures should be used (Hamill et al., 2000):

1. Probabilistic (skill) scores such as the Brier Score and/or the Continuous Ranked Probability Score (*CRPS*). These scores can provide an overall, single-number score for judging the quality of probabilistic forecasts. In this study the *CRPS* will be used, because this is a generalization of the Brier Score over more than one probability threshold (Hersbach, 2000; Wilks, 2006). Hersbach (2000) has shown that the *CRPS* can be decomposed into a reliability part and a

resolution/uncertainty part. However, Velázquez et al. (2010) mention that, because of the complexity of the *CRPS*, other ways to assess reliability are often used in parallel. Wilks (2006) also argues that single-number summaries of forecast performance can provide a convenient quick impression, but for a comprehensive evaluation of forecast quality the full joint distribution of forecasts and observations should be evaluated. It is considered that the *CRPS* is a suitable general score, but it does not provide thorough insight into reliability, sharpness, resolution and uncertainty of the ensemble flow forecasts. Therefore in addition to *CRPS* specific scores are used for these specific properties. In section 3.7.3 the CRPS is further explained.

2. Reliability diagrams, plotted together with a decomposition of the Brier score and a histogram of sample size (for sharpness). The Brier score will not be used here, because the *CRPS* will be used (see point 1 above). Reliability diagrams will be used to qualitatively evaluate reliability, sharpness and resolution. In section 3.7.4.3 the reliability diagram is further explained.
3. The Relative Operating Characteristic (ROC) curve, providing information about resolution of the ensemble forecast system. In section 3.7.6 the ROC curve is further explained.
4. Rank histograms to detect model bias and under- or overdispersion of the ensemble flow forecasting system. The rank histogram is further explained in section 3.7.4.2.

In addition to these scores the Relative Mean Absolute Error (*RMAE*) and the Relative Confidence Interval (*RCI*) will be used. This means that many evaluation scores will be used and when the results are examined it must be clear what each score measures. In section 3.7.3 to 3.7.7 the evaluation scores will be further explained, starting with the general evaluation scores (section 3.7.3), then the *RMAE*, rank histogram and reliability diagram to evaluate reliability (section 3.7.4), the reliability diagram to evaluate sharpness (section 3.7.5), the ROC to evaluate resolution (section 3.7.6) and the *RCI* to quantify the uncertainty that the ensemble flow forecasts contain (section 3.7.7).

3.7.3 General evaluation scores

The *CRPS* combines different properties of forecast quality. It is a frequently used score in meteorological sciences (Velázquez et al., 2010) and hydrological sciences (Bennett et al., 2014; Pappenberger et al., 2015; Velázquez et al., 2010). Pappenberger et al. (2015) argues that for most cases it is the recommended evaluation score. The *CRPS* is not limited to binary events, like several other scores to evaluate probabilistic forecasts (Hersbach, 2000). According to Hersbach (2000) the *CRPS* provides a broader view of performance. In section 3.7.3.1 it will be explained how the *CRPS* can be calculated and in section 3.7.3.2 an appropriate reference score is developed to calculate the skill of the ensemble flow forecasts.

3.7.3.1 Continuous Ranked Probability Score

In Figure 21 the concept of the *CRPS* is presented. With the *CRPS* the correspondence between the CDFs of the forecasts and observations is quantified (Pappenberger et al., 2015; Verkade et al., 2013). Instead of two options like in the Brier Score (event occurs or does not occur) the range of the predicted parameter is divided into more classes (Hersbach, 2000). With the *Continuous* Ranked Probability Score there is an infinite number of classes with zero width to evaluate all possible discrete events (Hersbach, 2000; Verkade et al., 2013). Hersbach (2000) argues that the *CRPS* can be considered as the integral of the Brier score over all possible probability threshold values. Advantages of the *CRPS* are that it is sensitive to the entire range of the outcome of interest and it does not require the introduction of predefined classes (Hersbach, 2000; Trinh et al., 2013). The score can be calculated with the following equation (Velázquez et al., 2010):

$$CRPS(F_t, X_r) = \int_{-\infty}^{\infty} \left(F_t(x) - H(x \geq X_r)\right)^2 dx \qquad [3.8]$$

$CRPS$ = Continuous Ranked Probability Score [unit of *x*]

$F_t(x)$ = Non-exceedance probability of event *x* in the forecasts (F_t is the CDF of the forecasts) [-]

$H(x \geq X_r)$ = Probability of the event according to the observations. Generally this is a Heavyside function, which equals 1 for ensemble members larger than the observed value and 0 for ensemble members lower than the observed values (Figure 21). [-]

x = Ensemble forecast member [unit of variable]

X_r = Reference observation or simulation [unit of variable]

A *CRPS* value of 0 indicates a perfect simulation (Hersbach, 2000; Velázquez et al., 2010; WMO, 2015; J. Ye et al., 2014), which can only be achieved in the case of a perfect deterministic forecast ($F_t(x) = H(x \geq X_r)$) (Hersbach, 2000). Equation 3.8 evaluates the correspondence between one observation-forecasts pair is evaluated. In practice the *CRPS* of all pairs is averaged (Hersbach, 2000; Velázquez et al., 2010; Verkade et al., 2013). The result is also denoted by *CRPS*. To bring the *CRPS* into practice the Matlab-script written by Shrestha (2014) is used. To determine the CDF of the forecasts an empirical CDF is developed.

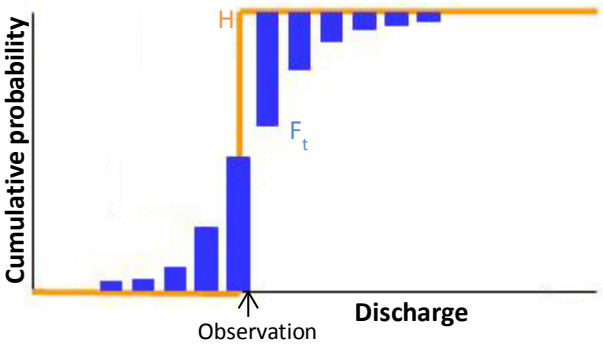


Figure 21: Concept of the Ranked Probability Score (Eumetcal, n.d.; Wilks, 2006)

A property of the *CRPS* is that in practice the value will not go to zero, but it becomes the average value of the studied outcome (with the same unit as the investigated parameter) (J. Ye et al., 2014). This means that the *CRPS* cannot directly be compared amongst different areas, seasons or flow categories (J. Ye et al., 2014). Comparison between different lead times is possible, because there are no different average discharge values expected with increasing lead time. The *CRPS* can be normalized by dividing it by the standard deviation of the investigated parameter observations (Trinh et al., 2013; J. Ye et al., 2014) or normalized against a reference *CRPS* (J. Ye et al., 2014). It turned out that after normalizing the *CRPS* with the standard deviation, the value is still correlated with discharge magnitude (Trinh et al., 2013), so the normalized *CRPS* can still not be compared among different forecast categories. Normalizing against a reference flow has the advantage that it does not only eliminate the effect of the magnitude of the investigated parameter, but it also compares the forecasts with a relevant reference forecast (J. Ye et al., 2014). In the next section this is further elaborated.

3.7.3.2 Continuous Ranked Probability Skill Score

To evaluate the skill of the ensemble flow forecasting system and compare them over different situations the score of the system relative to a reference system should be used (Demargne et al., 2010; Verkade et al., 2013). The Continuous Ranked Probability Skill Score (*CRPSS*) is defined as

follows (Bennett et al., 2014; Demargne et al., 2010; Renner et al., 2009; Velázquez et al., 2010; Verkade et al., 2013; J. Ye et al., 2014):

$$CRPSS = 1 - \frac{CRPS_{forecasts}}{CRPS_{reference}} \qquad [3.9]$$

The reference forecast set can be seen as alternative forecasts without using meteorological forecasts. A system with perfect skill gives a score of 1 and negative values indicate that the forecast system has a worse *CRPS* score than the reference (Demargne et al., 2010; J. Ye et al., 2014). The most useful reference forecast for the evaluation of flow forecasts is easy to determine and difficult to beat (Bennett et al., 2013; Pappenberger et al., 2015). It is common practice to apply hydrological persistency or hydrological climatology as a reference (Bennett et al., 2014; Pappenberger et al., 2015; Renner et al., 2009; J. Ye et al., 2014), but Pappenberger et al. (2015) conclude that this can result in an overestimation of forecast skill because other reference forecasts might be more difficult to beat. Bennett et al. (2013) show that hydrological climatology is easy to beat because discharge is strongly autocorrelated at short lead times and it can be biased because it is based on a different period. Hydrological persistency is easy to beat because persistence forecasts generally perform very poorly at longer lead times (> 1 day) (Bennett et al., 2013). Pappenberger et al. (2015) and Bennett et al. (2013) both advise to use an ensemble of past observation sequences of precipitation and temperature at the same calendar day over the past 20 years (so 20 ensemble members) to run the hydrological model and use this as an ensemble of reference discharges.

It will be investigated whether reference forecasts based on past meteorological observations (following the method of Pappenberger et al. (2015)), hydrological persistency or hydrological climatology are better to use, by comparing the *CRPS*. This is the same approach as Bennett et al. (2013) (in addition they used *NS* and a bias score) and Pappenberger et al. (2015) applied. With hydrological persistency the most recent discharge observation is used (at the forecast day – 1) as forecast for all lead times. With hydrological climatology the average observed discharge over the last 20 years on the same calendar day is used after applying a moving average window of 31 days, following Bennett et al. (2013). Hydrological persistency flows and hydrological climatology are deterministic variables. Hersbach (2000) shows that when two deterministic forecasts are compared the *CRPS* simplifies to the Mean Absolute Error (*MAE*).

In Figure 22 the *CRPS* values of the reference forecast sets are presented. One reference forecast set will be selected for all flow categories, so one general *CRPS* is calculated. It can be seen that the forecasts based on an ensemble of historical observations of precipitation and temperature has the best score, so these are most appropriate as reference flow. Even for a lead time of 0 days this gives a better *CRPS* than hydrological persistency. The discharge forecasts at a lead time of 0 days are totally generated by the initial storages at the beginning of this time step, so this shows that the updating procedure results in simulations closer to the observations than the observation of 1 day earlier. The performance of hydrological persistency also depends on how fast the hydrological circumstances in the catchment change.

The *CRPS* and *CRPSS* will be used to perform a general evaluation for each lead time, for the low, medium and high flow categories (defined in section 3.8) and different hydrological circumstances (defined in section 3.9).

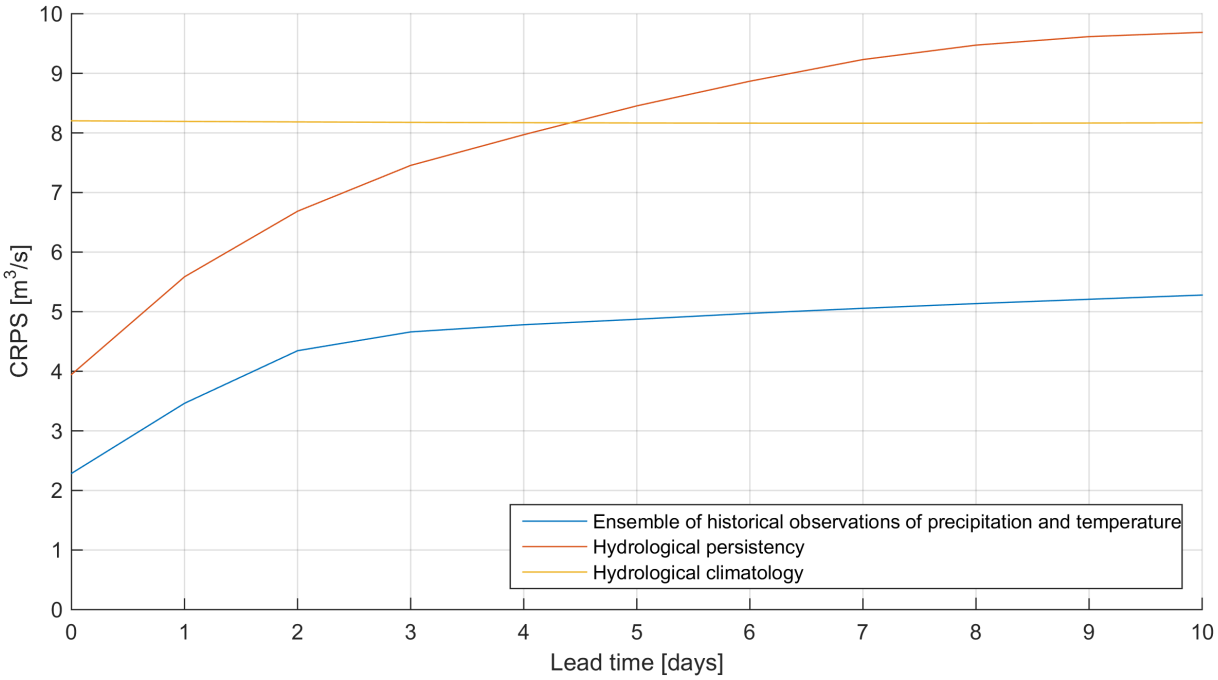


Figure 22: CRPS of three different reference forecast sets, evaluation period 2008-2013

3.7.4 Reliability

Reliability refers to the statistical consistency between simulations and observations (Candille & Talagrand, 2005; Velázquez et al., 2010) and whether uncertainty is correctly represented in the forecasts (Bennett et al., 2014). WMO (2015) defines reliability as the agreement between the forecasted probability of an event and the mean observed frequency of that event. The reliability scores differ in how they measure reliability, some are general over all flow categories (unconditional bias: *RMAE*, rank histogram and percentile-based evaluation) and others apply to pre-defined forecast situations (conditional bias: reliability diagrams). Percentile-based evaluation measures the same as the rank histogram (Olsson & Lindström, 2008) and the rank histogram provides a more clear interpretation, so the rank histogram will be used. Below the *RMAE* (section 3.7.4.1), rank histogram (section 3.7.4.2) and reliability diagram (section 3.7.4.3) are explained.

3.7.4.1 Relative Mean Absolute Error

The *RMAE* measures the average absolute difference between the mean of the ensemble forecasts and the reference flows, relative to the magnitude of the reference flows. The ensemble mean is commonly used for single-valued representation of the ensemble forecast (Demargne et al., 2010).

$$RMAE = \frac{\sum_{t=1}^{t=N} |X_r(t) - \overline{x(t)}|}{\sum_{t=1}^{t=N} X_r(t)} \qquad [3.10]$$

$RMAE$ = Relative Mean Absolute Error [-]
X_r = Reference observation or simulation [unit of variable]
$\bar{x}$ = Mean of the ensemble forecasts [unit of variable *x*]
t = Time step [day]
N = Total number of time steps [days]

The *RMAE* provides a measure of unconditional bias in the mean of the ensemble forecast (Verkade et al., 2013). The *RMAE* will be used as a basic score to evaluate the effect of pre- and post-processing. It will not be used to evaluate the flow forecasts, because of the principle that ensemble flow forecasts should be used as a set.

3.7.4.2 Rank histogram

The rank histogram enables to diagnose average errors in the mean and spread of the ensemble forecasts (Hamill, 2001). According to Wilks (2006) this is the most common approach to evaluate consistency of the ensemble forecasts. The consistency condition states that the true state is just one more member of the ensemble and it should be statistically indistinguishable from the forecast ensemble (Wilks, 2006). To construct the rank histogram the observed value is added to the simulated ensemble set, to obtain a new set of n+1 members (Velázquez et al., 2010). The rank associated with the observed value is determined and the rank histogram is constructed based on the resulting ranks (Velázquez et al., 2010). In case that (multiple) ensemble members and the observation have the same value, like 0 mm precipitation, the associated rank is determined randomly (like Hamill and Colucci (1998) also did).

In an ensemble prediction system with perfect spread each member is equally likely, so all ranks of the observation are equally likely and the rank histogram will be uniform (Hamill, 2001; Hersbach, 2000; Wilks, 2006; WMO, 2015; Zalachori et al., 2012). According to Hamill (2001), Velázquez et al. (2010) and Wilks (2006) an asymmetrical histogram is an indication of bias in the mean of the forecasts. U-shaped histograms indicate that the simulations are under-dispersed (overconfident), since this means that many of the observations are close to the outer range of the ensemble prediction (Hamill, 2001; Velázquez et al., 2010; Wilks, 2006). However, Hamill (2001) suggests that this can also be caused by conditional bias and/or imperfect reference flows. If the rank histogram is clock shaped the predictive distribution is over-dispersed (underconfident), since in that case many of the observations are close to the median of the ensemble forecast (Hamill, 2001; Velázquez et al., 2010; Wilks, 2006). The rank histogram is a visual way to assess the reliability of an ensemble prediction. It can be used to see if an ensemble forecast is calibrated well (Renner et al., 2009). Candille and Talagrand (2005) use a numerical indicator δ to reflect the flatness of the rank histogram. However, with the flatness coefficient of Candille and Talagrand (2005) it is impossible to compare different forecast sets, because the value is proportional to the length of the time series (Velázquez et al., 2010). This is a problem when the flatnesses of the different hydrological flow categories are compared (defined in section 3.8). Therefore the MAE will be used as flatness indicator, which measures the deviation of the relative frequencies per rank with the relative frequency in a uniform distribution, described in equation 3.11:

$$\delta = \frac{1}{n+1} \sum_{z=1}^{z=n+1} |f(z) - y| \qquad [3.11]$$

δ = Flatness coefficient [-]
$f(z)$ = Predicted relative frequency in rank *z* [-]
$y = \frac{1}{n+1}$ = Theoretical relative frequency [-]
n = number of ensemble members [-]

In a perfectly reliable forecast system the number of elements in each rank is equal to the theoretical number of elements (Velázquez et al., 2010). This gives an optimum value of δ equal to 0. It should be realized that the rank histogram and the flatness coefficient contain a random element if multiple ensemble members and the observation have the same value, like 0 mm precipitation. In addition, a uniform rank histogram is no guarantee that the ensemble forecasts are reliable at each point in the evaluation period (Hamill, 2001).

In this study the rank histogram will be used to examine whether the ensemble flow forecasting system has been calibrated well on bias and dispersion for all lead times and for different flow categories to see whether the dispersion of ensemble members is different for high and for low flows and whether conditional biases are present. Hamill (2001) explains that a different shape of the rank histogram over different flow categories indicates conditional bias in the ensemble forecasts. After all, the rank histogram is used to evaluate the effect of pre- and post-processing.

3.7.4.3 Reliability diagram

The reliability diagram is a common way to summarize and evaluate the reliability of probabilistic forecast systems (Bröcker & Smith, 2007), for example used by Olsson and Lindström (2008), Velázquez et al. (2010), Bennett et al. (2014) and Demirel et al. (2013a). This diagram consists of a plot of the observed relative frequency against the predicted probability for a certain event (Bröcker & 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). Wilks (2006) and WMO (2015) explain that the curve below the diagonal line indicates overforecasting, which means that forecasted probabilities are too high, and the curve above the diagonal indicates underforecasting (see Figure 23). The deviation from the diagonal line gives the conditional bias, which can be linked to resolution (see section 3.7.6) (Wilks, 2006; WMO, 2015).

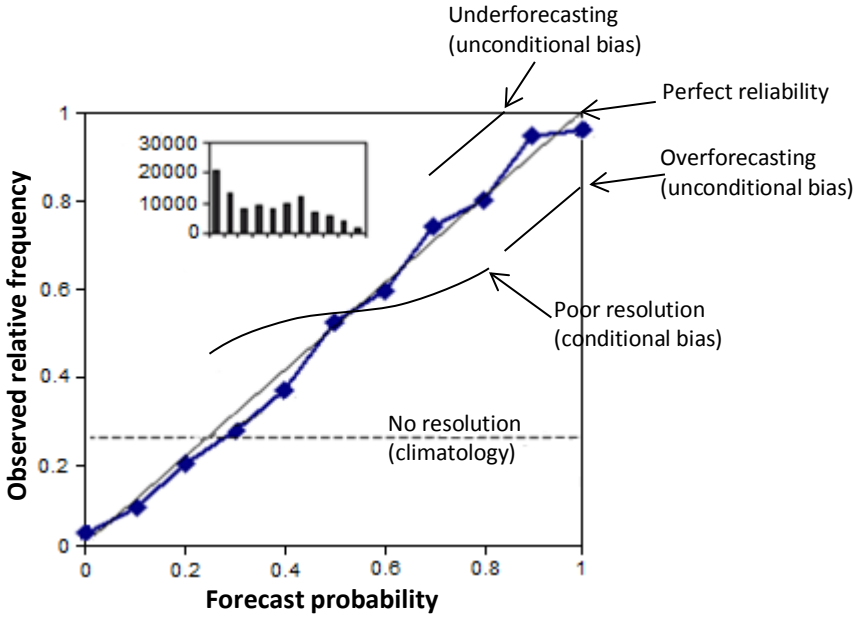


Figure 23: Interpretation of reliability diagram (Wilks, 2006; WMO, 2015)

The reliability diagram is implemented for the low and high flow categories that are defined in section 3.8. The forecast probability bins that are used are 0%-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100%, which was also used by Demirel et al. (2013a) and Bennett et al. (2014). Smaller bins are possible but this increases uncertainty of the points in the reliability diagram, because the bins will contain less time steps. Bennett et al. (2014) did not plot a point in the reliability diagram if the bin contains less than 100 forecasts, but this has not been applied here. It should however be realized that for points in the reliability diagrams for which a few forecasts were available, conclusions should be drawn cautiously. The bins to which this applies are mainly the medium probability bins (see the interpretation of the histogram in a reliability diagram in section 3.7.5).

The implementation of the reliability diagram in Matlab is based on an existing Matlab-script, developed by Skynet (2009). For low and high flow forecasts it is determined how many forecasts fall in a certain forecast probability bin (forecast probability) and how often this corresponds to the observations (observed relative frequency). Following Bröcker and Smith (2007) the observed frequencies are plotted against the average of forecast probabilities per bin instead of the bin average. Plotting against bin averages (so 10%, 30%, etc.) can cause substantial deviations from the diagonal (Bröcker & Smith, 2007).

In addition to reliability more properties of an accurate forecasting system can be derived from the reliability diagram. The sample size in each probability bin is often included as a histogram, and this shows the sharpness of the forecasts (Ranjan, 2009; Renner et al., 2009; WMO, 2015) (for more information see section 3.7.5). In addition, the flatter the reliability diagram the less resolution the forecasts have (WMO, 2015) (resolution is explained in section 3.7.6). According to Renner et al. (2009) the reliability diagram is a useful tool in the evaluation process, because it displays most forecast properties (when it includes the histogram) and it can be prepared for any threshold and any lead time. The reliability diagram will be used to qualitatively evaluate reliability, sharpness and resolution of the low flow forecasts and high flow forecasts for all lead times.

3.7.5 Sharpness

Ranjan (2009) and WMO (2015) define sharpness as the tendency to forecast probabilities of occurrence of an event near 0 or 1, as opposed to values clustered around the mean (climatological) probability. If an ensemble forecasting system always forecasts an event probability close to climatological probability, instead of close to 0 or close to 1, this forecasting system is not useful, although it might be well calibrated (Ranjan, 2009; Wilks, 2006). This is independent of the corresponding observations of forecasts (Wilks, 2006). The histogram accompanied with a reliability diagram, showing the sample size in each probability bin, indicates the sharpness of the forecast (Ranjan, 2009; Renner et al., 2009; WMO, 2015). For a sharp forecast, this histogram should be U-shaped (Ranjan, 2009). This histogram will be used to qualitatively assess the sharpness of the low flow and high flow forecasts. This approach is also used by Ranjan (2009).

3.7.6 Resolution

Resolution is the ability of the forecast model to correctly forecast the occurrence or non-occurrence of events (Demirel et al., 2013a; Martina et al., 2006). According to WMO (2015) resolution is the ability of the forecasting system to resolve the set of events into subsets with different frequency distributions. When event A was forecasted the distribution of outcomes should be different from the distribution of outcomes than when event B was forecasted (WMO, 2015). A flat curve in the reliability diagram indicates that the forecast system has no resolution on the considered event, since for all forecast probabilities bins the same observed relative frequency (climatology) is found (Wilks, 2006; WMO, 2015). So the reliability diagram will be used to qualitatively assess and compare the resolution of the forecast system for low flow and high flow forecasts. Contingency tables and ROC curves can be used to evaluate whether the forecast model correctly forecasts the occurrence and non-occurrence of events. It has been chosen to apply ROC as a score for this, because ROC is a score over more than one probability threshold like single contingency tables. In addition, WMO (2015) states that reliability diagrams and ROC are good partners, because the ROC is conditioned on the reference flows and the reliability diagram is conditioned on the forecasts.

3.7.6.1 Relative Operating Characteristic

The ROC is based on two-by-two contingency tables. Contingency tables are the basis to measure resolution, since they indicate the ability of the system to correctly discriminate between two alternative outcomes (Demargne et al., 2010; Velázquez et al., 2010). The combinations between observed, not observed, forecasted and not forecasted events are defined in Table 12. A forecasted event applies when a certain number of ensembles exceeds the flow threshold (probability threshold). With a low probability threshold more hits will be observed, but there will also be more false alarms (Komma et al., 2007). With the ROC this trade-off is evaluated (Komma et al., 2007; Mason & Graham, 2002). Although the ROC curve does indicate potential usefulness of the forecasting system (WMO, 2015), it cannot be used as a general evaluation score because specific numerical values of forecasts are not incorporated (Wilks, 2006).

Table 12: Contingency table

	Observed	**Not observed**
Forecasted	Hit	False alarm
Not forecasted	Miss	Correct negative

The skill of a forecasting system to predict events can be represented by the hit rate and the false alarm rate (Martina et al., 2006; Velázquez et al., 2010). The hit rate and false alarm rate can be calculated as follows (WMO, 2015):

$$Hit\ rate = \frac{hits}{hits + misses} \qquad [3.12]$$

$$False\ alarm\ rate = \frac{false\ alarms}{correct\ negatives + false\ alarms} \qquad [3.13]$$

For the hit rate a score of 1 is perfect and for the false alarm rate a score of 0 is perfect (WMO, 2015). To establish the ROC score 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 probability thresholds are usually defined according to the number of ensemble members that forecast the event (from 1 to the number of ensemble members) (Atger, 2001). Every pair in the ROC curve indicates the performance of a deterministic forecast that would be based on the fact that at least that number of ensemble members forecasts the considered event (Atger, 2001). 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 forecast 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. A limitation of using the *AUC* as performance measure is that for relatively good forecast systems the *AUC*s tend to be quite similar (Marzban, 2004).

The ROC score has been brought into practice by calculating the hit rate and false alarm rate for requested exceedance probabilities of 1 to 51 of the ensemble members that exceed the predefined low or high flow threshold. Also the two endpoints (0,0) and (1,1) are included (Fawcett, 2006; Wilks, 2006). (0,0) represents the strategy of never issuing a positive classification and (1,1) represents the

strategy of unconditionally issuing positive classifications (exceedance of 0 ensemble members is required) (Fawcett, 2006). After all, following Fawcett (2006) the *AUC* is calculated with a trapezoidal approximation. This score will be calculated for different lead times for the low and high flow thresholds that are defined in section 3.8.

3.7.7 Relative uncertainty of flow predictions

The *RCI* will be used to describe the uncertainty that the ensemble flow forecasts contain. The *RCI* is defined as follows (Demirel et al., 2013a):

$$RCI = \frac{1}{N} * \sum_{t=1}^{t=N} \frac{Q_{95}(t) - Q_5(t)}{Q_{50}(t)} * 100 \qquad [3.14]$$

RCI =Relative confidence interval [%]
Q_5 = 95% percentile of ordered ensemble flows [m^3/s]
Q_{95} = 5% percentile of ordered ensemble flows [m^3/s]
Q_{50} = 50% percentile of ordered ensemble flows [m^3/s]
t = Time step [day]
N = Total number of time steps [days]

This score quantifies how wide the ensemble predictions are and thus how uncertain the flow predictions are. The *RCI* will be used to quantify uncertainty in the ensemble predictions for different lead times and for different flow circumstances. This is not a true evaluation score, since ensemble members are not evaluated against reference flows and there is no good or bad score.

3.8 Evaluation of ensemble flow forecasting purposes

To fulfil the research objective the ensemble flow forecasting system will be evaluated for different purposes. This includes different lead times, from 1 day to 10 days ahead. At second the model will be evaluated for different hydrological flow categories to see how the model performs during low discharges, medium discharges and high discharges. To define the flow thresholds certain percentiles of observed discharges during the evaluation period (1-11-2007 until 31-10-2013) are used, to guarantee a sufficient number of data points in the evaluation. As low flow threshold Q_{75} (exceedance probability of 75%) is chosen, analogously to Demirel et al. (2013b). This guarantees a sufficient number of events in the evaluation and discharges below this threshold still affect river functions (Demirel et al., 2013b). Similarly to the definition of the low flow threshold, for high discharges the threshold Q_{25} is used. It should be realized that flows below Q_{75} and above Q_{25} are not necessarily very extreme discharges, but a sufficient number of data points is needed in the evaluation. All flows in between Q_{75} and Q_{25} are classified as medium flows. The flow categories are summarized in Table 13. A certain flow category applies when the observed discharge on a day falls in this category. The results of evaluation on different purposes are presented in section 4.3.

Table 13: Definition of hydrological flow categories

Flow category	Thresholds	Discharges of thresholds
Low flows	$Q_0 \leq Q_{75}$	$Q_0 \leq 2.76\ m^3/s$
Medium flows	$Q_{75} < Q_0 \leq Q_{25}$	$2.76\ m^3/s < Q_0 \leq 10.35\ m^3/s$
High flows	$Q_{25} < Q_0$	$10.35\ m^3/s < Q_0$

3.9 Evaluation of hydrological circumstances

After evaluation of the different hydrological flow categories the next step is to see whether there are differences in the performance of high flow forecasts (section 3.9.1) and low flow forecasts

(3.9.2) with different driving processes. This will not be done for medium flows, because these are non-extreme flows that are probably not clearly caused by a certain driving process. The observed discharges will be classified, so that independent of the ensemble forecasts or lead time a flow falls in a certain driving process category.

3.9.1 High flow producing processes

High flows can be caused by different driving processes. A range of variables, like rainfall regime, snowmelt and state of the catchment, can give rise to floods (Merz & Blöschl, 2003). Merz and Blöschl (2003) distinguish long-rain floods, short-rain floods, flash floods, rain-on-snow floods and snowmelt floods. Vormoor et al. (2015) used fluxes from the HBV model to classify events into rainfall (2/3 rainfall generated), snowmelt (2/3 snowmelt generated) and rainfall + snowmelt. Merz and Blöschl (2003) used time of the year, rainfall in the days before the event, snowmelt rates from a HBV model, snow water equivalents from the HBV model and soil moisture conditions from the HBV model to manually classify flood events into one of the mentioned high flow producing processes. Merz and Blöschl (2003) mention that the relationships between flood indicators and process types are very complex, so the development of quantitative rules is not straightforward. In this project fewer processes will be distinguished and characterization rules provide more insight into the actual characterization, so it has been chosen to establish quantitative characterization rules.

Flash floods cannot be distinguished in this study, because only daily observations and forecasts are available. Therefore short-rain floods will also contain flash floods. Rain-on-snow floods and snowmelt floods are considered as one category, because they cannot reliably be distinguished. In Table 14 the high flow producing processes are further explained.

In this study all high flows where snowmelt is involved will be classified as snowmelt generated floods, because when snow and rainfall occur together the snowpack plays an important role in the runoff process. All events in which no snowmelt is involved are classified as rainfall generated flows. If the precipitation at the day before the flow event is above the precipitation amount of 10 mm the event is classified as a short-rain flood. From the time series of observed precipitation, observed discharge and HBV model storages it has been found that with low initial storages a precipitation amount of 10 mm/day at one day before the flow event is able to cause a discharge event above the high flow threshold. 10 mm/day is not a very extreme amount of precipitation, but a flow threshold that is exceeded by 25% of the discharges is also not very extreme. All remaining events are classified as long-rain floods. The classification rules are related to fluxes and storages at one day before the event, because in the HBV model it takes one day before the rainfall and snowmelt fluxes end up in the fast runoff and slow runoff reservoir and can form runoff. The ensemble flow forecasts are evaluated for these high flow producing processes in section 4.4. In section 5.3 the classification procedure is discussed, inter alia the subjectivity in this classification and the problem of classifying high discharges that are generated by a combination of processes.

In Figure 24 the distribution of high flow producing processes over the year is presented. This seems a reliable distribution of the processes, with snowmelt flows in winter, short-rain floods mainly in summer and long-rain floods during the whole year.

Table 14: Characterization of the high flow producing processes

Process	Characterization	Characterization rules
Snowmelt floods	Snowmelt floods occur when there is a snowpack and a warm weather spell causes melting (Merz & Blöschl, 2003). With rain-on-snow floods the rainfall enhances snowmelt from an existing snow cover and due to antecedent snowmelt the catchment can already be saturated and the rainfall and melted snow water will directly form runoff (Merz & Blöschl, 2003). These floods are considered as one category, because both are related to snowmelt processes and in the HBV model snowmelt and precipitation falling as rainfall are strongly related (both based on temperature). It would not be possible to reliably distinguish between these processes. Therefore all high flow events where snow is involved are characterized as snowmelt floods.	• Snowpack present at day-1: The event is related to snowmelt.
Short-rain floods	Merz and Blöschl (2003) explain that rainfall of short duration and high intensity can saturate parts of the catchment. Flood runoff results from a combination of runoff from saturated areas, overland flow due to high rainfall intensities and fast subsurface flow (Merz & Blöschl, 2003). Wet antecedent conditions enhance the magnitude of this type of event (Merz & Blöschl, 2003). Flash floods also fall in this category.	• No snowpack present at day-1: The event is not related to snow. • Rainfall at day-1 above 10 mm: The event is caused by high rainfall at the day before.
Long-rain floods	Rainfall over several days or weeks can saturate the catchment. When the storage capacity of the catchment is exceeded any additional rain generates a flood event (Merz & Blöschl, 2003). This category applies when a flow event is not directly generated by snowmelt or high precipitation.	• No snowpack present at day-1: The event is not related to snow. • Rainfall at day-1 below 10 mm: The event is not caused by high rainfall at the day before.

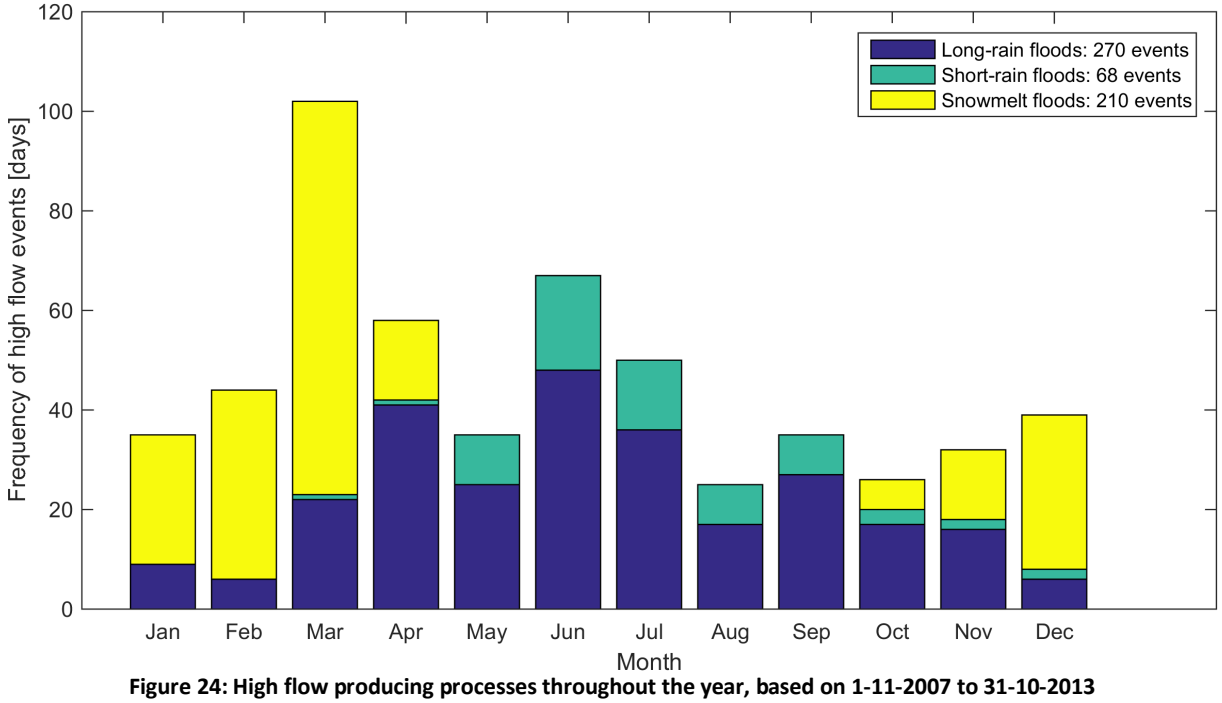


Figure 24: High flow producing processes throughout the year, based on 1-11-2007 to 31-10-2013

3.9.2 Low flow producing processes

The processes that cause low flows are snow accumulation and the combination of low rainfall and high evapotranspiration (precipitation deficit) during a period before the event. These processes are characterized in Table 15.

As for high flow producing processes also for low flow producing processes characterization rules are established to characterize all low flow events in the period of interest. If in the period before the day of the event the precipitation is accumulated as snow this is the reason that the fast runoff and slow runoff reservoir have a low actual storage. In the HBV model precipitation is stored in the snowpack then. In the case that no snowpack is present it is assumed that the low flow must be caused by low rainfall and high evapotranspiration. The evaluation results of the ensemble flow forecasts for these low flow producing processes are presented in section 4.5. In section 5.3 the classification of low flow producing processes is discussed.

In Figure 25 the distribution of low flow producing processes over the year is presented. Especially from December to February the low flows are caused by snow accumulation. In the rest of the year the low flows are mainly caused by low rainfall, with most low flow events in the months from August to October.

Table 15: Characterization of the low flow producing processes

Process	Characterization	Characterization rules
Snow accumulation	If temperature is low and precipitation falls as snow, this water is accumulated in the catchment and does not form runoff at that moment.	• Snowpack (HBV) at day-1: The event is related to snow accumulation.
Low rainfall/ high evapotrans-piration	If there falls no or very low rainfall and there is a high evapotranspiration during a long period the conditions in the catchment will become dry and discharge will be low.	• No snowpack (HBV) at day-1: The event is not related to snow accumulation.

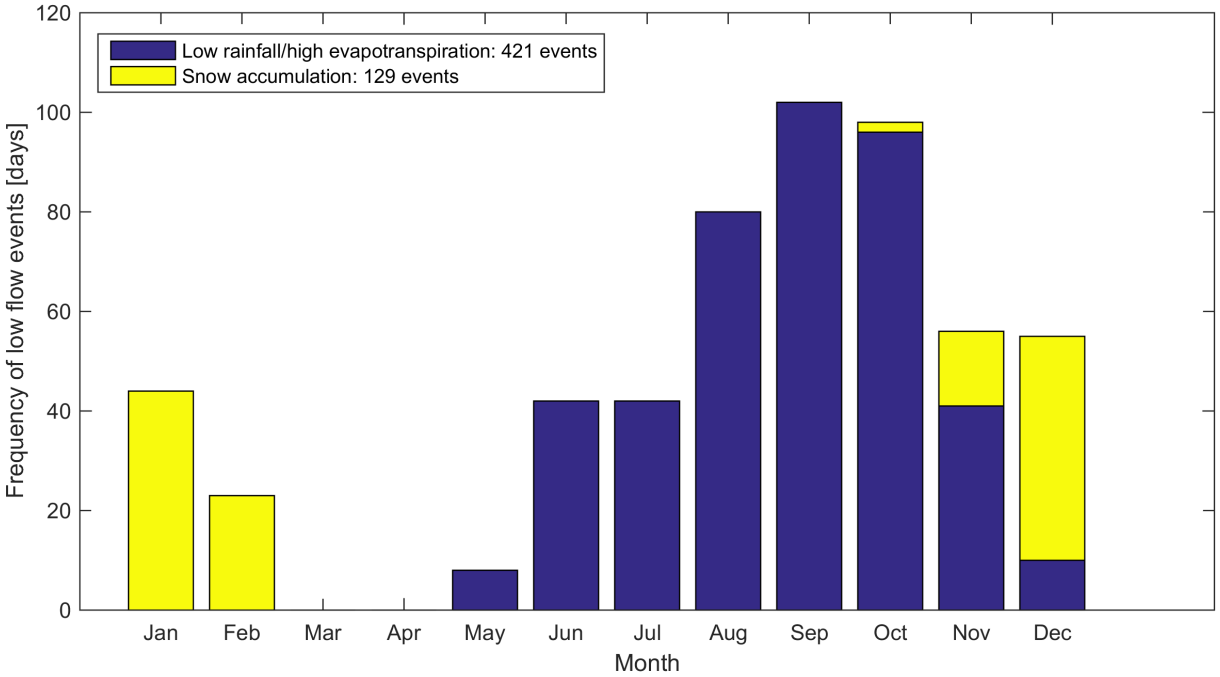


Figure 25: Low flow producing processes throughout the year, based on 1-11-2007 to 31-10-2013

4. Results

This chapter presents and explains the results. In section 4.1 the calibration and validation results are presented and in section 4.2 the pre- and post-processing strategies are evaluated. In section 4.3 the ensemble flow forecasts are evaluated for different purposes. In section 4.4 and section 4.5 the flow forecasts are evaluated for respectively high flow producing processes and low flow producing processes.

4.1 Calibration and validation results deterministic hydrological model

This section presents the results of the sensitivity analysis (section 4.1.1), the calibration and validation (section 4.1.2), an estimation of parameter uncertainty (section 4.1.3) and the model performance per year of the evaluation period (section 4.1.4). In section 4.1.5 the effect of different updating procedures is evaluated and in section 4.1.6 the hydrological model performance is examined for different purposes.

4.1.1 Results sensitivity analysis

The results of the sensitivity analysis will be used to see which parameters do not have a large influence on the model output and thus to which parameters a fixed value can be assigned in the second calibration round and in the uncertainty analysis. In Figure 26 the influence of parameters is presented. It must be noted that the picture of Figure 26 is slightly different every time the sensitivity analysis is run, but the relative position of the parameters is more or less the same. It follows that especially *FC*, α, K_f, *CFMAX* and *PERC* have a large influence on the model output. Also *TT*, *LP* and β have a considerable influence. Since with the method of Morris there is a risk that parameters are classified as non-influential while they are important (Song et al., 2015), also *TT*, *LP* and β are included in the calibration procedure.

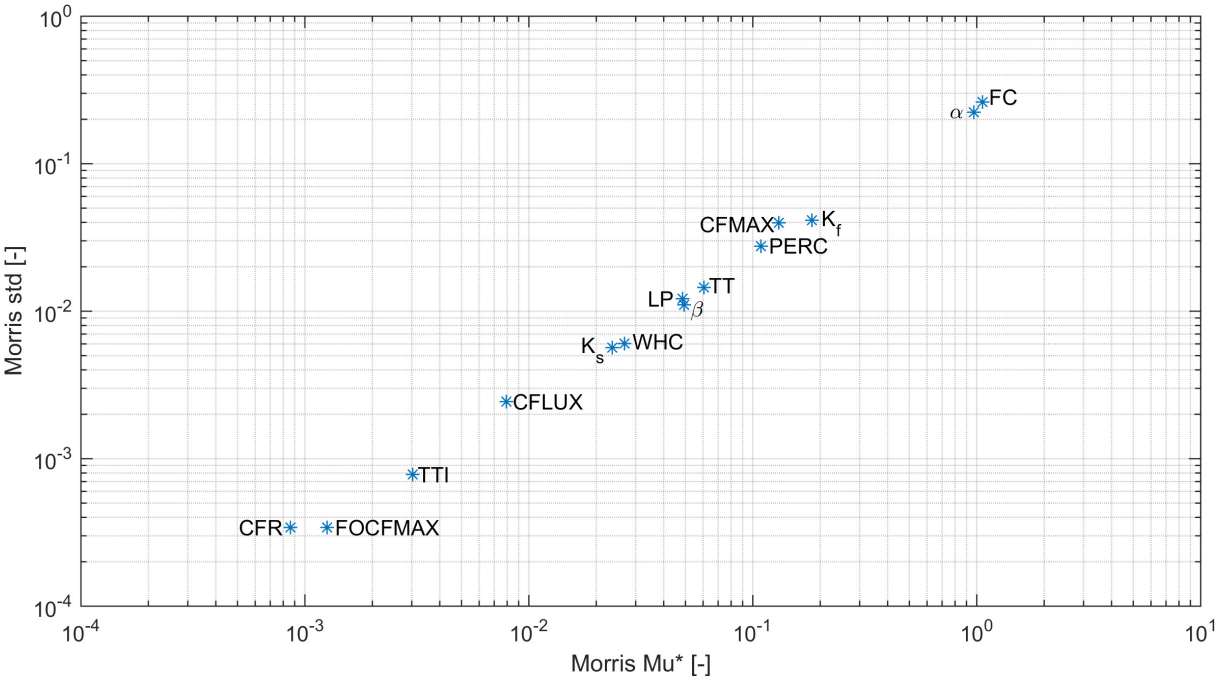


Figure 26: Result of the sensitivity analysis with the method of Morris. Based on this it has been chosen to calibrate on *FC*, α, K_f, *CFMAX*, *PERC*, *TT*, *LP* and β in the second calibration round.

4.1.2 Optimum parameter set

In Table 16 the resulting values of the objective function of the calibration and validation are presented, with uncorrected and corrected input data for elevation (section 3.1). By correction for elevation differences over the catchment, precipitation is corrected for underestimation and temperature is corrected for overestimation of the catchment averages. Without correction, these systematic errors would be (partly) captured in the calibration. The calibration and validation results have improved after elevation correction, so these systematic errors were not completely captured in the calibrated parameters. In addition, and this is more important in this study, the hydrological model will also be applied with the meteorological forecasts from ECMWF. So it must be prevented that in one of the input data sets a systematic bias is present, because the optimum model parameters are dependent on this bias. The resulting optimum parameter values are presented in Table 17. Over the calibration period larger values of the objective function are obtained, because the parameters are optimized for this period. The positive *RVE* values in the validation indicate that over the whole validation period the volume of the observed discharge is overestimated by the model. The values of the objective function of the validation period are rather low compared to the results of Van den Tillaart (2010), comparable with the results of Akhtar et al. (2009) and Knoben (2013) and higher than the results of Maat (2015).

DEGL is a global optimization procedure, which means that it is supposed to find the global optimum parameter set. However, by repeating the calibration procedure with 14 parameters with the same settings but different starting points of parameter values, this resulted in a slightly different parameter set. The same objective function is obtained over the calibration period, which is the reason that different parameter sets could result from the calibration. Due to the high number of parameters in the HBV model (14 parameters) overparameterisation may be present, which means that different parameter sets can give equally good output performances (Booij, 2005). This is also explained by Song et al. (2015) who state that in practice a large number of parameters (from tens to hundreds) leads to the curse of dimensionality where parameter estimation becomes a high-dimensional and mostly nonlinear problem. According to Song et al. (2015) it is often not feasible or necessary to include all model parameters in the calibration process. Therefore before the calibration, insensitive model parameters should be locked at a fixed value to facilitate more efficient calibration (Song et al., 2015). However, assigning a fixed value (from the first calibration round) to the insensitive parameters from the sensitivity analysis (see section 4.1.1) and applying the same calibration procedure to the 8 most sensitive parameters gives exactly the same result as their initial calibration with 14 parameters (with different starting points). You would expect that two calibrations with 14 parameters give a different optimum parameter set, but two calibrations with the 8 most sensitive parameters should give one optimum parameter set. Apparently the fixed values of the insensitive parameters already define the global optimum that will be found by DEGL. This indicates that the output is also sensitive to one or more of the parameters that were classified as insensitive.

Although it is undesirable to find a different parameter set each time the calibration is run, in this case it is accepted because the differences between the parameters sets from the two calibrations are quite small for most parameters. Only for the parameters *FOCFMAX*, *CFR* and *WHC* a very large difference was found (>10%). This can be explained by the fact that from the sensitivity analysis it appeared that these are insensitive parameters, so different values of these parameters will result in almost the same output. The difference of *CFMAX* between the two found optimum parameter sets

was about 10%, the difference between all remaining parameters was below 5% and for *FC*, *β*, *LP* and *TTI* even almost 0%. Also the *NS* and *RVE* over the validation period are almost the same (*NS* = 0.7657 and 0.7658 and *RVE* = 6.7164% and 6.7098%). During the validation period the maximum relative difference in simulated discharges between both parameter sets is 8.9% (absolute difference 0.0230 m^3/s) and the maximum absolute difference is 1.79 m^3/s (with relative difference 0.26%). Both parameter sets result in almost the same simulated discharges, so one of the parameter sets resulting from calibration with 14 parameters is accepted as the optimum parameter set.

Table 16: Calibration and validation results

Calibration run	Calibration (1-11-1971 to 31-10-2000)			Validation (1-11-2000 to 31-10-2013)		
	Y	***NS***	***RVE***	***Y***	***NS***	***RVE***
0. Calibration with uncorrected input data	0.78	0.78	0%	0.69	0.74	6.5%
1. Calibration with corrected input data (see section 3.1)	0.81	0.81	0%	0.72	0.77	6.7%

Table 17: HBV model parameter values from the calibration with corrected input data

Parameter	Parameter value	Unit
FC	122	mm
β	2.65	-
LP	1.00	-
α	0.267	-
K_f	0.232	d^{-1}
K_s	0.0782	d^{-1}
PERC	2.18	mm d^{-1}
CFLUX	0.875	mm d^{-1}
TT	0.129	^oC
TTI	7.00	^oC
CFMAX	1.75	mm $^oC^{-1}$ d^{-1}
FOCFMAX	0.879	-
CFR	0.222	-
WHC	$1.49*10^{-14}$	mm mm^{-1}

4.1.3 Parameter uncertainty

The highest objective function over the calibration period that followed from the Monte Carlo simulations is *Y* = 0.68, which confirms that DEGL is a smarter way to approach the global optimum than a simple Monte Carlo analysis. Only 0.82% of the parameter sets (50000 model runs) is assigned behavioural (*Y* > 0.5). This is a very small percentage compared to other studies. In the study of Jin et al. (2010) ~18% of the parameter sets is behavioural and in the study of Demirel et al. (2013a) ~9% of the parameter sets is behavioural. This percentage is dependent on several factors, like the threshold value for behavioural parameter sets, the parameter ranges and the used model and application.

The results of the GLUE analysis are presented in appendix 4. Sharp histogram distributions indicate well identifiable parameters, while flat distributions indicate a larger parameter uncertainty (Jin et al., 2010). For the parameters *FC*, *β*, *LP* and *α* the peak of behavioural parameter sets is well identifiable (Figure 52 in appendix 4) and except of *β* the distributions focus around the parameter values that were found by DEGL. These parameters have a relatively low parameter uncertainty. For the parameters K_f, *PERC*, *TT* and *CFMAX* the distributions are flatter, which indicates a larger parameter uncertainty of these parameters. This is also visible in the Monte Carlo simulations

scatterplots (Figure 51 in appendix 4), in which over a wide parameter range high objective functions are found for these parameters. The calibrated value of 1 for *LP* and the tendency of behavioural parameter values to 1 are treated in the discussion (section 5.1.2).

4.1.4 Hydrological model performance per year

In Table 18 the objective functions per hydrological year in the evaluation period are presented. In 2007 and 2008 the agreement between observed discharge and simulated discharge based on observed precipitation and temperature is very poor, especially in 2007. In the discussion (section 5.1.1) possible explanations are mentioned for this. The *NS* below 0 means that it would be even better to use the average discharge in this year than to use the output of the hydrological model. In Figure 27 observed and simulated discharge over 2007 and 2008 are shown. It is clearly visible that during a long period in 2007 observed discharge is overestimated by the model and therefore this year will be excluded from the evaluation period. The hydrological year 2008 will not be excluded, because in this year there is no period where errors are visibly consistently larger than in other years. At the discharge peaks large errors are made, but this also happens sometimes during other years (not for all peaks) and this is probably the reason for the bad *NS* value in 2008.

Table 18: Objective functions per hydrological year in the evaluation period

Hydrological year	*NS*	*RVE* [%]	*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
2007-2013	0.81	8.68	0.74
2009-2013	0.86	3.34	0.83

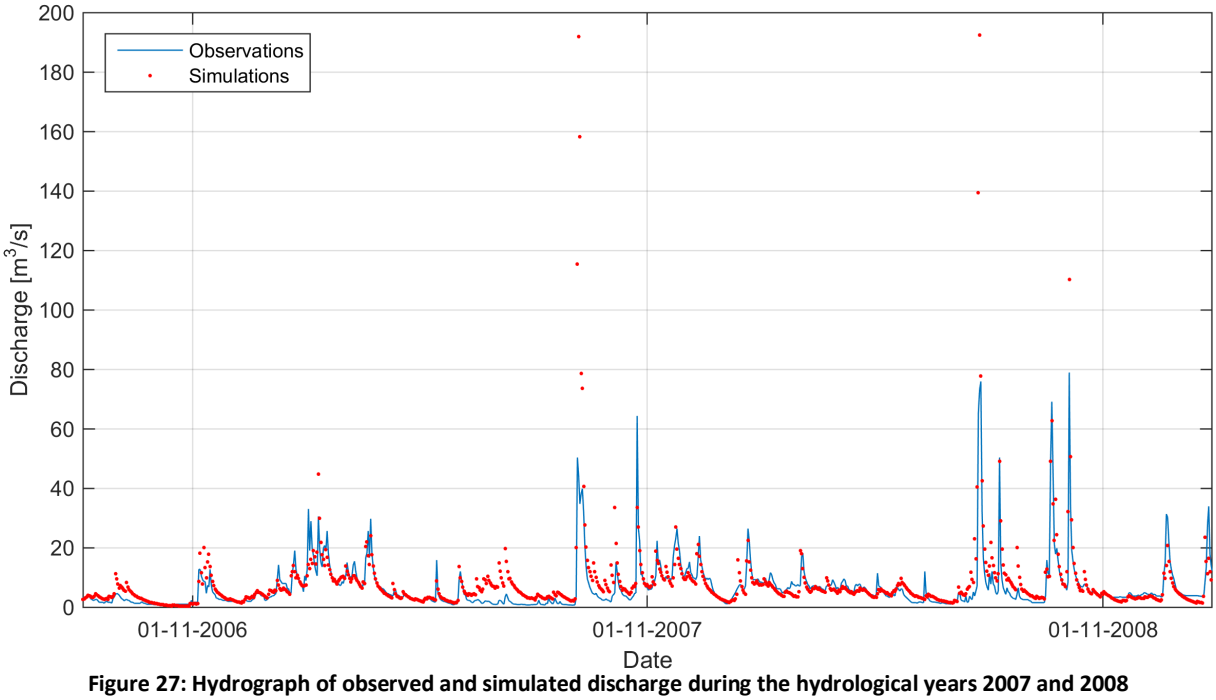


Figure 27: Hydrograph of observed and simulated discharge during the hydrological years 2007 and 2008

4.1.5 Effect of the updating procedure

The effect of the updating procedure will be examined by calculating the objective function over the validation period, for the different updating procedures that are explained in section 3.4. The effectiveness of the updating procedures will be assessed for the deterministic model with observed precipitation and temperature data, because this model is also used in the ensemble flow forecasting system. In Figure 28 the resulting *Y* is presented for different lead times. In this figure a lead time of 0 days means that the initial states are updated every day. A lead time of 10 days means that the initial states are updated and then the model is run 10 days without updating. To correct the overestimation of flows (positive *RVE* in Table 16), in general the updating procedure decreases the surface water storage and groundwater storage compared to simulated non-updated storages. It can be seen that for a lead time of 0 days there is a considerable improvement of the objective function. For larger lead times the improvement becomes smaller, because with larger lead times the initial conditions at day 0 are relatively less important and the summed effect of fluxes becomes more important.

The fraction of fast runoff shows considerable differences between the updating procedures. The differences in discharge are damped, because the outflow from the reservoirs is a part of the total storage in it. The updating procedure that uses total discharge and initial storage to determine *k* is sometimes better and sometimes worse than the simple relationship. According to Figure 28 in general it does not result in an improvement of updating of initial states compared to the simple model. Therefore it has been chosen to use the simplest method (relation between total discharge and *k*). Although the relationship of *k* with initial storage is clear (Figure 15), apparently it does not lead to an improved performance of the deterministic HBV model.

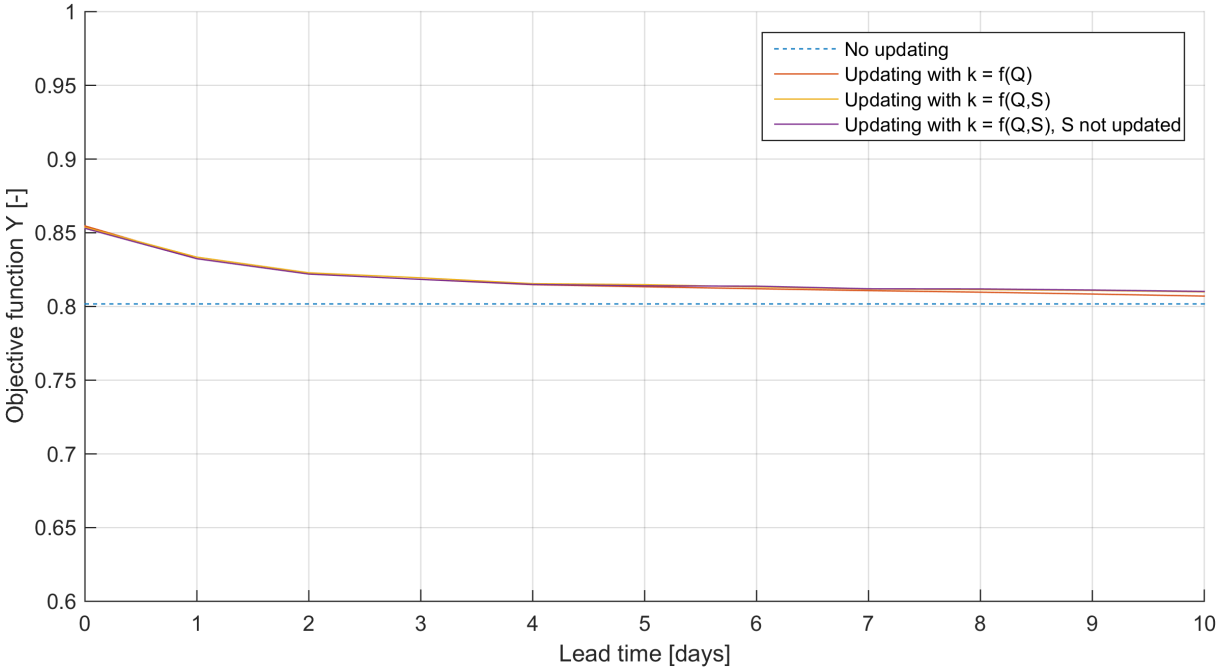


Figure 28: Resulting values of the objective function over the validation period after implementation of the three updating procedures. The lines of the three updating procedures are almost on top of each other.

4.1.6 Hydrological model performance for different purposes

In the ensemble flow forecasting system the hydrological model will be applied to different lead times and different hydrological flow categories. In Table 19 the performance of the deterministic hydrological model with input from observed precipitation and temperature data is presented for

different lead times and hydrological flow categories. The *NS* has not been used here, because the *NS* will not result in a fair comparison between the different hydrological flow categories. Therefore in addition to the *RVE* the *RMAE* has been used.

In Table 19 it can be seen that the results are much better for high flow simulations than for low flow simulations when no updating is applied. This means that the hydrological model parameters are primarily trained on high flows, which is the result of the chosen objective function (also see the discussion in section 5.1.2). Also with updating there is a clear difference between low and high flow simulations, with medium flow simulations in most cases in between these flow categories. The effect of updating on high flows and low flows is visualized in Figure 29 for an example period. It is visible that due to updating the simulations are getting closer to the observations. The updating procedure results in the largest improvement for low flow simulations. The effectiveness of the updating procedure depends on the autocorrelation of discharge, because the updating is based on observations of discharge of the previous time step. In low flow periods there is usually a high autocorrelation of discharge. The HBV model does not perform well for low flows so the performance deteriorates for larger lead times, but it is still much better than with the simulations without updating. The updating has less effect on high flows, because in these flows initial states have relatively less influence and the autocorrelation of discharge is usually less than in low flows.

In Figure 29 a result of the updating procedure that is based on a previous discharge observation can be seen after the small peak at 4-7-2012. The updating procedure adapts the initial states to this small discharge peak and as a result of this the simulated discharges that are based on updating with this peak are also larger than other simulated discharges. The falling limb is simulated much flatter than in the observations, so this effect stays until the maximum lead time of 10 days.

Table 19: Evaluation scores of low, medium and high flow simulations with updated initial states at a lead time of 0 days

Lead time	***RVE* [%]**			***RMAE* [-]**		
	Low flows	Medium flows	High flows	Low flows	Medium flows	High flows
No updating	43.3	7.3	1.8	0.71	0.43	0.33
0	3.2	4.7	2.2	0.11	0.16	0.25
1	6.4	7.2	2.6	0.19	0.21	0.29
2	8.6	8.8	2.5	0.23	0.25	0.31
3	11.5	9.6	2.3	0.29	0.28	0.32
4	13.6	10.1	2.2	0.33	0.30	0.32
5	15.9	10.4	2.0	0.37	0.31	0.32
6	18.2	10.4	2.0	0.41	0.32	0.32
7	19.2	10.5	2.0	0.43	0.34	0.32
8	20.6	10.3	2.1	0.45	0.35	0.32
9	22.9	10.1	2.1	0.49	0.35	0.32
10	24.0	10.0	2.1	0.50	0.36	0.32

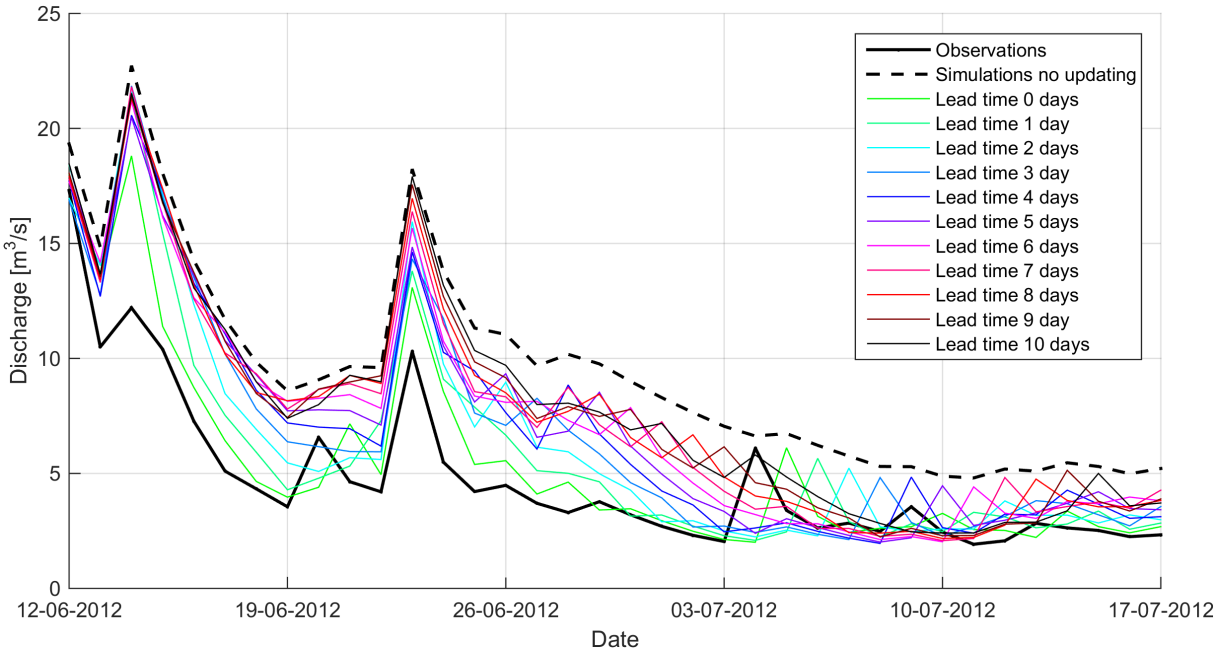


Figure 29: Example of the effect of updating at different lead times

4.2 Validation results of the processing strategies

In section 3.5 3 set-ups of QM in pre-processing and 4 strategies of combining pre- and post-processing are introduced. In this paragraph these set-ups (section 4.2.1) and strategies (section 4.2.2) are evaluated and the most appropriate set-up and strategy are chosen.

The effectiveness of the different QM set-ups and strategies will be evaluated with the CRPS. In section 3.7.3 it has been explained that this is a general evaluation score that combines different properties of forecast quality. This score is also used by other studies to evaluate the effectiveness of pre-processing and post-processing techniques, as single score (Kang et al., 2010) or as one of the evaluation scores (Madadgar et al., 2014; Pagano et al., 2013; Verkade et al., 2013; Zalachori et al., 2012). For an adequate evaluation of the performance of a forecasting system a set of evaluation scores is needed (also argued for in section 3.7), but in this study one of the processing strategies needs to be selected and it is considered that it is not needed to thoroughly analyse the performance on all properties of forecast quality. The set-up with the lowest *CRPS* over lead times from 0 days to 9 days will be used as pre-processing set-up and the strategy with the lowest *CRPS* will be used as processing strategy. In addition the *RMAE* (section 3.7.4.1) and rank histogram flatness coefficient (section 3.7.4.2) are used to show the effect of processing on the relative bias of the ensemble mean and on dispersion of the ensembles. The validation period is from 1-11-2007 to 31-10-2011.

4.2.1 Best pre-processing set-up

In Figure 30, Figure 31 and Figure 32 the evaluation results of the QM set-ups are presented. Regarding precipitation the three QM set-ups all result in an almost equal (little) improvement of the *CRPS*. The *RMAE* and especially the flatness coefficient improve considerably by QM. These results indicate that the correction has improved the precipitation forecasts. The *CRPS* and *RMAE* are almost equal for all QM set-ups and the flatness coefficient of QM separately for each lead time (set-up 2) is on average slightly better, so it has been chosen to apply set-up 2 to pre-process the precipitation forecasts. However the differences are very small, especially when it is remembered that there is a random element in the rank histograms of precipitation forecasts (explained in section 3.7.4.2).

Regarding temperature the *CRPS* of QM with separate lead times and two seasons (set-up 3) is slightly better than the other QM set-ups for most lead times and this is on average approved by the *RMAE* and flatness coefficient. Therefore it is chosen to use set-up 3 to correct the temperature forecasts, which was also expected based on Figure 18 (explained in section 3.5.3.1). A larger improvement is achieved for the temperature forecasts than for the precipitation forecasts. The reason for this is that in the temperature forecasts a more consistent bias is present, so the temperature forecasts are easier to pre-process based on a training period. This bias might be introduced by calculating the daily average temperature (see section 2.3). Over a large part of the cumulative probability domain the CDF of the pre-processed precipitation forecasts during the validation period is actually farther away from the CDF of the observations during the validation period than before the correction. In the temperature forecasts the bias is more consistent, so after pre-processing the CDF of the forecasts is closer to the CDF of the observations than before.

In appendix 5 the rank histograms of precipitation and temperature forecasts are presented for each lead time, before and after pre-processing. The rank histograms of uncorrected precipitation and temperature forecasts are substantial non-uniform. These U-shaped histograms indicate under-dispersion or conditional biases (Hamill, 2001), possibly as a result of under-dispersion and/or bias in the initial states of the meteorological model and as a result of the meteorological model itself. At larger lead times the flatness improves, which is a well-known feature of operational meteorological ensemble prediction systems (Candille & Talagrand, 2005). At larger lead times the influence of initial conditions is less and the spreading of the forecasts becomes larger. In general QM improves the flatness of the rank histograms considerably for both precipitation and temperature forecasts. Regarding precipitation, for small lead times (up to a lead time of 3 days) the histograms still show a non-uniform distribution after pre-processing. Regarding temperature, after pre-processing the rank histograms are non-uniform at all lead times and more non-uniform than the precipitation forecasts.

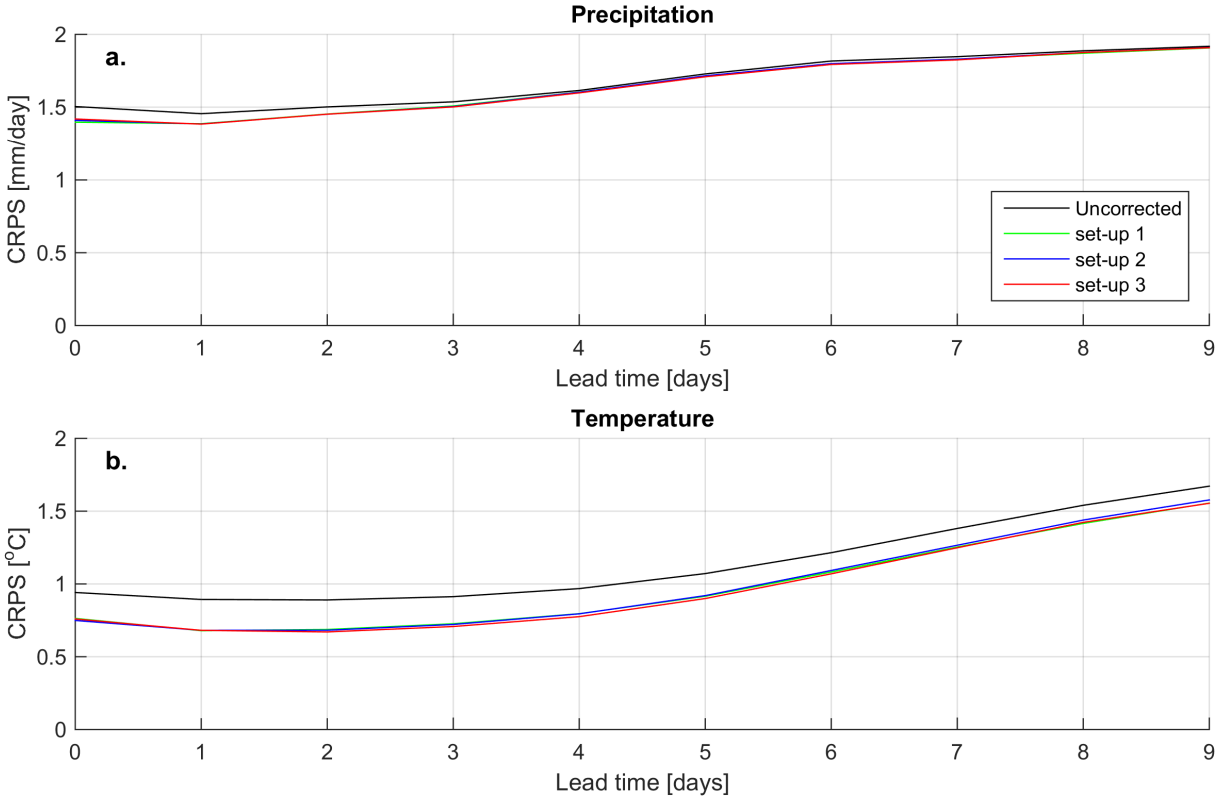


Figure 30: *CRPS* of the QM set-ups and uncorrected forecasts of precipitation (a) and temperature (b), over the validation period 2008-2011. Lines of the different set-ups are almost on top of each other.

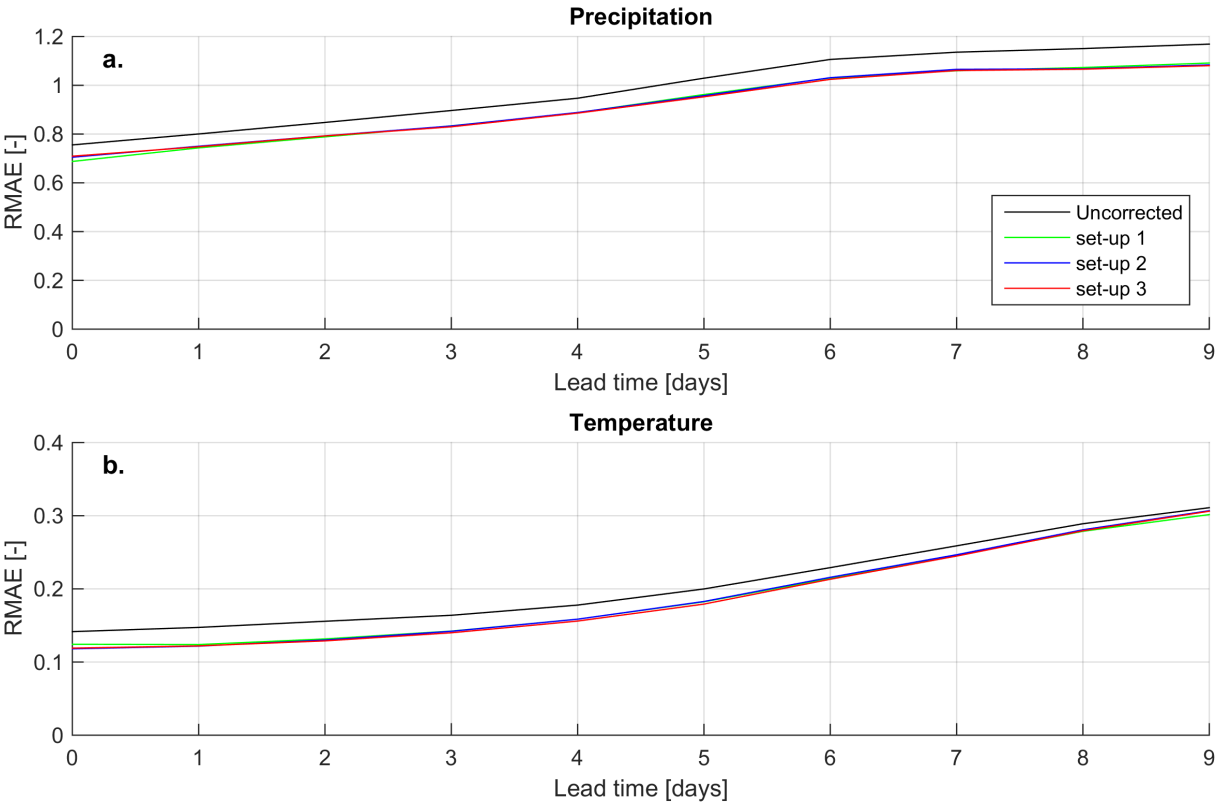


Figure 31: *RMAE* of the QM set-ups and uncorrected forecasts of precipitation (a) and temperature (b), over the validation period 2008-2011. Lines of the different set-ups are almost on top of each other.

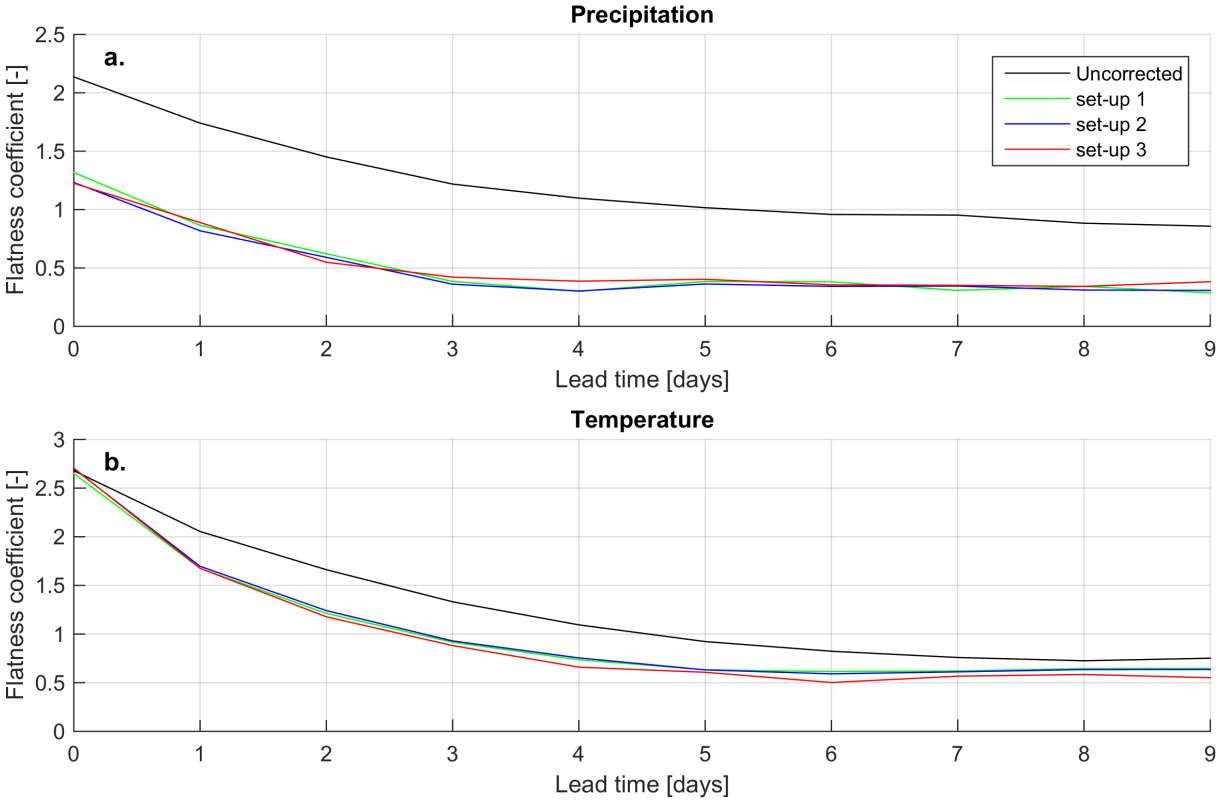


Figure 32: Rank histogram flatness coefficients of different QM set-ups and uncorrected forecasts of precipitation (a) and temperature (b) forecasts, over the validation period 2008-2011

4.2.2 Best processing strategy

The best QM set-up for precipitation and temperature forecasts from section 4.2.1 is applied in strategy 1 and 2 (see Table 11). In Figure 33, Figure 34 and Figure 35 the validation results of the different pre- and post-processing strategies are presented. It is very remarkable that strategy 0 (no correction) results in the best *CRPS* values. Since the flatness coefficient has also worsened by post-processing, the spread of the flow forecasts is worse after post-processing. There is not a large effect on the *RMAE*, so on average the ensemble mean flow forecasts have not clearly been improved or got worse. The best performing processing strategy over the validation period will be applied to the flow forecasts, so no correction at all (strategy 0). Since no processing will be applied, the training period does not need to be excluded from the evaluation period of the ensemble flow forecasts. So the ensemble flow forecasts can be evaluated over the period 2008-2013.

In section 3.5.3 it has already been mentioned that the effectiveness of QM depends on whether during the validation period the same bias is present between the CDF of observations and the CDF of forecasts as during the training period. In Figure 36 it is visible that this bias is not the same over the whole cumulative probability domain for all years. Especially during the hydrological year 2010 (green line) the bias is different, but also for other years on some parts of the cumulative probability domain the correction is in the other direction than during the validation period. Over the cumulative probability domains 0 - 0.2 and 0.85 - 1 the bias during the validation period is in the other direction than during the training period. In addition between a cumulative probability of about 0.4 and 0.7 the bias during the training period is much larger than during the validation period, so in this range the correction is in the good direction but the correction is too large. As a result of these points many corrected forecasts deviate further away from the observations than the uncorrected forecasts. QM usually functions effectively in cases with distant CDFs (consistent bias over whole cumulative

probability domain) (Madadgar et al., 2014), which is clearly not the case here. Madadgar et al. (2014) describe that when two distributions are relatively close, which will be the case for a well-calibrated model, this deficiency of QM becomes more significant. This is further elaborated in the discussion in section 5.2.

In appendix 6 the effects of the different processing strategies on the *CRPS*, *RMAE* and flatness coefficient during the training period are shown. These figures show the potential of processing with QM when a consistent bias is present. Over the training period strategy 3 with seasonal distinction gives the best performance.

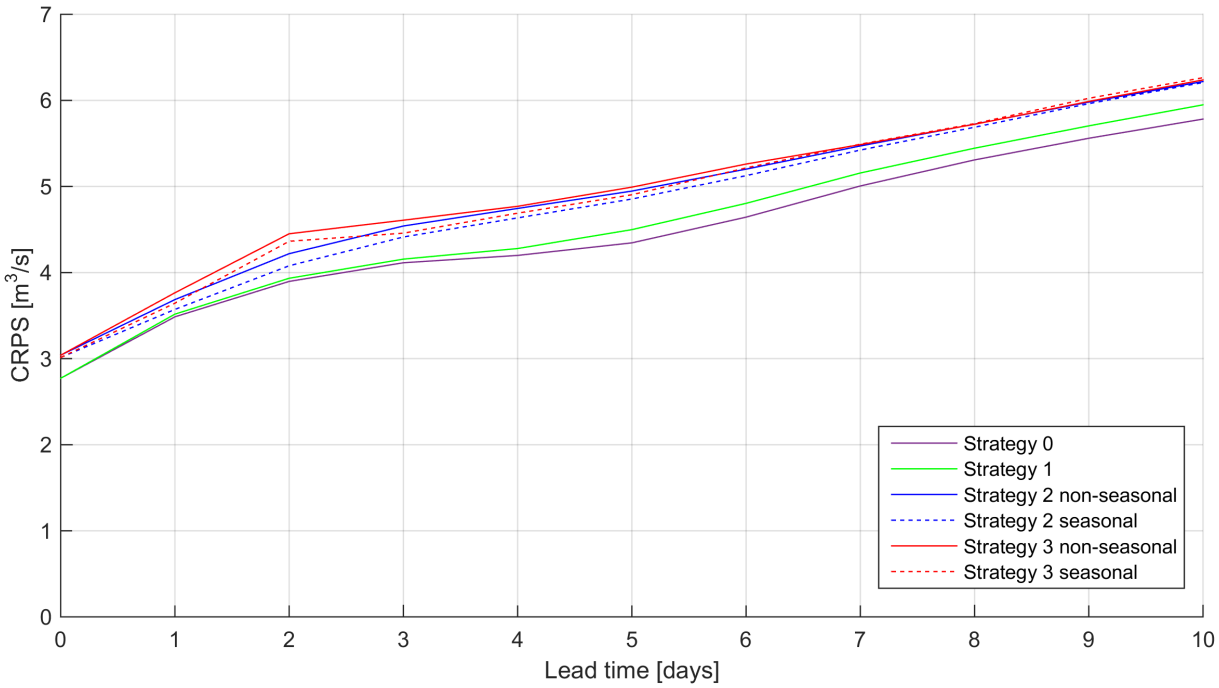


Figure 33: *CRPS* of the post-processing strategies, over the validation period 2008-2011

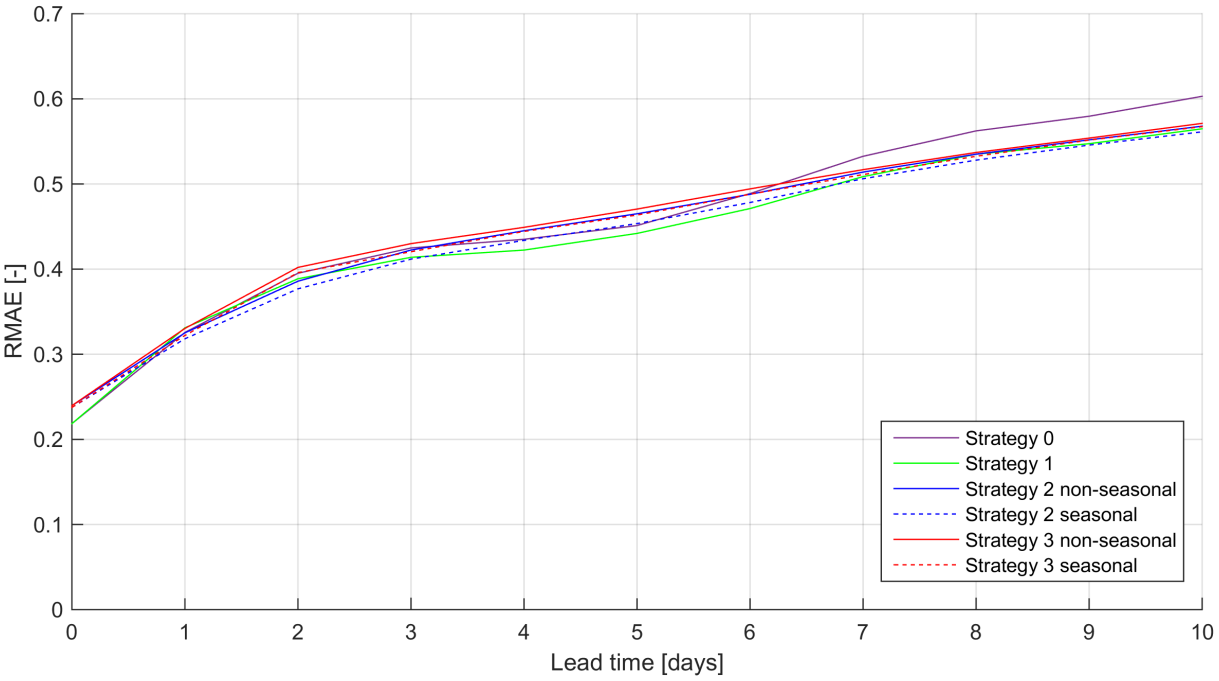


Figure 34: *RMAE* of the post-processing strategies, over the validation period 2008-2011

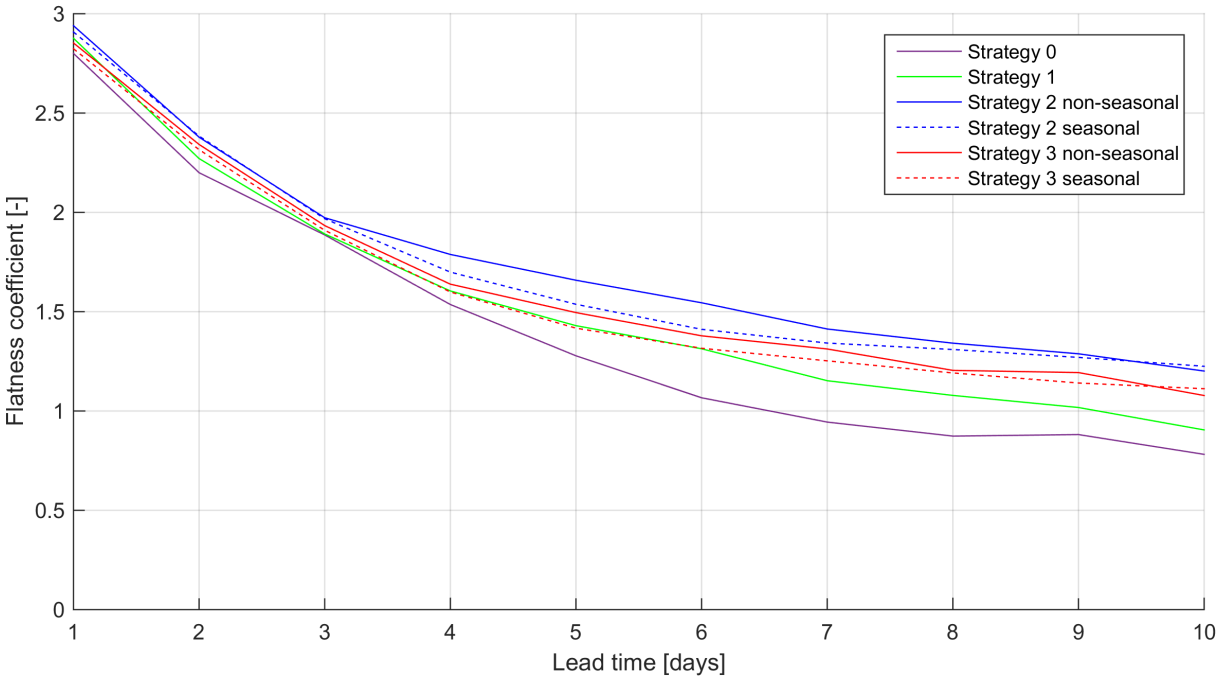


Figure 35: Rank histogram flatness coefficients of the post-processing strategies, over the validation period 2008-2011

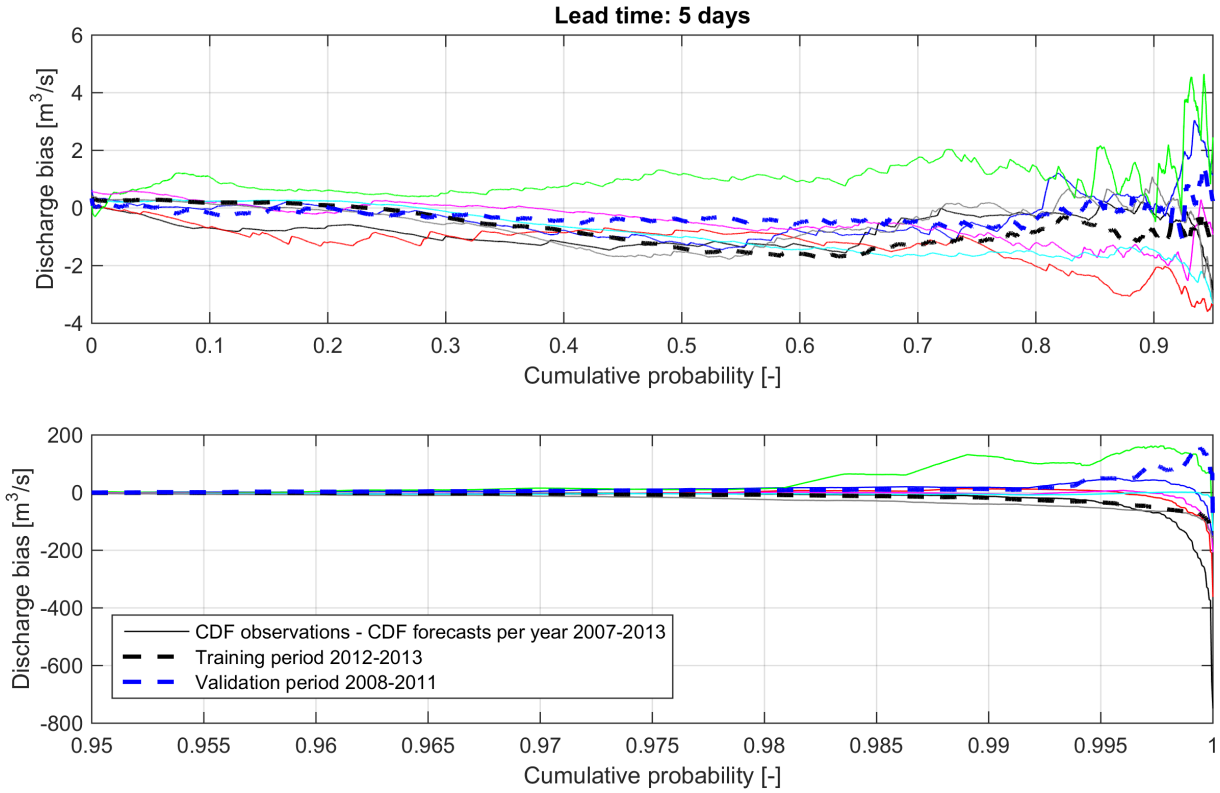


Figure 36: Difference between the CDFs of the observations and the CDFs of the uncorrected flow forecasts per hydrological year. This example is for a lead time of 5 days.

4.3 Evaluation of purposes of ensemble flow forecasts

For research question 2 the ensemble flow forecasts are evaluated for lead times until 10 days and for the low, medium and high flow categories. The hydrological year 2007 has been excluded from the evaluation period (see section 4.1.4), so the ensemble flow forecasts are evaluated over the period from 1-11-2007 to 31-10-2013. Figure 37 presents an example hydrograph forecast. The high flow event at a lead time of 4 days is generated by a rainfall amount of 15 mm at a lead time of

2 days (see Figure 6). From lead times of 1 day and more the flow forecasts are spread, because at a lead time 0 days the discharge is totally determined by the initial states. The spread of meteorological forecasts has effect on the flow forecasts from a lead time of 1 day.

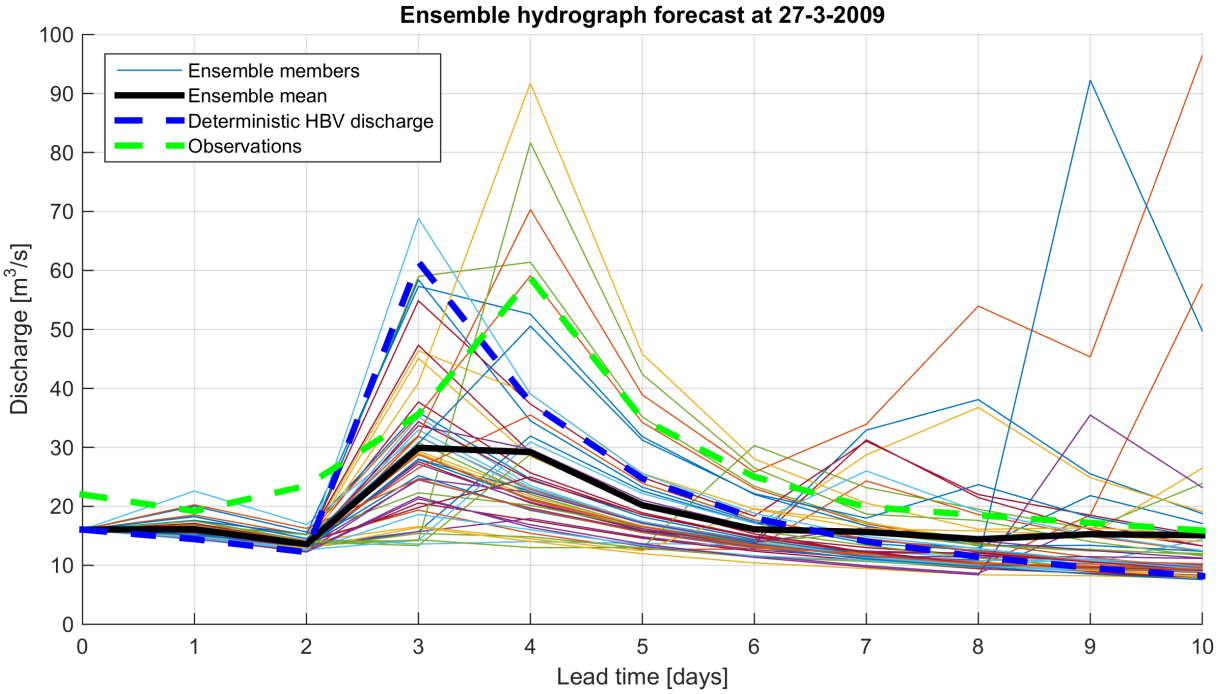


Figure 37: Example of an ensemble flow forecast (see Figure 6 for the precipitation forecast at the same day)

4.3.1 General evaluation

In Figure 38 the *CRPSS* against the reference forecasts is presented. For all flows in general the *CRPSS* is clearly positive for all lead times, so on average the flow forecasts are better than the reference forecasts. The forecast skill is maximum between a lead time of 2 and 5 days. Both the flow forecasts and the reference forecasts are simulated with the same hydrological model and initial conditions, so skill is only generated by the meteorological forecasts compared to historical meteorological observations. This also explains that the values of the CRPSS are not very high, because the hydrological model and the initial conditions are usually very important in the simulation of discharge. For a lead time of 0 days the forecasts have no skill, because both the flow forecasts and the reference forecasts are totally determined by the initial conditions on the forecast day.

Regarding low flows, for small lead times (1-2 days) reference forecasts are better or have almost the same performance. Initial conditions are very important for low flows (explained below Figure 38) and the same initial conditions are used for the reference forecasts and the flow forecasts, so it is difficult to generate skill for low flow forecasts at small lead times. The negative skill at small lead times indicates that for small lead times historical observations of precipitation and temperature are even better forecasts than meteorological ensemble forecasts from ECMWF for this category of flows. Possibly low flow periods are often in the same period of the year, so that historical observations are reasonable forecasts. In addition it often happens that the precipitation and/or temperature observations are totally missed by the meteorological forecasts, especially at short lead times (see the rank histograms of precipitation in appendix 5). After all, meteorological models tend to forecast drizzle instead of zero precipitation (also see section 3.5.3). For lead times of 3 days and more there is a positive skill of low flow forecasts and this is increasing with lead time, although the skills are still not very high. Apparently from a lead time of 3 days the accumulated effects of the

meteorological forecasts are more skilful than historical observations. Regarding medium flows the flow forecasts do not have a clear positive skill for all lead times, which can be explained by the fact that historical discharges are often around the medium discharge so historical observations are a good approximation for these flows. The high flow forecasts have a clear positive skill. Initial conditions are relatively less important in these events, so the flow forecasts and reference forecasts can easier deviate. In addition these events will be less well captured in historical observations and thus in the reference forecasts, because high flow periods are often shorter and not bounded to specific days like low flow periods.

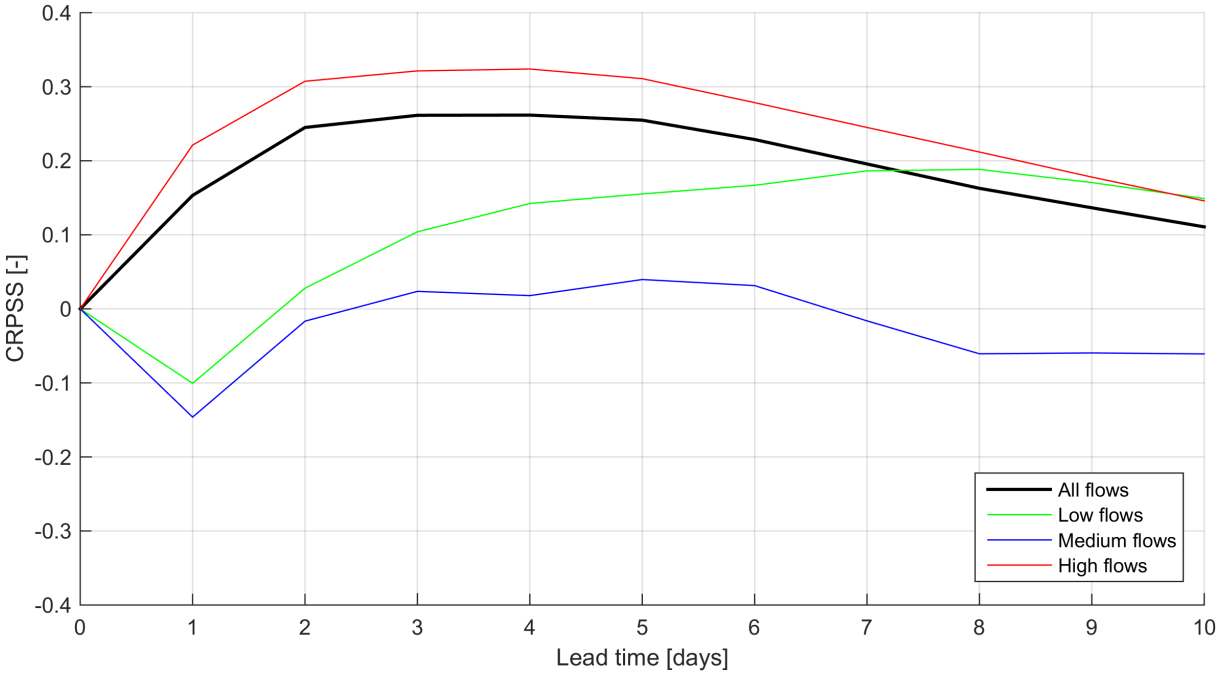


Figure 38: Skill of the flow forecasts, expressed by the CRPS of the flow forecasts compared to the CRPS of the reference forecasts

In Figure 39a the *CRPS* of the flow forecasts is presented and in Figure 39b the relative contribution of meteorological forecast errors (high ratio) and hydrological model errors (low ratio) is presented. The *CRPS* increases with lead time for all different flow categories, so the performance of the flow forecasts compared to observations decreases with lead time for all flow categories (Figure 39a). As explained in section 3.7.3.1 the *CRPS* of the different flow categories cannot directly be compared.

In Figure 39b it can be seen that the relative contribution of meteorological forecast errors increases with lead time, although the model error increases with lead time (Table 19). This is caused by two effects. In the first place the performance of the meteorological forecasts deteriorates with lead time (see Figure 30) and the errors in the meteorological forecasts accumulate with lead time. In the second place the effect of the initial conditions at the forecast day becomes smaller (part of the hydrological model error), because more water is added to the system. It is expected that the effect of initial condition errors will last longer in larger basins, because more water is already in the system and precipitation needs more time to form discharge at the outflow point.

By comparing the different flow categories it can be seen that for high flow forecasts the meteorological forecast error is relatively more important, while for low flow forecasts the hydrological model error is relatively more important. In the first place initial conditions have less influence on high flows. Relatively more water is added to the runoff system by rainfall or snowmelt,

while low flows are mainly generated by the initial conditions. In the second place the hydrological model performance is better for high flows than for low flows, especially at larger lead times (see Table 19), so that meteorological forecast errors are relatively more important. The difference in quality of the meteorological forecasts in low and high flow situations is unknown.

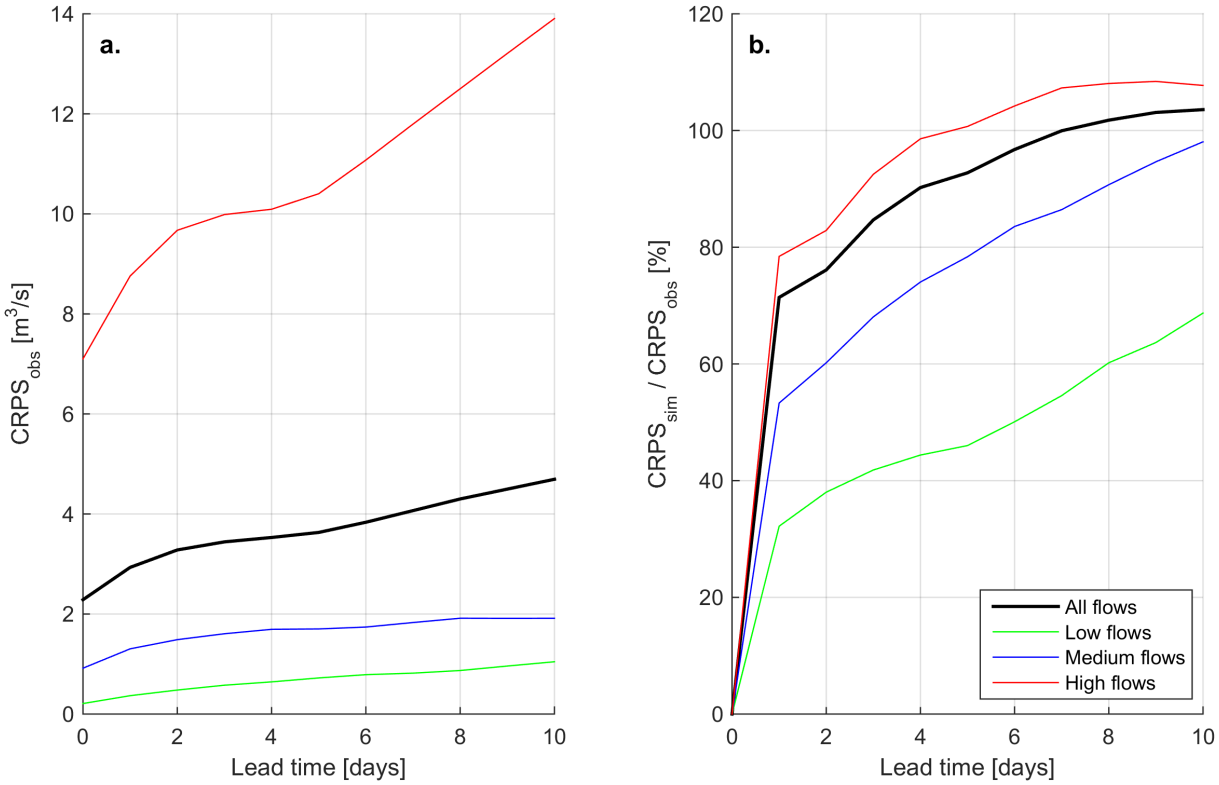


Figure 39: a. *CRPS* against observations. b. Ratio of errors in meteorological forecasts (*CRPS$_{sim}$*) to meteorological + model errors (*CRPS$_{obs}$*)

4.3.2 Evaluation of reliability

Reliability of the ensemble flow forecasting system is evaluated with the rank histogram (section 4.3.2.1) and the reliability diagram (section 4.3.2.2).

4.3.2.1 Rank histogram

In Figure 40 the flatness coefficients of the flow forecasts at different lead times are plotted and in appendix 7 the corresponding rank histograms are presented. The rank histograms are far from flat for all lead times. The U-shaped rank histograms in appendix 7 indicate under-dispersion and/or conditional bias in the flow forecasts. The flatness coefficients of the flow forecasts are higher than the flatness coefficients of the precipitation and temperature forecasts. This must be the result of the initial hydrological model conditions and the hydrological model. The flatness of the rank histogram of flow forecasts improves for larger lead times, so the spread of the ensemble flow forecasts improves with lead time. This could be expected (Bennett et al., 2014; Pagano et al., 2013), because with ensemble meteorological forecasts the dominant source of uncertainty at larger lead times is incorporated (precipitation), while the dominant source of uncertainty at shorter lead times (hydrological processes) is not incorporated (Bennett et al., 2014).

For most lead times the flatness coefficient of high flows is better than the flatness coefficient of low and medium flows. The meteorological forecasts are relatively more important in high flow forecasts than in low flow forecasts and the flatness coefficient of the meteorological forecasts is better, so

this (partly) explains the lower rank histogram flatness. In addition, the results in section 4.1.6 suggest that in low flow forecasts bias from the hydrological model is more present than in high flow forecasts. In appendix 7 the rank histograms of the low, medium and high flow categories are presented. It is clearly visible that a conditional bias is present in the flow forecasts. High flows are in general underestimated by the ensemble flow forecasting system and this effect is increasing with lead time. On the other hand, low flows are overestimated. This can be the result of biased meteorological forecasts and/or hydrological model bias.

It is very remarkable that the flatness coefficient of high flow forecasts increases again after a lead time of 6 days, while all other patterns (including precipitation and temperature forecasts) do not show this. This might be caused by an increased conditional bias in high flow forecasts at larger lead times.

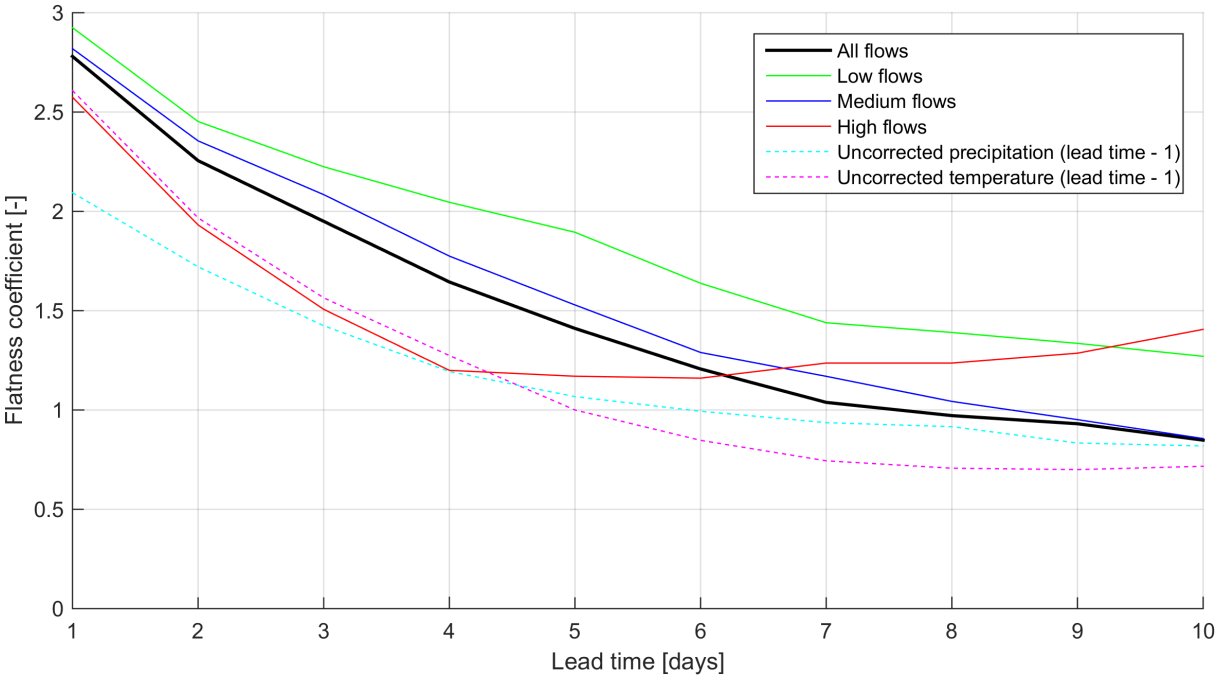


Figure 40: Rank histogram flatness coefficients. The flatness coefficients of the precipitation and temperature forecasts refer to one lead time earlier.

4.3.2.2 Reliability diagram

In Figure 41 the reliability diagrams of high and low flow forecasts are presented for all lead times. Just like the rank histograms, also the reliability diagrams do not show a good reliability, especially for small lead times. From the reliability diagrams it appears that for low flow forecasts the observed relative frequency is in general larger than the forecast probability, so observed relative frequencies are underestimated. The probability bins 0 - 0.2 and 0.8 - 1 are closest to the diagonal line, which makes sense because when the forecast probability is close to 0 or 1 the observation will most often correspond to this.

Regarding the high flow forecasts it is very remarkable that observed relative frequencies are overestimated, while it followed from the rank histograms that high flows are in general underestimated by the hydrological model (conditional bias). This is possible because to construct a rank histogram all observations and forecasts are directly compared, while to construct the reliability diagram the observations and forecasts are compared to the low flow threshold and high flow threshold.

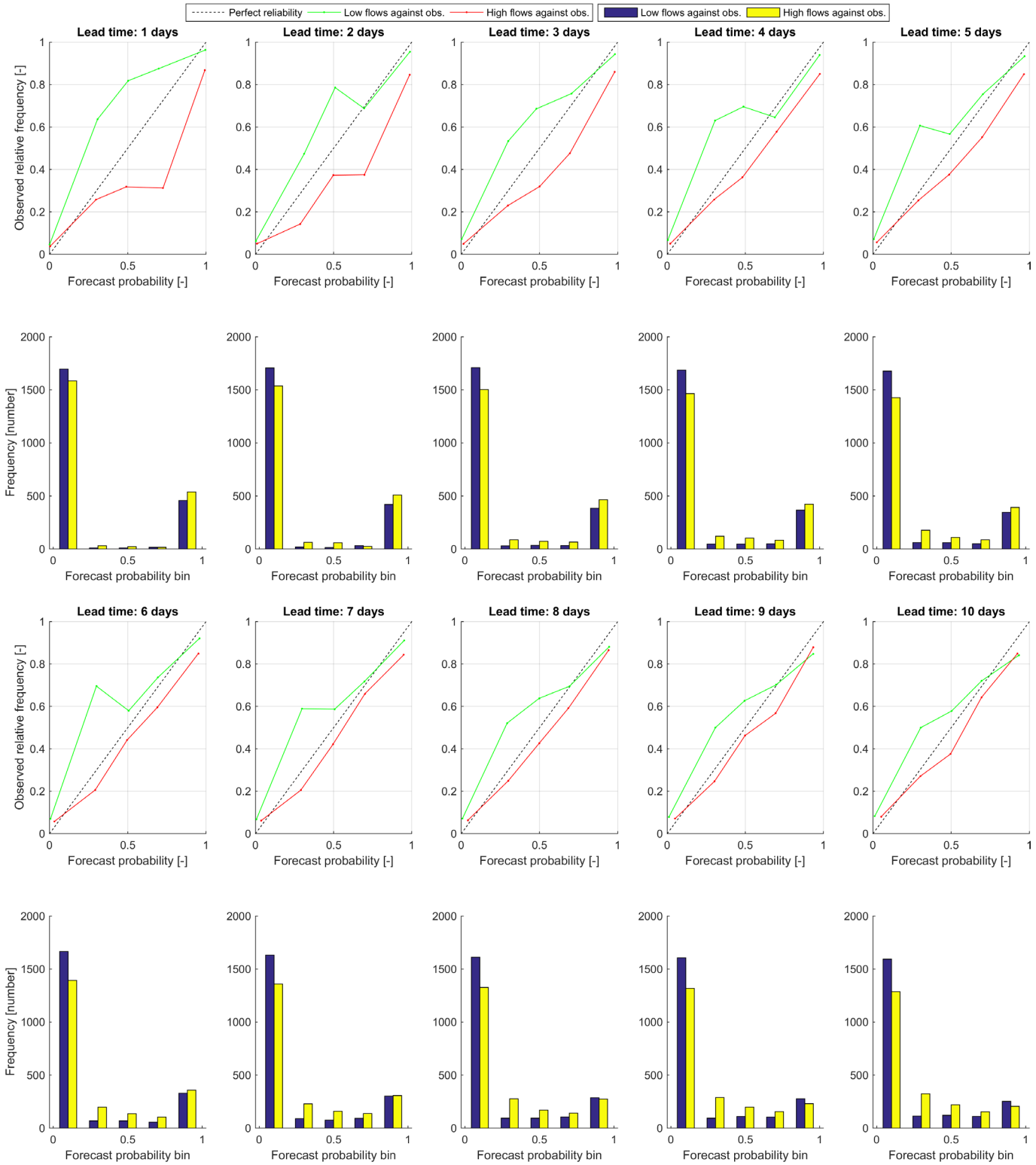


Figure 41: Reliability diagrams (top) and histograms of sample size per bin (under) of low and high flow forecasts for lead times from 1 to 10 days

4.3.3 Evaluation of sharpness

From the frequency histograms accompanied with the reliability diagrams (Figure 41) it can be seen that forecast probabilities close to 0 and 1 most often occur, which qualitatively indicates a good sharpness of the ensemble flow forecasting system. The sharpness decreases with lead time, also see section 4.3.5 for an explanation about this.

4.3.4 Evaluation of resolution

The reliability diagrams in Figure 41 do not show flat curves so it can be qualitatively concluded that the ensemble flow forecasts have a good resolution for both high and low flows. For different forecast probabilities also different relative observed frequencies are found. So if for a certain ensemble forecast the forecast probability of a high flow is close to 1, it is likely that the observation is also above the threshold instead of an observed relative frequency that is always close to a certain value independent of whether the forecast probability is close to 0 or close to 1.

In appendix 8 the ROC curves of low flow forecasts and high flow forecasts are presented and in Figure 42 the *AUC* values of the curves are presented for lead times from 1 day to 10 days. All *AUC* values are above 0.8, which is indicative for good prediction systems according to Buizza et al. (1999). A high *AUC* value means that the hit rate is high compared to the false alarm rate. Low flow forecasts and high flow forecasts perform more or less equal.

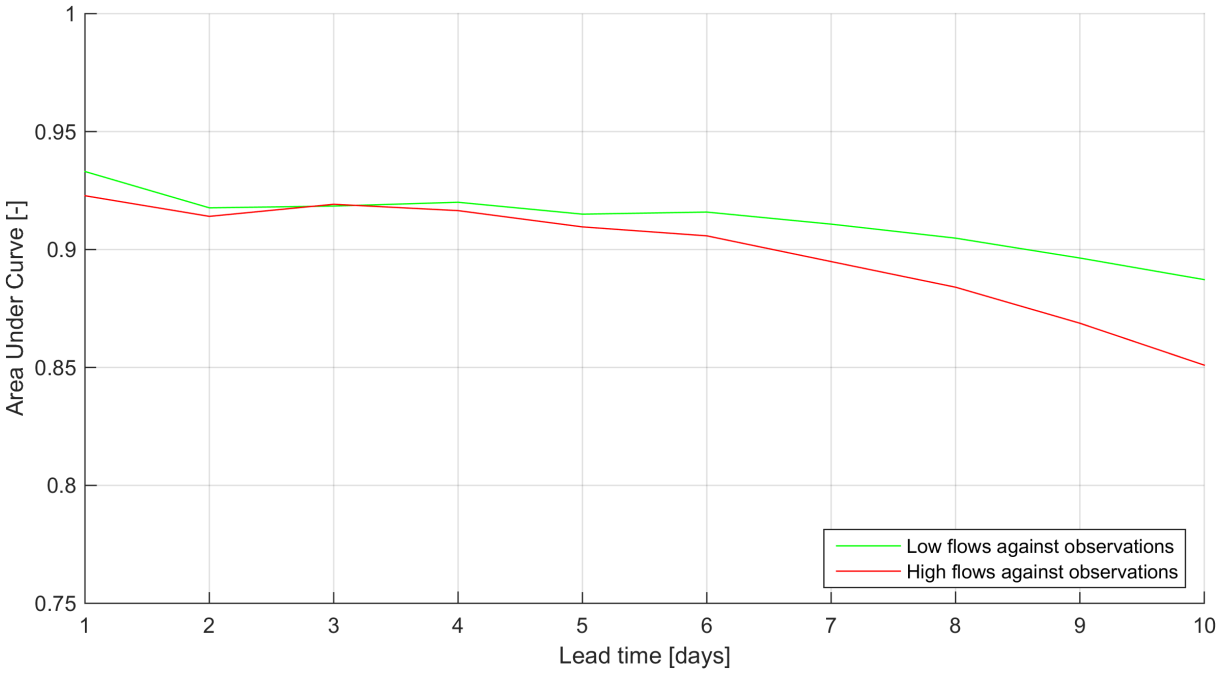


Figure 42: *AUC* of ROC curves for low and high flow forecasts for lead times from 1 to 10 days

4.3.5 Relative uncertainty of predictions

In Figure 43 it can be seen that the uncertainty in the flow forecasts as a result of uncertainty in the meteorological input increases with lead time. The relative uncertainty at a lead time of 0 days is 0, because all ensemble flow forecast members are equal at this lead time. The relative uncertainty of high flow forecasts is larger than the relative uncertainty of medium flow and low flow forecasts.

The relative uncertainty of the ensemble forecasts is related to the reliability and sharpness of the ensemble forecasts. Especially the rank histograms provide insight into this relation. If the spread is larger (larger relative uncertainty) the probability that the observation will be smaller or larger than all ensemble members is smaller, so the under-dispersion will be less pronounced. This means that if additional sources of uncertainty like initial condition uncertainty and model parameter uncertainty are included the *RCI* will increase and the spread in the rank histograms will also improve. On the other hand the sharpness will probably get worse.

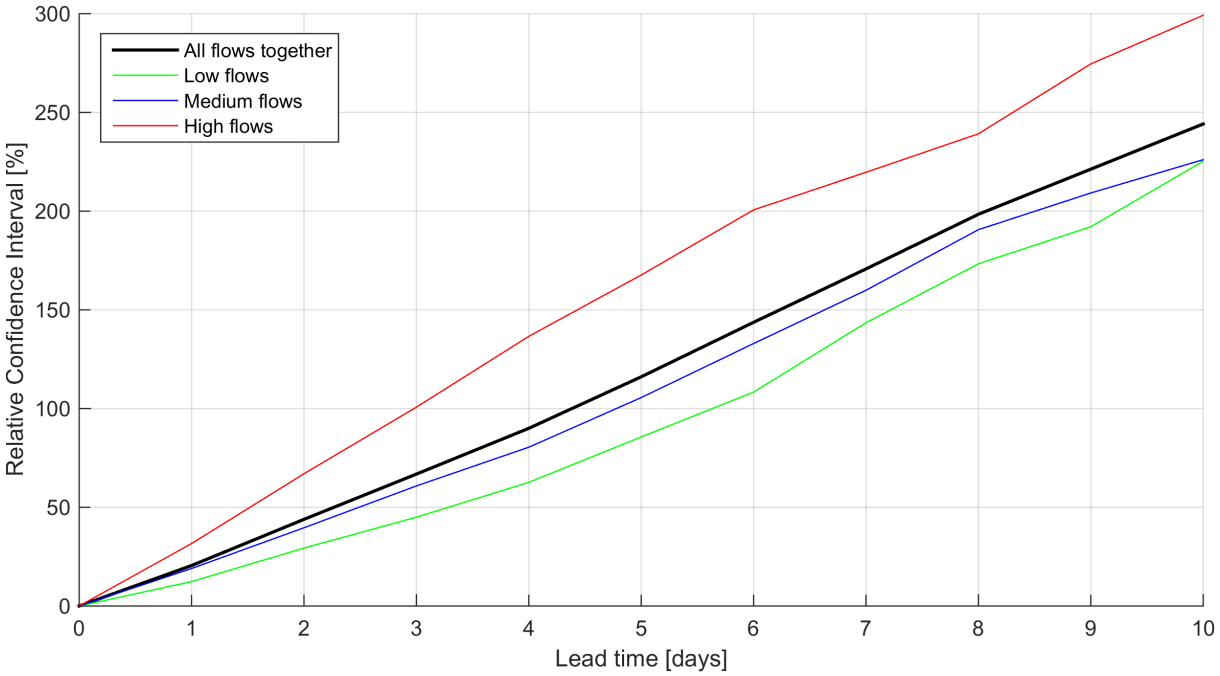


Figure 43: *RCI* for different flow categories for lead times from 1 day to 10 days

4.4 Evaluation of high flow producing processes

The performance of high flow forecasts is further investigated for different high flow producing processes. The high flow producing processes are defined in section 3.9.1. The skill of the different high flow producing processes is presented in Figure 44a and the relative contribution of errors from the meteorological forecasts and from the hydrological model is presented in Figure 44b.

4.4.1 Short-rain floods

In Figure 44a it can be seen that at short lead times high flows generated by short-rain have a large positive skill, so for these flows the precipitation forecasts have much added value compared to historical meteorological observations. Two effects play a role in this. At first, by definition short-rain floods are mainly generated by rain that has fallen one day before the high flow event and the flow is relatively less influenced by the initial conditions. Figure 44b shows that in short-rain flood forecasts the errors in meteorological forecasts are very dominant in the total error, which indicates the importance of meteorological forecasts in these events. As a result, differences between the ensemble flow forecasts and the reference forecasts can be larger and it is better possible to generate forecast skill. At second, short-rain floods are by definition caused by short and heavy rain. Such events do occur mainly in certain periods of the year, but the specific days of these events are different each year. These rainfall events are therefore less well captured in historical meteorological observations. For larger lead times the skill of short-rain generated floods decreases again. This must be the result of a decreased performance of the forecasts of extreme precipitation for larger lead times.

In Figure 44b it is very remarkable that the ratio between the CRPS against perfect forecasts and the *CRPS* against observations is above 100%. This means that the forecasts are closer to the observations than to the perfect forecasts. An example of this is presented in Figure 37 at a lead time of 3 days. This is caused by the fact that in the HBV model always a lag time of 1 day between a high rainfall and the discharge peak is present, while in reality this lag time varies between 1 and 3 days (see section 2.2.3). Especially in a small catchment the timing of the rainfall event on the day (early

or late) is very important in this. The timing of the observed precipitation peak and the forecasted precipitation peak can be shifted a day with regard to each other. This can cause that the peak of the flow forecasts better corresponds to the observations than the peak of the perfect flow forecasts, and the results in Figure 44b indicate that this is regularly the case for this kind of events.

4.4.2 Long-rain floods

The skill of long-rain floods is low for small lead times (Figure 44a). By definition for these events the initial conditions are much more important than for short-rain floods. This is also represented by the larger contribution of model errors (Figure 44b). Since the initial conditions are important it is difficult to deviate from the reference forecasts and thus to generate skilful flow forecasts at small lead times. In addition the results indicate that the meteorological forecasts do not result in positive skill compared to historical meteorological observations for small lead times. With larger lead times the accumulation of rainfall in the catchment becomes important, which is confirmed by a larger contribution of meteorological forecast errors (Figure 44b). Skill is generated from a lead time of 3 days and the highest skill is obtained for lead times between 6 and 10 days.

4.4.3 Snowmelt floods

Regarding snowmelt generated floods Figure 44a shows that the highest skill is obtained for forecasts between 2 and 8 days. For the generation of snowmelt generated floods the initial conditions (also see the relatively large model error contribution in Figure 44b) and temperature forecasts are important. Apparently at lead times of 0 and 1 days the temperature forecasts are not better than the historical observations. For lead times larger than 8 days the skill decreases again. Since the reference forecasts start with the same initial conditions and the same model is used, this must be the result of a decreasing skill of temperature forecasts for larger lead times. This is confirmed by an increasing contribution of meteorological forecast errors in the total error (Figure 44b).

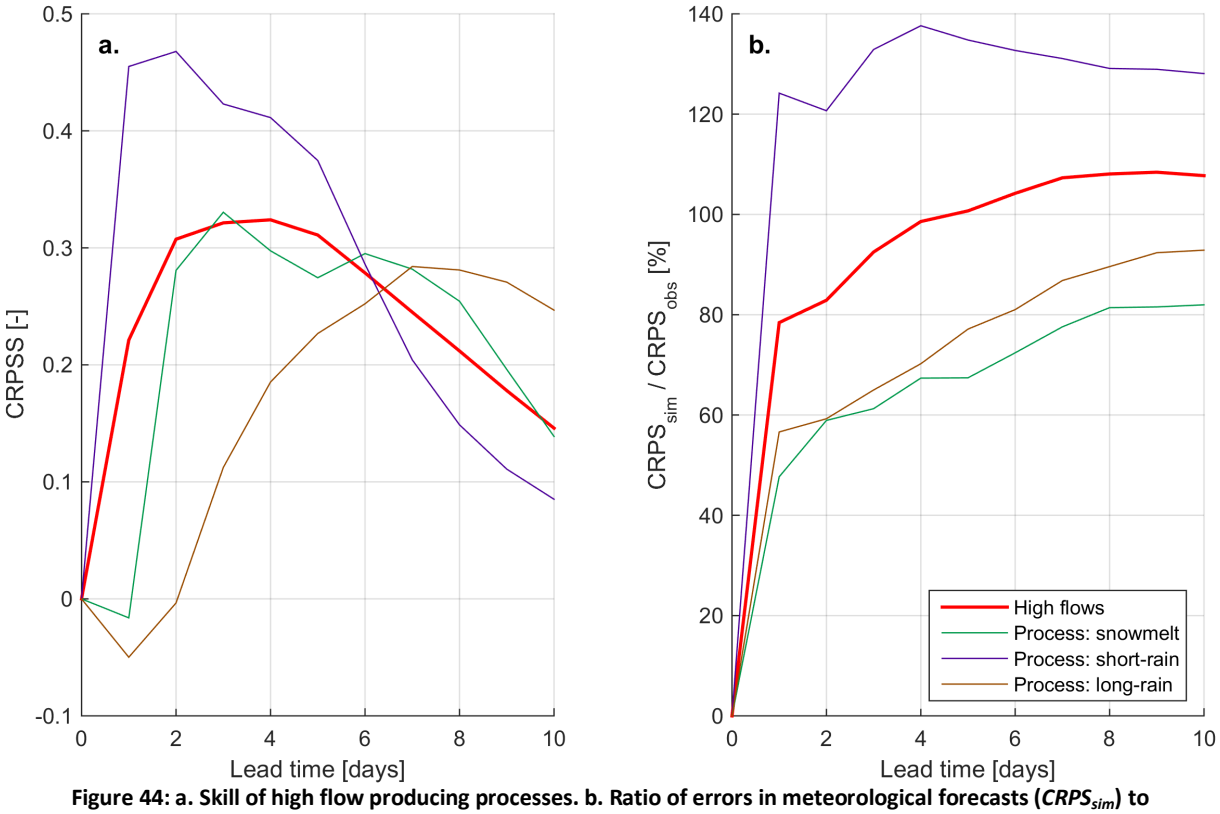


Figure 44: a. Skill of high flow producing processes. b. Ratio of errors in meteorological forecasts ($CRPS_{sim}$) to meteorological + model errors ($CRPS_{obs}$).

4.5 Evaluation of low flow producing processes

The performance of low flow forecasts is further investigated for different low flow producing processes that are defined in section 3.9.2. In Figure 45a the skill of low flow producing processes is presented and in Figure 45b the relative contribution of errors from the meteorological forecasts and from the hydrological model is presented. Figure 45a shows that the low skill of low flow forecasts is mainly caused by low rainfall/high evapotranspiration generated low flow forecasts. The skill of low flows that are caused by snow accumulation is relatively high. For both snow accumulation and low rainfall generated low flows errors from the HBV model and initial conditions make up a large part of the total error (Figure 45b). These errors are also present in the reference forecasts, so this cannot explain the large difference in skill between snow accumulation and low rainfall generated low flows.

4.5.1 Low rainfall/high evapotranspiration generated low flows

For low rainfall generated low flows the precipitation forecasts are important. Apparently the precipitation forecasts before low flow events are worse than historical observations for small lead times. This might have to do with the fact that low rainfall periods often occur in the same period of the year and with the tendency of (uncorrected) meteorological models to forecast drizzle instead of zero precipitation (also see section 3.5.3). For larger lead times the skill increases, so the accumulated precipitation forecasts from ECMWF are better than from historical observations at larger lead times. The fact that the influence of initial conditions at the forecast day decreases for larger lead times (also see Figure 45b) adds to the skill at larger lead times.

4.5.2 Snow accumulation generated low flows

For snow accumulation generated low flows temperature forecasts are especially important. The results indicate that temperature forecasts provide more skill than historical observations of temperature on the same calendar day. For larger lead times (lead time of 9 and 10 days) the skill decreases, like this was also the case with snowmelt generated floods. So also with low flows there is a decreasing skill of temperature forecasts with larger lead times. However, even for a lead time of 10 days the snow accumulation generated low flows and snowmelt generated high flows are skilful forecasts.

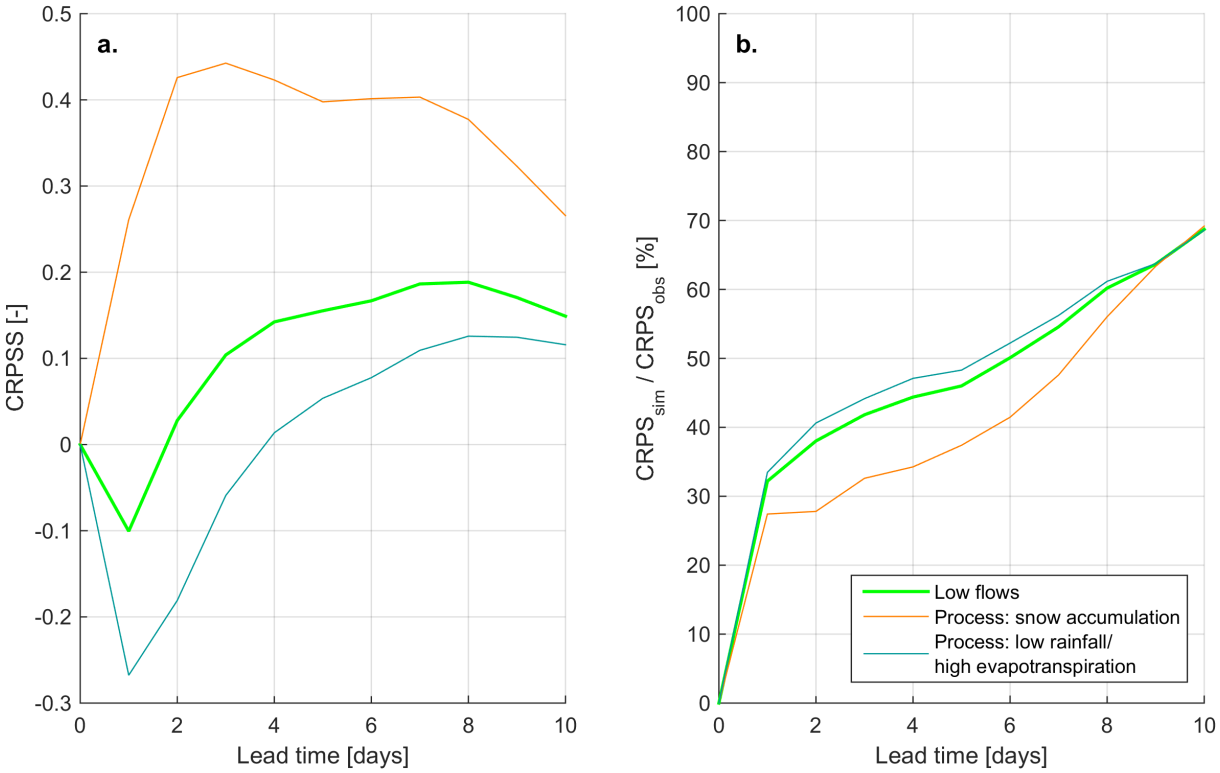


Figure 45: a. Skill of low flow producing processes b. Ratio of errors in meteorological forecasts ($CRPS_{sim}$) to meteorological + model errors ($CRPS_{obs}$).

5. Discussion

In this chapter the research methodology will be reviewed critically (section 5.1 to 5.3) and the results are compared with other studies (section 5.4).

5.1 Input data and calibration of the hydrological model

5.1.1 Input data

Calibration of the HBV model is based on observations of precipitation, temperature and discharge starting from 1971. The quality of the observations over this period is unknown. It has been found that the performance of the hydrological model over the hydrological years 2007 and 2008 is worse than during other years of the evaluation period (Table 18). Also with another model (Data Based Mechanistic methodology (DBM)) with the same observation data the performance was worse during these years (Kiczko et al., 2015). In 2007 during a low flow period of at least 3 months the observed discharge is largely below the simulated discharge and this continues during high flow periods in 2008 (Figure 27). It is expected that this is the result of errors in the discharge, precipitation and/or temperature observations and/or human influence, because for each year the same hydrological model is used and it is unlikely that in this period different hydrological processes are taking place that are not captured well by the HBV model and by the DBM model. This period is very notable, but also during the rest of the validation period and during the calibration period observation errors can be present. On top of this there can be systematic errors in the observations. Errors in the observations have effect on the calibrated model parameters and as a consequence, on the ensemble flow forecasts. This is the indirect effect of observation errors on the flow forecasts.

The precipitation and temperature observations are corrected for elevation differences over the catchment. In the simple correction method that has been applied some assumptions have been made, which are mentioned in section 3.1. Data from only 5 measurement stations have been used to calculate the precipitation gradient and temperature lapse rate. This is a small number of stations compared to other studies. Li et al. (2013) found that when the number of measurement stations is less than 16, the resulting temperature lapse rates show a great variation, which indicates a large uncertainty in the used correction factors. For temperature the global standard atmospheric lapse rate has been used, because the number of measurement stations is small and the temperature lapse rate does not vary significantly over the year. On the contrary the applied precipitation gradients have been based on the measurement stations, because the precipitation gradient clearly varies over the year. It is considered that in this case the determined precipitation gradients for two periods of the year based on 5 measurement stations are more reliable than one default value from literature. Without correction the systematic errors in the precipitation and temperature observation data would be partly captured in the calibration. However, since the hydrological model is also used with another dataset this must be prevented.

In this study meteorological forecasts from ECMWF are used. According to Persson and Andersson (2013) the ECMWF forecasts may represent a land elevation that is significantly different from the real elevation in a grid. If this systematic bias is present this should have been captured by a pre-processing step with QM. QM of the meteorological forecasts did result in an improvement of the meteorological forecasts, but this did not result in an improvement of the flow forecasts (also see section 5.2). This means that still a systematic bias might be present in the meteorological forecasts. In addition, although ECMWF generates an ensemble of meteorological forecasts it should be

realized that ECMWF also uses a meteorological model with certain assumptions, and as a result the meteorological uncertainty might be underrepresented. This can be one of the reasons of under-dispersive meteorological and flow forecasts. To incorporate additional meteorological uncertainty the meteorological forecasts from other forecast systems can be included to establish a 'super-ensemble'. This approach is used by Fleming et al. (2015) and He et al. (2009) and Bennett et al. (2014), Bougeault et al. (2010) and Ranjan (2009) mention this as an option to improve the forecasts.

5.1.2 Calibration procedure

The calibration procedure contains several choices regarding the objective function, calibration period, sensitivity analysis technique, calibration algorithm and parameter ranges. These subjective choices will affect the calibrated parameter values and the model results.

An important choice that directly influences the parameter values is the objective function. The objective function *Y* that has been used is a combination of *RVE* and *NS*. It appears that the *RVE* is almost zero after calibration, so on top of this *NS* almost exclusively determines the value of *Y*. However, *NS* is more sensitive to errors in high discharges so the calibration is skewed to high discharges. Table 19 confirms that the error in low flow simulations is relatively larger than in high flow simulations. Using another objective function or calibrating separately on different flow categories would result in different optimum parameter sets.

Due to the large number of parameters in the HBV model (14 parameters) the problem of overparameterisation might be present (Booij, 2005). This problem is observed in this study, since with two calibrations with 14 parameters two different optimum parameter sets are obtained with the same values of *Y*. This is undesirable, because the simulated discharges will be different with two different parameter sets. In an attempt to circumvent this problem a second calibration round has been applied with only the 8 of the 14 most sensitive parameters. However in the second calibration round the same parameter values were found as in the calibration with 14 parameters for both independent calibrations. This means that the fixed values of the classified insensitive parameters already define the optimum parameter set that has been found. This might be caused by the risk that is reported for the method of Morris to classify important parameters as non-influential (type II error) (Song et al., 2015). In this project one of the calibrated parameter sets is accepted as the optimum deterministic parameter set, because the hydrological model is used as a tool, the differences between most parameters are small and the relative and absolute differences between simulated discharges produced with the two different optimum parameter sets are very small. However, it should be realized that this kind of subjectivity increases uncertainty in the flow forecasts and in the evaluation results.

The calibrated value of the parameter *LP* is 1 (see Table 17), equal to the upper boundary of the parameter range. Also in the GLUE analysis the majority of behavioural *LP* values is near 1. This means that the actual evapotranspiration will only reach potential evapotranspiration if the soil moisture storage is completely saturated. In many other applications of the HBV model the value of *LP* is lower, for example 0.26 and 0.39 are found by Van den Tillaart (2010), 0.23 and 0.87 by Maat (2015) and 0.48 by Demirel et al. (2013a). This value of *LP* suggests that something is wrong with the water balance of the system. Precipitation observations are too low, potential evapotranspiration values are too high or discharge observations are too high. Regarding the potential evapotranspiration, Rao et al. (2011) and references by Rao et al. (2011) found that the method of

Hamon tends to underestimate potential evapotranspiration from forests, although correction factors has been applied for this. On the other hand, in the study of Prudhomme and Williamson (2013) the method of Hamon showed an overestimation of potential evapotranspiration during a large part of the year (spring, summer and autumn).

5.1.3 Update state procedure

The update state procedure that has been applied assumes that one relationship between the fraction of fast runoff and observed discharge can be used. In Figure 12 it can be seen that this relationship contains a large uncertainty. Because this uncertainty is not included, the total uncertainty in the flow forecasts might be underestimated.

The updating procedure direct model storage updating has been chosen, because it is a simple method and in earlier studies it has been found that simple updating models (not direct model storage updating) produce nearly as good results as complex Kalman Filters. It can be researched whether in this application the direct model storage updating produces the same results as more sophisticated updating procedures. Examples are the (Ensemble) Kalman Filter that has been applied to update initial states in several studies on flow forecasting, autoregressive approaches (Xiong & O'Connor, 2002), artificial neural networks (Xiong & O'Connor, 2002), particle filter approaches (Weerts & El Serafy, 2006) and updating of flow forecasts with an error model (Bennett et al., 2014; Piotrowski & Napiorkowski, 2012).

5.2 Pre- and post-processing of forecasts

It turned out that processing with QM was not effective in improving the ensemble flow forecasts. QM has several limitations which might have played a role in this. At first, each variable is corrected independently, while bias in precipitation might be related to bias in temperature (Boé et al., 2007). At second, the temporal autocorrelation properties of the forecasts are not corrected. If for example the simulated wet periods are too short, they remain too short after correction (Boé et al., 2007; Verkade et al., 2013). At third, the spatial correlation (correlation between locations) is not corrected (Boé et al., 2007). The space-time covariability of the meteorological forecasts is important, because the hydrological model integrates the forcing in time and in space (Clark et al., 2004). It is expected that this limitation does not play a large role here, because a relatively small catchment is studied, a lumped model is applied and catchment-average forecasts are corrected against catchment-average observations. When multiple sub-catchments are combined in one system, this can become a problem. At fourth, by QM all ensemble members are corrected independently and the original spreading of the ensemble members is changed. To improve this, an extra procedure should be introduced that takes the spreading of the ensemble members into account.

An essential point for pre- and post-processing in general is the length of available observation and forecast data. The effectiveness of QM depends on whether during the validation period the same bias is present between the CDF of the observations and the CDF of the forecasts as during the training period. In Figure 36 it is shown that the correction of flow forecasts is too large or even in another direction than is required during the validation period over large parts of the cumulative probability domain. A result of this is that in some cases the corrected forecasts even deviate further away from the observations than uncorrected forecasts. QM usually functions effectively in cases with distant CDFs (consistent bias over whole cumulative probability domain) (Madadgar et al., 2014), which is clearly not the case here. If two distributions are relatively close, which will be the

case for a well-calibrated model, this deficiency of QM becomes more significant (Madadgar et al., 2014). It has been shown that the evaluation results of the flow forecasts improve over the training period after post-processing (appendix 6), which shows the potential of processing with QM when a consistent bias is present. Possibly with a longer period of data individual deviating years (like 2010) will have less influence and a more consistent bias can be obtained over the whole cumulative probability domain. In addition it is possible to distinguish different weather patterns to potentially obtain a more consistent bias for different weather patterns. An additional problem, however, is that the joint distribution of observations and forecasts is often nonhomogeneous in time, for example due to improvement of forecasting systems over time (Verkade et al., 2013).

QM is a relatively simple pre- and post-processing technique. Several previous medium-term meteorological and flow forecast studies have applied more sophisticated pre- and post-processing techniques, which are based on other kind of relationships between forecasts and observations and additional predictors. To ascribe realistic spatial and temporal patterns the Schaake shuffle can be used (Clark et al., 2004). Wetterhall et al. (2012) states that the most appropriate method also depends on the objective of the study and that it is important to test different bias correction techniques prior to use one. It should however be realized that the application of all techniques is limited by the limited period of available forecast data and the non-consistent bias over this period.

5.3 Evaluation methodology

Evaluation of the ensemble flow forecasts is an important part of this study, but several reservations should be taken into account. At first, discharge observations are considered as 'truth', but the discharge observations also contain measurement and sampling errors (WMO, 2015). This is the direct effect of observation errors (also see section 5.1) on the evaluation of flow forecasts.

At second, the period of available meteorological forecast data is quite short (6 years). This is the reason that the defined thresholds for low and high flows are not very extreme, because a sufficient number of events is required to evaluate the results. It should be realized that flows below Q_{75} and above Q_{25} are not necessarily very extreme events. If a longer period of data would be available the thresholds for low and high flows could be more extreme. In addition, a longer period of data would provide more confidence in the conclusions.

At third, several assumptions have been made to establish quantitative rules for the classification of low flow and high flow producing processes. It has been assumed that the snow accumulation history before an event is embedded in the initial snowpack storage of the day before the event and if snow is involved the event is classified as a snowmelt flood or snow accumulation low flow. This is of course a strong simplification, because when there is a snowpack in the model there is not necessarily a snowpack over the whole catchment and in addition snow accumulation is not necessarily the most important process in these situations (also see examples described below). Snowmelt floods and rain-on-snow floods are considered as one category, because both processes are related to snowmelt and in the HBV model snowmelt and rainfall are strongly related (both depend on temperature). It is not possible to reliably distinguish between these processes. If no snowpack is present, it has been assumed that the low flow event or high flow event is caused by low/high rainfall. To distinguish between short-rain floods and long-rain floods a semi-arbitrary rainfall threshold amount of 10 mm is used. The threshold of 10 mm is a subjective element in the characterization (see section 3.9.1 for the motivation). When rainfall was below 10 mm and there

was no snowpack present at one day before the event it has been assumed that the high flow must have been caused by long-rain. This has large consequences for the classification of high flows that are caused by a combination of processes. For example, snowmelt has caused saturated conditions in the catchment but if after some time the snowpack has disappeared, a low rainfall can cause high flows. This event will be classified as a long-rain flood because a small amount of rain is the direct driver of the flood, although the wet initial conditions are caused by snowmelt. Another example is that a period of rain causes wet conditions and if on top of this an extreme daily rainfall of at least 10 mm occurs the event will be classified as a short-rain flood, while the long-rain certainly has a role in this flood. These kind of combined processes often occur in practice and they will cause the highest floods, but it is difficult to properly classify them. Another point is that only recent information (one day before the flow event) is used to classify the processes, although it has been assumed that the snow accumulation history is embedded in the snowpack storage from the HBV model. The lag time between precipitation peaks and discharge peaks is not always 1 day, like this is incorporated in the HBV model and the characterization rules. This has the consequence that for example the high flow at the day after a high rainfall amount is classified as a short-rain high flow, while the discharge peak might come one day later and this day will be classified as a long-rain high flow (see for example Figure 6 and Figure 37). Using only recent information is even a larger simplification for low flow classification, because low flow processes usually happen over a longer term. These assumptions must be kept in mind when the evaluation results are interpreted. Nevertheless the classification is based on data that are available and, although it has many limitations, it is considered that it is appropriate for the objective of this study. Another option to examine the performance on different processes could be to use different seasons. However, using these categories instead of seasons provides more insight into the underlying processes and it is for example not necessarily the case that a high flow event in spring is generated by snowmelt (also see Figure 24).

At fourth, statistical significance of the evaluation scores has not been tested in this study. This involves the risk of type I errors, which means that it is concluded that a supposed effect exists when in fact it does not exist (Davis, 2002). In literature several statistical tests are published to test the significance of evaluation scores, like the significance of non-uniformity of rank histograms (Hamill, 2001), confidence intervals around ROC curves (Fawcett, 2006; Wilks, 2006), statistical significance of the *AUC* (Mason & Graham, 2002) and confidence intervals around reliability diagrams (Bröcker & Smith, 2007; Wilks, 2006). Including these tests would provide more confidence in the evaluation results. To include these tests this study requires an extension with a more statistical point of view.

In this study a combination of 5 evaluation scores is used to evaluate the different properties of forecast quality. The different evaluation scores also have limitations in them, which are mentioned in section 3.7. Regarding the *CRPSS* it has been found that the *CRPSS* of all flows in general seems to follow the pattern of the *CRPSS* of high flows, so probably the high flows have the largest influence on the value of the *CRPSS*. Instead or besides these scores other scores could be used, which could have led to different conclusions. Possibilities of other scores are the Brier (Skill) Score, percentile-based evaluation, Probability Integral Transform curves, discrimination diagrams, cost-loss functions and the error-spread score (a new score developed by Christensen et al. (2015)).

5.4 Evaluation results

To examine the performance of the developed ensemble flow forecasting system relative to other ensemble flow forecasting systems the results are compared with previous studies. A similarity between all found studies is that the performance of the flow forecasts decreases with lead time. It is impossible to directly compare the general score *CRPS* to other studies, because *CRPS* is dependent on the magnitude of the evaluated variable. Bennett et al. (2014) also used historical observations of precipitation and temperature to generate reference forecasts, so the *CRP* Skill Score (*CRPSS*) can be compared. The skills of the flow forecasts in this study have the same magnitude and show the same pattern over lead times from 1 to 9 days as in the study of Bennett et al. (2014). The flow forecasting system of Bennett et al. (2014) is different on several points, like different meteorological forecasts with a larger spatial resolution, post-processing of meteorological forecasts, a different hydrological model and an updating procedure of flow forecasts, nevertheless the skill of the forecasts is very comparable. The forecast skill of high flows in the study of Roulin and Vannitsem (2005), expressed by the Brier Skill Score, is in winter higher and in summer comparable with the skill values that has been found in this study. However, the forecast skill decreases for higher flow thresholds (from Q_{20} to Q_{10} and Q_5) (Roulin & Vannitsem, 2005), so this might also be the case with the flow forecasting system in this study. In the introduction multiple studies are mentioned that investigated the contribution of errors in the meteorological forecasts and the hydrological model to errors in the flow forecasts. The results in previous studies do correspond to the findings in this study. The performance on the different hydrological processes cannot be compared with previous studies, because no studies have been found that have investigated the performance of an ensemble flow forecasting system on different hydrological processes.

Ignoring hydrological model and initial condition uncertainty can result in under-dispersive and overconfident flow forecasts (Bennett et al., 2014; Pagano et al., 2013), especially for short lead times (Bennett et al., 2014). The rank histograms of the flow forecasts are more non-uniform than the rank histograms of the precipitation and temperature forecasts, so this must be the result of the hydrological model and the initial conditions. The additional under-dispersion can potentially be improved by including hydrological model parameter and initial condition uncertainty. In appendix 9 it is shown that incorporating model parameter uncertainty by a GLUE analysis results in general in a better performance of the flow forecasts at small lead times, especially for medium and high flow forecasts. The flatness of the rank histograms has improved by including model parameter uncertainty, but the flow forecasts are still under-dispersed. After all the uncertainty of the flow forecasts (expressed by *RCI*) has increased, especially at small lead times. In appendix 9 the results are discussed more extensively. The GLUE analysis that has been applied contains a large subjectivity and many assumptions, but this simple implementation shows the effect of including model parameter uncertainty. With another threshold for behavioural model parameter selection the results are very different. For example with a lower threshold ($Y = 0.3$) the relative frequencies of the most extreme ranks in the rank histograms become slightly better and the *RCI* increases. Possibly this lower threshold is required to correctly incorporate the model parameter uncertainty, but this requires further research.

6. Conclusions and recommendations

The research questions are answered in the conclusions in section 6.1. The results are further explained in chapter 4. From the conclusions and the discussion follow several recommendations that are mentioned in section 6.2.

6.1 Conclusions

In section 1.5 three research questions have been formulated. In the first research question the ensemble flow forecasting system is set-up for the Biała Tarnowska catchment, in the second research question the performance of the flow forecasting system is investigated for different purposes and in the third research question the performance is investigated for different hydrological circumstances.

> *Research question 1: What is the most appropriate set-up of input data, the hydrological model and the calibration procedure to obtain an ensemble flow forecasting system for the Biała Tarnowska catchment?*

In the first research question it is investigated how the ensemble flow forecasting system should be developed to generate ensemble flow forecasts for the Biała Tarnowska catchment. This consists of several steps. In general it can be concluded that the most elaborated techniques are not always the best techniques.

To calibrate the lumped deterministic hydrological (HBV) model precipitation, temperature and discharge observation data have been used. The precipitation and temperature observation data are corrected for differences in elevation between the observation stations and the rest of the catchment, because precipitation and temperature are dependent on elevation. The calibration and validation results are better with the corrected input data than with the uncorrected input data, which means that the systematic underestimation of precipitation and systematic overestimation of temperature were not totally captured in the calibrated parameters. The hydrological model is calibrated with DEGL, because this is an efficient method to find the global optimum deterministic parameter set. The calibration has resulted in a *Y* value of 0.72 over the validation period, with *NS* equal to 0.77 and *RVE* equal to 6.7%. The performance of the deterministic HBV model with observed precipitation and temperature is much better for high flow simulations than for low flow simulations.

To improve the representation of the current situation in the catchment the initial model conditions are updated based on available discharge observations. As updating procedure direct model storage updating has been chosen, because this is a simple approach and in earlier studies it has been found that simple updating procedures produce nearly as good results as complex Kalman Filtering algorithms. The simplest implementation of direct model storage updating and more elaborated implementations in which also non-updated initial conditions are used result in almost the same improvement of the deterministic simulations, so it has been chosen to use the simplest approach. The updating results in a considerable improvement of especially low flow simulations and this effect is still noticeable after 10 days without updating.

Ensemble precipitation and temperature forecasts from ECMWF are used as input data to the hydrological model to generate flow forecasts. In theory raw meteorological forecasts need to be pre-processed, because the scale at which they are generated is usually coarser than the application,

the forecasts are provided as grid values and because of under-dispersivity and systematic bias in the meteorological forecasts. It has been considered that downscaling to correct for the first two points is not required, because a lumped hydrological model is used and the scale of the meteorological forecast grids is already comparable with the scale of the catchment. To correct for under-dispersivity and systematic bias a pre-processing step should be applied. Post-processing of flow forecasts is required, because hydrological models also introduce simulation biases that degrade forecast quality. It has been chosen to apply QM as bias and dispersion correction technique, because it is a simple method and it has often been advised and used by previous studies in both pre- and post-processing procedures. For the pre-processing step it turned out that for the precipitation forecasts the best set-up is to apply QM separately to each lead time and for the temperature forecasts the best set-up is to apply in addition also separate relationships for the summer and winter season. However, the best flow forecasts were obtained if no pre- or post-processing is applied at all.

> *Research question 2: How does the ensemble flow forecasting system perform for different purposes and how does this relate to errors from meteorological input data and the hydrological model?*

The purposes that are evaluated are lead times from 1 day to 10 days and low, medium and high flow forecasts. The performance of the flow forecasts deteriorates with lead time in terms of the CRPS. The skill of the flow forecasts is determined with respect to the performance of a reference forecast set. It has been found that reference forecasts based on an ensemble of historical observations of precipitation and temperature on the same calendar day over the past 20 years are the most appropriate reference forecasts.

In general the skill of the flow forecasts is positive and maximum between lead times of 2 and 5 days. The forecast skill is very different for the low, medium and high flow categories. The low flow forecasts do not have skill until a lead time of 2 days and after that they show a small positive skill. They even have a small negative skill for small lead times, which means that it would be better to use the reference flow forecasts than the flow forecasts that are generated by the meteorological forecasts. Apparently the historical observations of temperature and precipitation are in this case better than the meteorological forecasts from ECMWF. The medium flow forecasts provide no skill. The highest skills are obtained for high flow forecasts. High flow forecasts can easier deviate from the reference forecasts than low flow forecasts, because initial conditions are less important. In addition, historical meteorological observations will be less good compared to the meteorological forecasts, because high flow periods are in general shorter and not bounded to specific days like low flows.

It applies to all flow categories that the relative contribution of meteorological forecast errors increases and the relative contribution of hydrological model errors (including initial conditions) decreases with lead time. In low flow forecasts errors from the hydrological model are relatively more important, while in high flow forecasts errors from the meteorological forecasts are relatively more important.

Three properties of forecast quality of ensemble forecast systems are reliability, sharpness and resolution. The sharpness of the forecasts is good, because forecast probabilities of high and low flows are most often close to 0 and 1, instead of forecast probabilities close to the mean (climatological) probability. The ROC curves show a good resolution with areas under the ROC curve

well above 0.8, which is indicative for good prediction systems (Buizza et al., 1999). This means that the hit rates of high and low flow forecasts are high compared to the false alarm rates. However, the rank histograms and reliability diagrams show that the reliability of the flow forecasting system is not good, especially for small lead times. The rank histograms show a clear non-uniform pattern. This indicates under-dispersion and/or conditional bias in the flow forecasts. Since the general performance in terms of the *CRPS* is better at small lead times, the spreading of the ensemble flow forecasts is not good at small lead times but the ensemble members are on average closer to the observations than at larger lead times. The rank histograms of the low, medium and high flow forecasts show that conditional bias is present. High flows are in general underestimated and low flows are overestimated. The reliability improves with lead time. With increasing lead time the relative uncertainty of the ensemble forecasts increases and as a result of this the reliability increases and the sharpness decreases.

> *Research question 3: How does the ensemble flow forecasting system perform for different hydrological circumstances and how does this relate to errors from meteorological input data and the hydrological model?*

Three different high flow producing processes have been distinguished, which are short-rain floods, long-rain floods and snowmelt floods. The forecast skill compared to the reference forecast set is very different for these processes. The highest skill is obtained for the short-rain flood forecasts, but the skill decreases considerably for lead times larger than 5 days. In the generation of long-rain floods and snowmelt floods initial conditions are by definition much more important, which is also reflected in the higher relative contribution of hydrological model errors. This means that it is more difficult for the flow forecasts to deviate from the reference forecasts and thus to generate skilful forecasts at small lead times. The flow forecasting system generates skilful long-rain flood forecasts for lead times of 3 days and more and skilful snowmelt flood forecasts for lead times of 2 days and more. The skill of snowmelt flood forecasts is higher than the skill of long-rain flood forecasts until a lead of 6 days, so it can be concluded that the temperature forecasts that generate snowmelt floods provide more skill compared to historical observations than the meteorological forecasts that generate long-rain floods. With larger lead times the skill of the flood forecasts decreases again, which must be the result of a decreasing skill of meteorological forecasts compared to historical observations for larger lead times. This is confirmed by an increasing contribution of meteorological forecast errors in the total error. The skill of short-rain flood forecasts decreases the most and at the shortest lead time (from a lead time of 6 days).

As low flow producing processes low rainfall/high evapotranspiration generated low flows and snow accumulation generated low flows have been distinguished. The low skill of low flow forecasts is mainly caused by low rainfall/high evapotranspiration generated low flow forecasts, while the skill of snow accumulation low flows is relatively high (also compared to the skill of high flow forecasts). It can be concluded that temperature forecasts are more skilful than precipitation forecasts for the generation of these events compared to historical observations on the same calendar day. Regarding the low rainfall/high evapotranspiration generated low flows for lead times until 4 days there is even a negative forecast skill, so the meteorological forecasts are worse than historical observations. For both snowmelt floods and snow accumulation generated low flows there is decreasing skill from a lead time of 8 days, so there is a decreasing skill of temperature forecasts at large lead times.

6.2 Recommendations

The discussion and conclusions lead to several recommendations for the use of and further research to ensemble flow forecasting systems, which are mentioned below. The most important recommendations are recommendations 1, 5 and 9.

1. The developed ensemble flow forecasting system can be used to generate skilful flow forecasts for high flows and for low flows that are generated by snow accumulation in the Biała Tarnowska. Medium flows and low flows that are generated by low rainfall/high evapotranspiration are not skilfully forecasted, compared to reference flow forecasts that are based on historical observations of precipitation and temperature on the same calendar day.
2. It is expected that recommendation 1 also applies to comparable catchments, but it is recommended to also research other catchments to be able to draw more general conclusions. If it is possible it is also recommended to evaluate the flow forecasting system with a longer period of data, to increase confidence in the conclusions and to test more extreme high and low flow thresholds, before the system is potentially applied operationally.
3. Potentially the ensemble flow forecasting system of the Biała Tarnowska catchment can be combined with systems for other catchments to generate flow forecasts at a larger hydrological scale (e.g. the Dunajec catchment). When multiple sub-catchments are combined in one flow forecasting system the spatial correlation of meteorological forecasts should get attention in a pre-processing step.
4. When the ensemble flow forecasting system is used for all flows together, like in this study, it is recommended to choose another objective function or to calibrate separately on high and low flows.
5. It is recommended to start further development of the ensemble flow forecasting system with a consideration of what the objectives of the system are. When this or another flow forecasting system is (further) developed with the objective to generate high flow forecasts it is recommended to focus further research mainly on improving the meteorological forecasts. This can be obtained with improved meteorological forecasts (like the higher resolution forecasts from COSMO-LEPS (Renner et al., 2009)) or by improving the pre-processing step. When the flow forecasting system will be applied to forecast low flows it is recommended to focus further research at first mainly on the hydrological model performance. As mentioned in the discussion the calibration was skewed to high discharges, so it is expected that an ‘easy’ improvement of the forecasts can be achieved when the hydrological model would be calibrated on low flow situations. Besides improvement of the hydrological model, further research should be done to improve the meteorological forecasts, especially the precipitation forecasts (problem of drizzle).
6. With QM a relatively simple pre- and post-processing technique has been applied. It is recommended to do further research to other pre- and post-processing techniques to improve the meteorological and flow forecasts.
7. When the flow forecasting system is applied exclusively on low or high flow forecasts the reference forecast set should be reconsidered. Probably there will be another most appropriate reference forecast set when the system is only applied to low or high flows, instead of all discharges.
8. The updating procedure results in a large improvement of the discharge simulations from the model, so it is advised to always use an updating procedure in an ensemble flow forecasting system. The procedure that has been used results in a considerable improvement of the

simulations, but it can be researched whether other updating procedures could result in a larger improvement.

9. In this study a simple implementation of GLUE to incorporate hydrological model parameter uncertainty has resulted in an increase of the reliability of the flow forecasts. It is recommended to further research how hydrological model and initial condition uncertainty should be included in the ensemble flow forecasting system to improve the reliability of the flow forecasts. An alternative way to improve the reliability is by post-processing of the flow forecasts (Olsson & Lindström, 2008; Pagano et al., 2013), although post-processing was not successful in this study.
10. Not only the ensemble flow forecasts show non-uniform rank histograms and bad reliability, but also the precipitation and temperature forecasts do show this (to a less extent). So to improve the reliability of the flow forecasts it is recommended to also research how the reliability of the meteorological forecasts can be improved, potentially by including meteorological forecasts from other forecast systems ('super-ensemble') or by pre-processing.
11. In this study meteorological forecasts until a maximum lead time of 10 days have been used, following the definition of medium-range forecasts. However, via the TIGGE data portal ECMWF forecasts are available until a lead time of 15 days, so these data can be used to extend the ensemble flow forecasting system until a lead time of 15 days.
12. This study has focused on the use of ensemble meteorological forecasts to generate ensemble flow forecasts. The study could be extended with investigating the use of deterministic meteorological forecasts, for example the high-resolution forecasts (HRES) of ECMWF. In section 1.1 theoretical advantages of ensemble forecasts have been mentioned, and by also investigating the use of deterministic meteorological forecasts the added value of ensemble forecasts could be confirmed or weakened.
13. To increase confidence in the conclusions, this study should be extended with a more statistical point of view to incorporate statistical tests for the evaluation scores.

References

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

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, *517*, 913–922. doi:10.1016/j.jhydrol.2014.06.035

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

Bennett, J. C., Robertson, D. E., Shrestha, D. L., & Wang, Q. J. (2013). Selecting reference streamflow forecasts to demonstrate the performance of NWP-forced streamflow forecasts. In J. Piantadosi, R. S. Anderssen, & J. Boland (Eds.), *MODSIM2013, 20th International Congress on Modelling and Simulation* (pp. 2611–2617). Adelaide, Australia: Modelling and Simulation Society of Australia and New Zealand. Retrieved from http://www.mssanz.org.au/modsim2013/L8/bennett.pdf

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

Bergström, S. (1995). The HBV model. In V. P. Singh (Ed.), *Computer Models of Watershed Hydrology* (pp. 443–476). Highlands Ranch, Colorado USA: Water Resources Publications.

Bergström, S., & Lindström, G. (2015). Interpretation of runoff processes in hydrological modelling-experience from the HBV approach. *Hydrological Processes*, *29*(16), 3535–3545. doi:10.1002/hyp.10510

Beven, K. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*(1), 41–51. doi:10.1016/0309-1708(93)90028-E

Blasone, R. S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A., & Zyvoloski, G. A. (2008). Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Advances in Water Resources*, *31*(4), 630–648. doi:10.1016/j.advwatres.2007.12.003

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

Booij, M. J. (2005). Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, *303*(1-4), 176–198. doi:10.1016/j.jhydrol.2004.07.013

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., … Worley, S. (2010). The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, *91*(8), 1059–1072. doi:10.1175/2010BAMS2853.1

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

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

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

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

Burgers, G., Van Leeuwen, P. J., & Evensen, G. (1998). Analysis Scheme in the Ensemble Kalman Filter. *Monthly Weather Review*, *126*(6), 1719–1724. doi:10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling and Software*, *22*(10), 1509–1518. doi:10.1016/j.envsoft.2006.10.004

Campolongo, F., Saltelli, A., & Cariboni, J. (2011). From screening to quantitative sensitivity analysis. A unified approach. *Computer Physics Communications*, *182*(4), 978–988. doi:10.1016/j.cpc.2010.12.039

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, *131*(609), 2131–2150. doi:10.1256/qj.04.71

Carpenter, T. M., Sperfslage, J. A., Georgakakos, K. P., Sweeney, T., & Fread, D. L. (1999). National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *Journal of Hydrology*, *224*(1-2), 21–44. doi:10.1016/S0022-1694(99)00115-8

Chen, H., Yang, D., Hong, Y., Gourley, J. J., & Zhang, Y. (2013). Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting. *Advances in Water Resources*, *59*, 209–220. doi:10.1016/j.advwatres.2013.06.010

Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015). Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 538–549. doi:10.1002/qj.2375

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

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3-4), 613–626. doi:10.1016/j.jhydrol.2009.06.005

Cloke, H. L., Pappenberger, F., Van Andel, S. J., Schaake, J., Thielen, J., & Ramos, M. H. (2013). Preface: Hydrological ensemble prediction systems. *Hydrological Processes*, *27*(1), 1–4. doi:10.1002/hyp.9679

Das, S., Abraham, A., Chakraborty, U. K., & Konar, A. (2009). Differential Evolution Using a Neighborhood-Based Mutation Operator. *IEEE Transactions on Evolutionary Computation*, *13*(3), 526–553. doi:10.1109/TEVC.2008.2009457

Davis, J. C. (2002). *Statistics and Data Analysis in Geology* (3rd ed.). New York, New York USA: John Wiley & Sons.

De Jong, S., Wanders, N., & De Roo, A. (2012). Satellieten helpen overstromingen te voorspellen. *Geografie*, *21*(9), 10–13.

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

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

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

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

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

Dobrowski, S. Z., Abatzoglou, J. T., Greenberg, J. A., & Schladow, S. G. (2009). How much influence does landscape-scale physiography have on air temperature in a mountain environment? *Agricultural and Forest Meteorology*, *149*(10), 1751–1758. doi:10.1016/j.agrformet.2009.06.006

ECMWF. (2012). Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter*, *133*, 11–13. Retrieved from http://old.ecmwf.int/publications/newsletters/pdf/133.pdf

Eumetcal. (n.d.). Computation of the Rank Probability Score (RPS) - Accuracy. Retrieved April 20, 2015, from www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_prob_forec/uos2b/uos2b_ko1.htm

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

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

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

Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, *32*(2), 135–154. doi:10.1007/BF00547132

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

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

Hamill, T. M., Mullen, S. L., Snyder, C., Toth, Z., & Baumhefner, D. P. (2000). Ensemble forecasting in the Short to Medium Range: Report from a Workshop. *Bulletin of the American Meteorological Society*, *81*(11), 2653–2664. doi:10.1175/1520-0477(2000)081<2653:EFITST>2.3.CO;2

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

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

Herman, J. D., Kollat, J. B., Reed, P. M., & Wagener, T. (2013). Technical Note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models. *Hydrology and Earth System Sciences*, *17*(7), 2893–2903. doi:10.5194/hess-17-2893-2013

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

Intergovernmental Panel on Climate Change. (2014). *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*. (Core Writing Team, R. K. Pachauri, & L. A. Meyer, Eds.). Geneva, Zwitzerland. Retrieved from http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full.pdf

Jin, X., Xu, C. Y., Zhang, Q., & Singh, V. P. (2010). Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. *Journal of Hydrology*, *383*(3-4), 147–155. doi:10.1016/j.jhydrol.2009.12.028

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

Kiczko, A., Romanowicz, R. J., Osuch, M., & Pappenberger, F. (2015). Adaptation of the integrated catchment system to on-line assimilation of ECMWF forecasts. In R. J. Romanowicz & M. Osuch (Eds.), *Stochastic Flood Forecasting System - The Middle River Vistula Case Study* (pp. 173–186). Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-18854-6_11

Klemes, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. doi:10.1080/02626668609491024

Knoben, W. J. M. (2013). *Estimation of non-stationary hydrological model parameters for the Polish Welna catchment*. MSc thesis, University of Twente, Enschede, the Netherlands.

Komma, J., Blöschl, G., & Reszler, C. (2008). Soil moisture updating by Ensemble Kalman Filtering in real-time flood forecasting. *Journal of Hydrology*, *357*(3-4), 228–242. doi:10.1016/j.jhydrol.2008.05.020

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

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. doi:10.1016/j.jcp.2007.02.014

Li, H., Luo, L., Wood, E. F., & Schaake, J. (2009). The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research: Atmospheres*, *114*(D4). doi:10.1029/2008JD010969

Li, X., Wang, L., Chen, D., Yang, K., Xue, B., & Sun, L. (2013). Near-surface air temperature lapse rates in the mainland China during 1962-2011. *Journal of Geophysical Research: Atmospheres*, *118*(14), 7505–7515. doi:10.1002/jgrd.50553

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

Lu, J., Sun, G., McNulty, S. G., & Amatya, D. M. (2005). A Comparison of Six Potential Evapotranspiration Methods for Regional Use in the Southeastern United States. *Journal of the American Water Resources Association*, *41*(3), 621–633. doi:10.1111/j.1752-1688.2005.tb03759.x

Maat, W. H. (2015). *Simulating discharges and forecasting floods using a conceptual rainfall-runoff model for the Bolivian Mamoré basin*. MSc thesis, University of Twente, Enschede, the Netherlands.

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

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

Martinec, J., Rango, A., & Roberts, R. (2008). *Snowmelt Runoff Model (SRM) User's Manual*. (E. Gómez-Landesa & M. P. Bleiweiss, Eds.). Retrieved from http://www.nmworkssouthern.nmsu.edu/pubs/research/weather_climate/SRMSpecRep100.pdf

Marzban, C. (2004). The ROC Curve and the Area under It as Performance Measures. *Weather and Forecasting*, *19*(6), 1106–1114. doi:10.1175/825.1

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145–2166. doi:10.1256/003590002320603584

Merz, R., & Blöschl, G. (2003). Regional flood risk - what are the driving processes? In G. Blöschl, S. Franks, M. Kumagai, K. Musiake, & D. Rosbjerg (Eds.), *Water Resources Systems-Hydrological Risk, Management and Development* (pp. 49–58). Wallingford, UK: International Association of Hydrological Sciences. Retrieved from http://hydrologie.org/redbooks/a281/iahs_281_049.pdf

Ministerie van Verkeer en Waterstaat, Ministerie van Volkshuisvesting, Ruimtelijke Ordening en Milieubeheer, & Ministerie van Landbouw, Natuur en Voedselkwaliteit. (2009). *Beleidsnota Waterveiligheid 2009-2015*. Den Haag, the Netherlands. Retrieved from http://www.rijksoverheid.nl/documenten-en-publicaties/notas/2009/12/22/beleidsnota-waterveiligheid-2009-2015.html

Moradkhani, H., Sorooshian, S., Gupta, H. V., & Houser, P. R. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources*, *28*(2), 135–147. doi:10.1016/j.advwatres.2004.09.002

Morris, M. D. (1991). Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics*, *33*(2), 161–174. doi:10.2307/1269043

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

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. doi:10.1016/0022-1694(70)90255-6

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

Osuch, M., Romanowicz, R. J., & Booij, M. J. (2015). The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics. *Hydrological Sciences Journal*. doi:10.1080/02626667.2014.967694

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2 - Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. *Journal of Hydrology*, *303*(1-4), 290–306. doi:10.1016/j.jhydrol.2004.08.026

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

Paiva, R. C. D., Collischonn, W., Bonnet, M. P., & De Gonçalves, L. G. G. (2012). On the sources of hydrological prediction uncertainty in the Amazon. *Hydrology and Earth System Sciences*, *16*(9), 3127–3137. doi:10.5194/hess-16-3127-2012

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., & Smith, L. (2006). Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*, *106*, 10–17. Retrieved from http://old.ecmwf.int/publications/newsletters/pdf/106.pdf

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

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., … Salamon, P. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble predictions. *Journal of Hydrology*, *522*, 697–713. doi:10.1016/j.jhydrol.2015.01.024

Pechlivanidis, I. G., Jackson, B. M., Mcintyre, N. R., & Wheater, H. S. (2011). Catchment scale hydrological modelling : a review of model types , calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. *Global Nest Journal*, *13*(3), 193–214. Retrieved from http://journal.gnest.org/sites/default/files/Journal Papers/193-214_778_Pechlivanidis_13-3.pdf

Penning-Rowsell, E. C., Tunstall, S. M., Tapsell, S. M., & Parker, D. J. (2000). The benefits of Flood Warnings: Real But Elusive, and Politically Significant. *Journal of the Chartered Institution of Water and Environmental Management*, *14*(1), 7–14. doi:10.1111/j.1747-6593.2000.tb00219.x

Persson, A., & Andersson, E. (2013). *User guide to ECMWF forecast products*. Retrieved from http://old.ecmwf.int/products/forecasts/guide/user_guide.pdf

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

Piotrowski, A. P., & Napiorkowski, J. J. (2012). Product-Units neural networks for catchment runoff forecasting. *Advances in Water Resources*, *49*, 97–113. doi:10.1016/j.advwatres.2012.05.016

Prudhomme, C., & Williamson, J. (2013). Derivation of RCM-driven potential evapotranspiration for hydrological climate change impact analysis in Great Britain: A comparison of methods and associated uncertainty in future projections. *Hydrology and Earth System Sciences*, *17*(4), 1365–1377. doi:10.5194/hess-17-1365-2013

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

Rao, L. Y., Sun, G., Ford, C. R., & Vose, J. M. (2011). Modeling Potential Evapotranspiration of Two Forested Watersheds in the Southern Appalachians. *American Society of Agricultural and Biological Engineers*, *54*(6), 2067–2078. Retrieved from http://www.srs.fs.usda.gov/pubs/ja/2011/ja_2011_rao_001.pdf?

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

Romanowicz, R. J., Osuch, M., & Grabowiecka, M. (2013). On the choice of calibration periods and objective functions: A practical guide to model parameter identification. *Acta Geophysica*, *61*(6), 1477–1503. doi:10.2478/s11600-013-0157-6

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

Schaake, J., Pailleux, J., Thielen, J., Arritt, R., Hamill, T., Luo, L., … Pappenberger, F. (2010). Summary of recommendations of the first workshop on Postprocessing and Downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15-18 June 2009. *Atmospheric Science Letters*, *11*(2), 59–63. doi:10.1002/asl.267

Serban, P., & Askew, A. J. (1991). Hydrological forecasting and updating procedures. In F. H. M. Van der Ven, D. Gutknecht, D. P. Loucks, & K. A. Salewicz (Eds.), *Hydrology for the Water Management of Large River Basins* (pp. 357–369). Wallingford, UK: International Association of Hydrological Sciences. Retrieved from http://hydrologie.org/redbooks/a201/iahs_201_0357.pdf

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

Shen, Z. Y., Chen, L., & Chen, T. (2012). Analysis of parameter uncertainty in hydrological and sediment modeling using GLUE method: A case study of SWAT model applied to Three Gorges Reservoir Region, China. *Hydrology and Earth System Sciences*, *16*(1), 121–132. doi:10.5194/hess-16-121-2012

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

Shin, M. J., Guillaume, J. H. A., Croke, B. F. W., & Jakeman, A. J. (2013). Addressing ten questions about conceptual rainfall-runoff models with global sensitivity analyses in R. *Journal of Hydrology*, *503*, 135–152. doi:10.1016/j.jhydrol.2013.08.047

Shrestha, D. L. (2014). Continuous rank probability score. Melbourne, Australia: Commonwealth Scientific and Industrial Research Organization. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/47807-continuous-rank-probability-score

Singh, V. P. (1995). *Computer Models of Watershed Hydrology*. Highlands Ranch, Colorado USA: Water Resources Publications.

Skynet. (2009). Reliability Diagram for calibration of two-class predictor. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/25704-reliability-diagram-for-calibration-of-two-class-predictor

Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., & Xu, C. (2015). Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. *Journal of Hydrology*, *523*, 739–757. doi:10.1016/j.jhydrol.2015.02.013

Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, *44*. doi:10.1029/2008WR006822

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

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

Thielen, J., Bogner, K., Pappenberger, F., Kalas, M., Del Medico, M., & De Roo, A. (2009b). Monthly-, medium-, and short-range flood warning: testing the limits of predictability. *Meteorological Applications*, *16*(1), 77–90. doi:10.1002/met.140

Thiemann, M., Trosset, M., Gupta, H., & Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research*, *37*(10), 2521–2535. doi:10.1029/2000WR900405

Tracton, M. S., & Kalnay, E. (1993). Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects. *Weather and Forecasting*, *8*(3), 379–398. doi:10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2

Trinh, B. N., Thielen-del Pozo, J., & Thirel, G. (2013). The reduction continuous rank probability score for evaluating discharge forecasts from hydrological ensemble prediction systems. *Atmospheric Science Letters*, *14*(2), 61–65. doi:10.1002/asl2.417

Van den Tillaart, S. P. M. (2010). *Influence of uncertainties in discharge determination on the parameter estimation and performance of a HBV model in Meuse sub basins*. MSc thesis, University of Twente, Enschede, the Netherlands.

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

Verbunt, M., Walser, A., Gurtz, J., Montani, A., & Schär, C. (2007). Probabilistic Flood Forecasting with a Limited-Area Ensemble Prediction System: Selected Case Studies. *Journal of Hydrometeorology*, *8*(4), 897–909. doi:10.1175/JHM594.1

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

Vormoor, K., Lawrence, D., Heistermann, M., & Bronstert, A. (2015). c. *Hydrology and Earth System Sciences*, *19*(2), 913–931. doi:10.5194/hess-19-913-2015

Vrugt, J. A., Gupta, H. V., Bouten, W., & Sorooshian, S. (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, *39*(8). doi:10.1029/2002WR001642

Weerts, A. H., & El Serafy, G. Y. H. (2006). Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resources Research*, *42*(9). doi:10.1029/2005WR004093

Werner, M. G. F., Schellekens, J., & Kwadijk, J. C. J. (2006). Flood early warning systems for hydrological (sub) catchments. In M. G. Anderson & J. J. McDonnell (Eds.), *Encyclopedia of Hydrological Sciences*. John Wiley & Sons. doi:10.1002/0470848944.hsa022

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

Wilks, D. S. (2006). *Stastistical methods in the atmospheric sciences* (2nd ed.). Oxford, UK: Elsevier.

Wilks, D. S., & Hamill, T. M. (2007). Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, *135*(6), 2379–2390. doi:10.1175/MWR3402.1

World Meteorological Organization. (2015). Forecast Verification: Issues, Methods and FAQ. WWRP/WGNE Joint Working Group on Verification. Retrieved March 12, 2015, from http://www.cawcr.gov.au/projects/verification/

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

Wood, A. W., Leung, L. R., Sridhar, V., & Lettenmaier, D. P. (2004). Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Climatic Change*, *62*(1-3), 189–216. doi:10.1023/B:CLIM.0000013685.99609.9e

Xiong, L., & O'Connor, K. M. (2002). Comparison of four updating models for real-time river flow forecasting. *Hydrological Sciences Journal*, *47*(4), 621–639. doi:10.1080/02626660209492964

Yazdi, J., Salehi Neyshabouri, S. A. A., & Golian, S. (2014). A stochastic framework to assess the performance of flood warning systems based on rainfall-runoff modeling. *Hydrological Processes*, *28*(17), 4718–4731. doi:10.1002/hyp.9969

Ye, A., Duan, Q., Schaake, J., Xu, J., Deng, X., Di, Z., … Gong, W. (2015). Post-processing of ensemble forecasts in low-flow period. *Hydrological Processes*, *29*(10), 2438–2453. doi:10.1002/hyp.10374

Ye, J., He, Y., Pappenberger, F., Cloke, H. L., Manful, D. Y., & Li, Z. (2014). Evaluation of ECMWF medium-range ensemble forecasts of precipitation for river basins. *Quarterly Journal of the Royal Meteorological Society*, *140*(682), 1615–1628. doi:10.1002/qj.2243

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

Young, P. C. (2002). Advances in real-time flood forecasting. *Philosophical Transactions of the Royal Society A*, *360*(1796), 1433–1450. doi:10.1098/rsta.2002.1008

Zak, S. K., & Beven, K. J. (1999). Equifinality, sensitivity and predictive uncertainty in the estimation of critical loads. *Science of the Total Environment*, *236*(1-3), 191–214. doi:10.1016/S0048-9697(99)00282-X

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., & Gailhard, J. (2012). Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Advances in Science and Research*, *8*, 135–141. doi:10.5194/asr-8-135-2012

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

Zappa, M., Rotach, M. W., Arpagaus, M., Dorninger, M., Hegg, C., Montani, A., … Wunram, C. (2008). MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems. *Atmospheric Science Letters*, *9*(2), 80–87. doi:10.1002/asl.183

Zhang, F., Zhang, H., Hagen, S. C., Ye, M., Wang, D., Gui, D., … Liu, J. (2014). Snow cover and runoff modelling in a high mountain catchment with scarce data: Effects of temperature and precipitation parameters. *Hydrological Processes*, *29*(1), 52–65. doi:10.1002/hyp.10125

Zhao, L., Duan, Q., Schaake, J., Ye, A., & Xia, J. (2011). A hydrologic post-processor for ensemble streamflow predictions. *Advances in Geosciences*, *29*, 51–59. doi:10.5194/adgeo-29-51-2011

Zheng, Y., & Keller, A. A. (2007). Uncertainty assessment in watershed-scale water quality modeling and management: 1. Framework and application of generalized likelihood uncertainty estimation (GLUE) approach. *Water Resources Research*, *43*(8). doi:10.1029/2006WR005345

Table of contents appendices

Appendix 1: HBV model description

The hydrological model that is used to simulate discharge is a HBV model, provided by IGF PAN in Matlab. The used HBV model matches to a large extent to the model that has been used by Knoben (2013). The description below is based on the provided HBV model in Matlab and the report of Knoben (2013). Figure 46 presents the structure of the applied HBV model. Below the figure the symbols of parameters, storages and fluxes are declared and the equations that are used in the HBV model are given. The model works with a daily time step and the output of the model is discharge in m^3/s, representing the discharge on a certain calendar day.

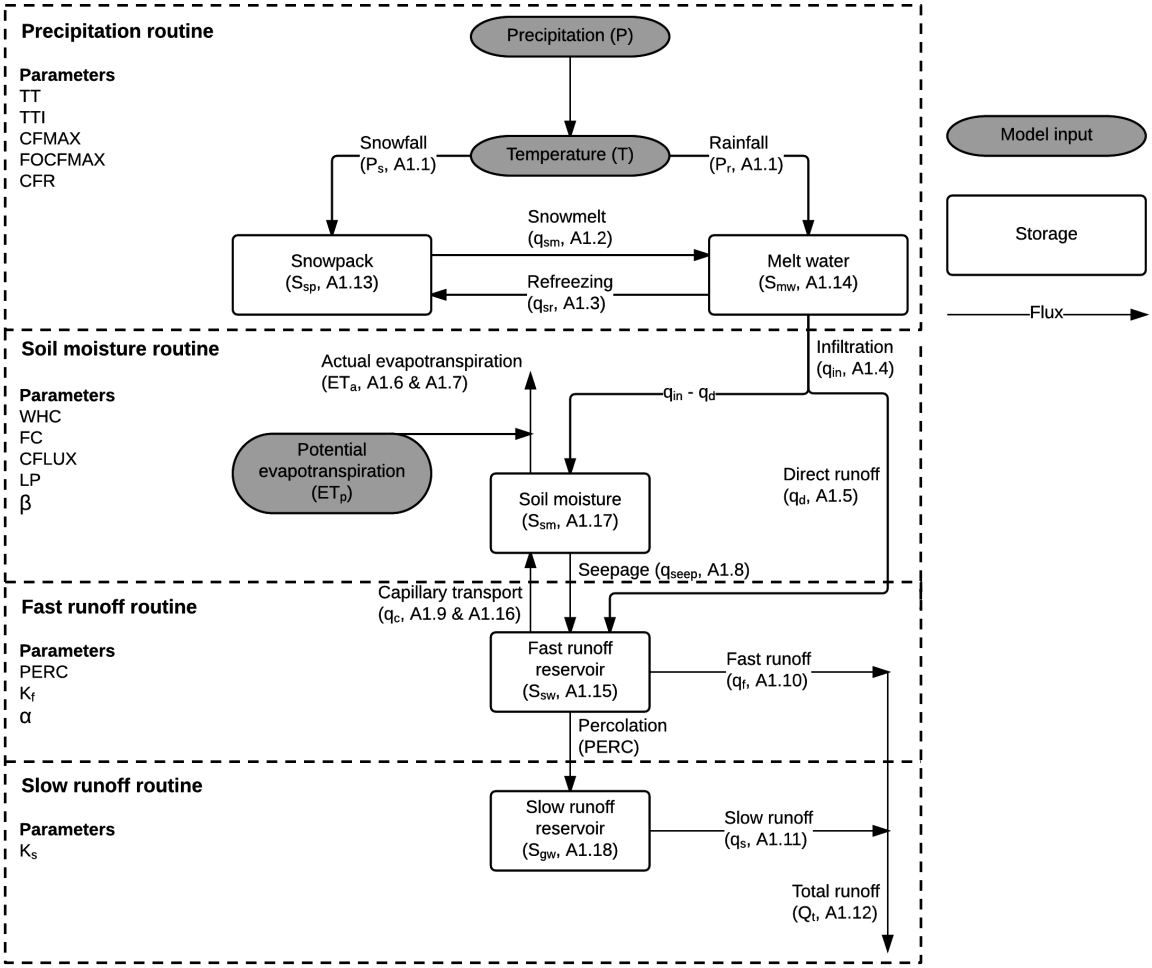


Figure 46: Structure of the applied HBV model. Numbers correspond to equation numbers. Figure is adopted from Knoben (2013).

Table 20: Parameter symbols in the HBV model. The parameter ranges (parameter minimum and maximum) are adopted from an earlier application of the HBV model to the same catchment by IGF PAN.

Parameter	Explanation (Knoben, 2013; Lindström et al., 1997)	Parameter minimum	Parameter maximum	Unit
FC	Field capacity, maximum soil moisture storage	0.10	1000	mm
β	Non-linearity parameter of seepage	0.01	7	-
LP	Limit for potential evapotranspiration	0.10	1	-
α	Non-linearity parameter of fast runoff	0.10	3	-
K_f	Fast runoff parameter	0.0005	0.3	d^{-1}
K_s	Slow runoff parameter	0.0005	0.3	d^{-1}
PERC	Rate of percolation	0.001	6	mm d^{-1}
CFLUX	Rate of unlimited capillary rise	0	4	mm d^{-1}
TT	Threshold temperature for which 50% of precipitation occurs as snow and 50% as rain	-3	4	^{o}C
TTI	Interval length between which precipitation occurs partly as snow and partly as rain	0	7	^{o}C
CFMAX	Snowmelt rate	0	20	mm $^{o}C^{-1}$ d^{-1}
FOCFMAX	*CFMAX* parameter corrected for forests	0	1	-
CFR	Refreezing factor	0	0.8	-
WHC	Water holding capacity of snow	0	1	mm mm^{-1}
FFO	Percentage of forest to correct snowmelt, snow refreezing and evapotranspiration.	Not calibrated: 11%		%
FFI	Percentage of non-forest.	Not calibrated: 89%		%
CEVPFO	Correction factor for evaporation in forest areas.	Not calibrated: 1.15		-

Table 21: Storage symbols in the HBV model

Storage	Unit	Explanation (Knoben, 2013; Lindström et al., 1997)
S_{sp}	mm	Snowpack
S_{mw}	mm	Melt water storage
S_{sm}	mm	Soil moisture storage
S_{sw}	mm	Fast runoff reservoir. This is also called surface water storage.
S_{gw}	mm	Slow runoff reservoir. This is also called groundwater storage.

Table 22: Flux symbols in the HBV model

Flux	Unit	Explanation (Knoben, 2013; Lindström et al., 1997)
P	mm/d	Precipitation
P_s	mm/d	Snowfall
P_r	mm/d	Rainfall
q_{sm}	mm/d	Snowmelt. Theoretical snowmelt flux is not limited by current snowpack.
q_{sr}	mm/d	Snow refreezing. Theoretical snow refreezing flux is not limited by current melt water storage.
q_{in}	mm/d	Total infiltration leaving the melt water storage
q_d	mm/d	Direct runoff to fast runoff reservoir
ET_{p0}	mm/d	Potential evapotranspiration, calculated by method of Hamon (see appendix 2)
ET_p	mm/d	Potential evapotranspiration corrected for different evapotranspiration rates from forests
ET_a	mm/d	Actual evapotranspiration
q_{seep}	mm/d	Seepage from soil moisture to fast runoff reservoir
q_c	mm/d	Capillary rise from fast runoff reservoir to soil moisture
PERC	mm/d	Percolation from fast runoff reservoir to slow runoff reservoir
q_f	mm/d	Outflow from fast runoff reservoir
q_s	mm/d	Outflow from slow runoff reservoir
Q_t	m^3/s	Total runoff from the catchment

A1.1 Precipitation and snow accumulation routine

Precipitation is an incoming flux into the hydrological system. Based on temperature it is determined whether precipitation occurs as rainfall, snowfall or partly as rainfall and partly as snowfall. With temperature equal to parameter *TT*, 50% of the precipitation occurs as rainfall and 50% as snowfall. The division of rainfall and snowfall is visualized in Figure 47 and described in equation A1.1.

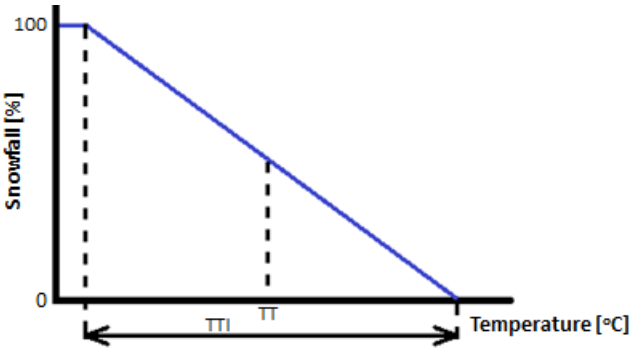


Figure 47: Division of precipitation in rainfall and snowfall (Knoben, 2013)

$$\begin{cases} P_s(t) = P(t) & T(t) < TT - \frac{TTI}{2} \rightarrow 100\%\ snow \\ P_r(t) = 0 & \\ P_s(t) = P(t) * \frac{\frac{TT + TTI}{2} - T}{TTI} & \\ P_r(t) = P(t) * \frac{T - \frac{TT - TTI}{2}}{TTI} & T \geq TT - \frac{TTI}{2}\ and\ T < TT + \frac{TTI}{2} \rightarrow partly\ snow, partly\ rain \\ P(t) = 0 & \\ P_r(t) = P(t) & T \geq TT + \frac{TTI}{2} \rightarrow 100\%\ rain \end{cases} \quad [A1.1]$$

Snow is directed into the snowpack storage and rainfall into the melt water storage. The interaction between these storages in the form of snowmelt and refreezing fluxes is described by equation A1.2 and equation A1.3.

$$q_{sm,theoretical}(t) = \begin{cases} CFMAX * (FFO * FOCFMAX + FFI) * (T(t) - TT) & T(t) > TT \\ 0 & T(t) \leq TT \end{cases} \quad [A1.2]$$

$$q_{sr,theoretical}(t) = \begin{cases} CFR * (FFO * FOCFMAX + FFI) * CFMAX * (TT - T) & T(t) < TT \\ 0 & T(t) \geq TT \end{cases} \quad [A1.3]$$

These are theoretical snowmelt and refreezing fluxes, because the real fluxes are limited by the actual storage in respectively the snowpack storage and snowmelt storage.

A1.2 Soil moisture routine

The soil moisture storage provides an estimation of the wetness of the catchment (Bergström, 1995). Precipitation seeps into the soil moisture routine via the infiltration flux. Infiltration is described by equation A1.4. Infiltration is based on available water in the snowmelt storage, but part of this water is temporarily retained in the snowpack. It is prevented that infiltration can become negative.

$$q_{in}(t) = \max\left(\left(S_{mw}(t) + q_{sm}(t) + P_r(t) - q_{sr}(t) - WHC * S_{sp}(t)\right), 0\right) \quad [A1.4]$$

By equation A1.4 the total infiltration is obtained. The part of the infiltration that would exceed the field capacity (parameter *FC*) forms direct runoff to the fast runoff reservoir and $q_{in} - q_d$ flows into the soil moisture storage.

$$q_d(t) = \max\left((q_{in}(t) + S_{sm}(t) - FC), 0\right) \quad [A1.5]$$

Outgoing fluxes from the soil moisture storage are actual evapotranspiration and seepage. To calculate the actual evapotranspiration at first potential evapotranspiration is adopted from the method of Hamon based on temperature input (see appendix 2). Evapotranspiration is higher from forests so at first potential evapotranspiration is corrected for the percentage of forest in the catchment (equation A1.6). The actual evapotranspiration is limited by the actual soil moisture storage (water that is available for evapotranspiration), *LP* and the potential evapotranspiration (equation A1.7).

$$ET_p(t) = ET_{p0}(t) * (FFO * CEVPFO + FFI) \qquad [A1.6]$$

$$ET_a(t) = \min\left(ET_p(t)\frac{S_{sm}(t)}{LP * FC}, ET_p(t)\right) \qquad [A1.7]$$

Seepage (equation A1.8) flows from the soil moisture storage into the fast runoff reservoir. Seepage cannot be below 0 (flux from fast runoff reservoir to soil moisture storage is described by capillary transport).

$$q_{seep}(t) = \max\left(\left(\left(\frac{S_{sm}(t)}{FC}\right)^{\beta} * \left(in(t) - q_d(t)\right)\right), 0\right) \qquad [A1.8]$$

Capillary rise is the flow from the fast runoff reservoir to the soil moisture storage (equation A1.9). Capillary rise is limited by the actual soil moisture storage.

$$q_c(t) = CFLUX * \frac{FC - S_{sm}(t)}{FC} \qquad [A1.9]$$

A1.3 Runoff generation routine (fast and slow runoff)

From the fast runoff reservoir and slow runoff reservoir respectively fast runoff (equation A1.10) and slow runoff (equation A1.11) are released. The fast runoff and slow runoff together make total runoff, i.e. discharge in the river in m^3/s (equation A1.12). Percolation from the fast runoff reservoir to the slow runoff reservoir is described by the calibrated parameter *PERC*.

$$q_f(t) = K_f * S_{sw}(t)^{1+\alpha} \qquad [A1.10]$$

$$q_s(t) = K_s * S_{gw}(t) \qquad [A1.11]$$

$$Q_t(t) = \frac{\left(q_f(t) + q_s(t)\right) * A}{\frac{86400}{10^{-3}}} \qquad [A1.12]$$

All fluxes in the model are based on storages in the current time step. Precipitation will not have effect on discharge in the same time step, so there is always a lag time of at least 1 day. This corresponds to observations of precipitation and discharge, where in general a lag time of 1 to 3 days is observed (see section 2.2.3).

A1.4 Storages

The last step is to determine the storages for the next time step. This is based on water balances, so the storages for the next time step are the storages of the current time step plus and minus the fluxes that go in and out of these storages.

A1.4.1 Snowpack

Incoming fluxes into the snowpack are snowfall and refreezing of melt water. There is one outgoing flux, namely snowmelt. This gives the water balance in equation A1.13. Snowpack storage cannot become below 0, because snowmelt is already limited by $S_{sp}(t)$.

$$S_{sp}(t+1) = S_{sp}(t) + P_s(t) + q_{sr}(t) - q_{sm}(t) \qquad [A1.13]$$

A1.4.2 Melt water storage

Incoming fluxes into the melt water storage are rainfall and snowmelt, while outgoing fluxes are refreezing of melt water and outflow from the melt water storage as infiltration (equation A1.14). Melt water storage cannot become below 0, because snow refreezing and infiltration are limited by the actual melt water storage.

$$S_{mw}(t+1) = S_{mw}(t) + P_r(t) + q_{sm}(t) - q_{sr}(t) - q_{in}(t) \qquad [A1.14]$$

A1.4.3 Fast runoff reservoir

The water balance to calculate the new fast runoff reservoir storage is given in equation A1.15. The new fast runoff reservoir is calculated before the new soil moisture storage, because negative fast runoff reservoir storage is corrected with capillary rise (equation A1.16). Actual percolation is limited to the amount of water that flows into the fast runoff reservoir (direct runoff + seepage).

$$S_{sw}(t+1) = \max\left(S_{sw}(t) + max\left(q_d(t) + q_{seep}(t) - PERC\right), 0\right) - q_f(t) - \min\left(S_{sw}(t), q_c(t)\right), 0) \qquad [A1.15]$$

$$q_c(t) = \begin{cases} \max\left(S_{sw}(t) + max\left(\left(q_d(t) + q_{seep}(t) - PERC\right), 0\right) - q_f(t), 0\right) & S_{sw}(t+1) = 0 \\ \min(\mathrm{S_{sw}}(t), q_c(t)) & S_{sw}(t+1) \neq 0 \end{cases} \qquad [A1.16]$$

A1.4.4 Soil moisture storage

Now the capillary rise is known the new storage in the soil moisture reservoir can be calculated (equation A1.17). Soil moisture storage cannot become below 0 or higher than the field capacity (*FC*), because all fluxes are limited to these lower and upper limits.

$$S_{sm}(t+1) = S_{sm}(t) + q_{in}(t) - q_d(t) - q_{seep}(t) + q_c(t) - ET_a(t) \qquad [A1.17]$$

A1.4.5 Slow runoff reservoir

The last storage that needs to be updated is the slow runoff reservoir, defined in equation A1.18. This storage cannot become below 0 because slow runoff is limited by the current slow runoff reservoir storage.

$$S_{gw}(t+1) = S_{gw}(t) + \min\left(q_d(t) + q_{seep}(t), PERC\right) - q_s(t) \qquad [A1.18]$$

Appendix 2: Potential evapotranspiration by method of Hamon

Potential evapotranspiration is one of the required inputs to the HBV model. Potential evapotranspiration is defined as the amount of water that could evaporate and transpire from a vegetated landscape without restrictions other than the atmospheric demand (Lu et al., 2005). With the method of Hamon temperature measurements or temperature forecasts are used to calculate potential evapotranspiration. In Table 23 the inputs are explained, coefficient values are provided in Table 24 and in Table 25 other variables are characterized.

Table 23: Inputs to the method of Hamon

Input variable	(Value) [Unit]	Explanation
T	^{o}C	Daily mean air temperature based on temperature observations or forecasts
J	-	Julian day of the year. This is used for day length. In a leap year 29 February has been given the same Julian day as 28 February.
φ	49.75^{o}	Representative latitude of the catchment

Table 24: Coefficients in the method of Hamon

Coefficient	Value	Explanation
a	6.108	Same value as used by Lu et al. (2005) and Rao et al. (2011)
b	17.26939	Same value as used by Lu et al. (2005) and Rao et al. (2011)
c	237.3	Same value as used by Lu et al. (2005) and Rao et al. (2011)
k_{pec}	1.2	Calibration coefficient. Same value as used by Lu et al. (2005).

Table 25: Variables in the method of Hamon

Variable	Unit	Explanation (Lu et al., 2005; Rao et al., 2011)
d	rad	Solar declination
Ω_s	rad	Sunset solar angle
ω	hrs	Daytime length
e_s	mbar	Saturated vapour pressure at a given temperature
ρ_{sat}	g/m^3	Saturated vapour density at a given temperature
ET_{p0}	mm/day	Potential evapotranspiration

Below the equations of the method of Hamon are given, adopted from a Matlab script provided by IGF PAN and also described by Lu et al. (2005) and Rao et al. (2011). The calculation of potential evapotranspiration is based on air temperature and day length (Prudhomme & Williamson, 2013; Rao et al., 2011). The equations to calculate the day length are also provided by Prudhomme and Williamson (2013). At first the solar declination is determined in equation A2.1 and the sunset solar angle is determined in equation A2.2.

$$d = 0.4093 * \sin\left(\frac{2 * J * \pi}{365} - 1.405\right) \qquad [A2.1]$$

$$\Omega_s = \operatorname{acos}\left(-tan\left(\frac{2 * \pi * \varphi}{360}\right) * tan(d)\right) \qquad [A2.2]$$

With the sunset solar angle the number of hours per day can be determined, by equation A2.3.

$$\omega = \frac{24 * \Omega_s}{\pi} \qquad [A2.3]$$

Temperature is used to calculate the saturated vapour pressure (equation A2.4) and saturated vapour density (equation A2.5).

$$e_s = a * \exp\left(b * \frac{T}{T + c}\right) \qquad [A2.4]$$

$$\rho_{sat} = \frac{216.7 * e_s}{T + 273.3} \qquad [A2.5]$$

Now the potential evapotranspiration can be calculated, based on daytime length and saturated vapour density (equation A2.6). The potential evapotranspiration is set to 0 when the temperature is below 0. This is not done in all studies where the Hamon method is applied, but in some studies like Lu et al. (2005) this has been done.

$$ET_{p0} = \begin{cases} \frac{0.1651 * \omega * \rho_{sat} * k_{pec}}{12} & T > 0 \\ 0 & T \leq 0 \end{cases} \qquad [A2.6]$$

According to Rao et al. (2011) and other studies (references in Rao et al. (2011)) the uncorrected Hamon method largely underestimates potential evapotranspiration in forests. A solution could be to apply a correction factor for the potential evapotranspiration from forests (Rao et al., 2011). It appeared that correction factors fluctuate largely between different forest types and during the year (Rao et al., 2011). The maximum and minimum correction factors that Rao et al. (2011) found are summarized in Table 26. In line with IGF PAN and Lu et al. (2005) a calibration coefficient of 1.2 (k_{pec}) is used to increase the potential evapotranspiration over the whole catchment. In addition a correction factor for evaporation from forests of 1.15 is used, adopted from an earlier application of the method of Hamon to this catchment by IGF PAN (see equation A1.6). Looking at Table 26 this seems a small correction.

Table 26: Required correction factors of potential evapotranspiration calculated with the method of Hamon for forests (Rao et al., 2011)

	Conifer forest	Deciduous forest
Maximum correction factor	2.6 (in March)	1.9 (in March)
Minimum correction factor	1.5 (in August)	1.0 (in December)
Annual correction factor	1.9 ± 0.5	1.3 ± 0.5

Actual evapotranspiration is also limited by water availability in the catchment. This is incorporated in the HBV model (equation A1.7).

Appendix 3: Contour plots of *k* as a relation of discharge and storage

In this appendix the contour plots of the fraction of fast runoff (*k*) as a relation of both simulated discharge and initial surface water storage are presented. For the lowest category of flows (Figure 48) and the medium category of flows (Figure 49) it can be seen that for the same discharge and increasing initial surface water storage the fraction of fast runoff increases. The other way around, with the same initial storage but increasing discharge the fraction of fast runoff decreases. This is because there is also correlation between discharge and surface water storage and usually with a higher discharge there is also higher surface water storage. For the lowest category of flows the maximum difference between the plotted function and the points is 0.04 and for the medium category this is 0.02. The relationship in Figure 50 for high flows is more difficult to see, because of the large domain of discharge and initial storage. The relationship looks like in Figure 49, but with less spread in *k*. For larger discharges there is even less spread, but this is also because there are only a few points here.

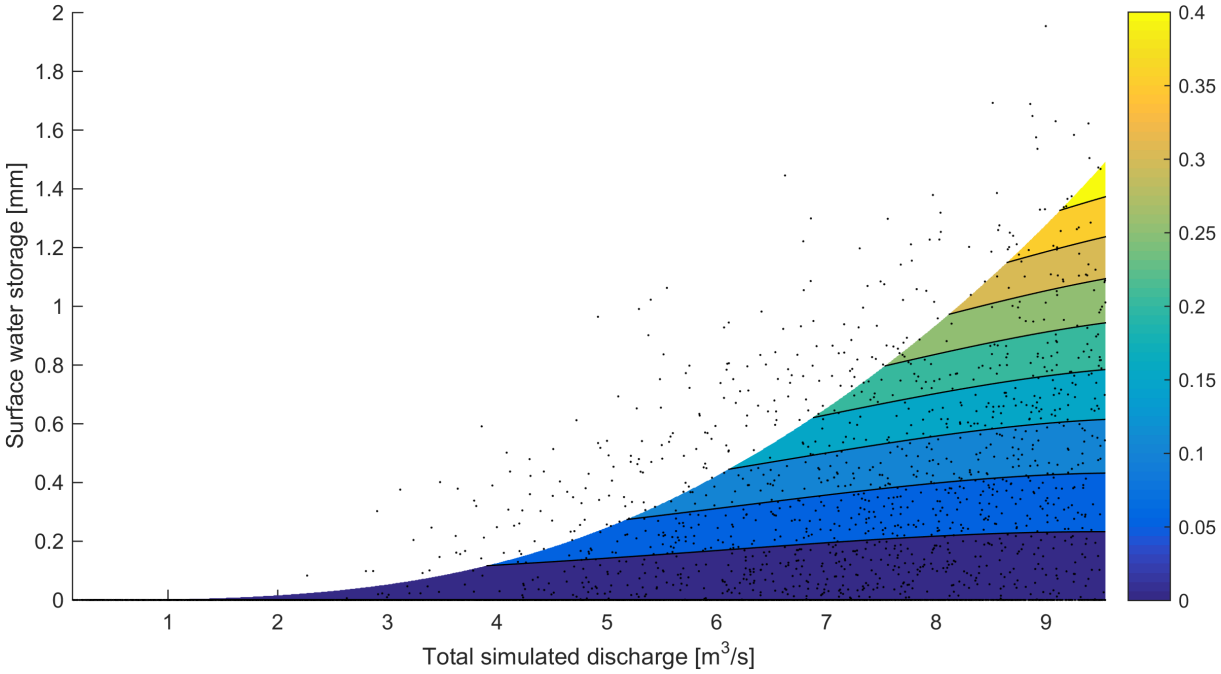


Figure 48: Contour plot of *k* (fraction of fast runoff) as a function of simulated discharge and surface water storage for the lowest category of discharge ($Q_0 \leq Q_{25}$). The contours are bounded by the 5% and 95% confidence lines as explained in section 3.4.3

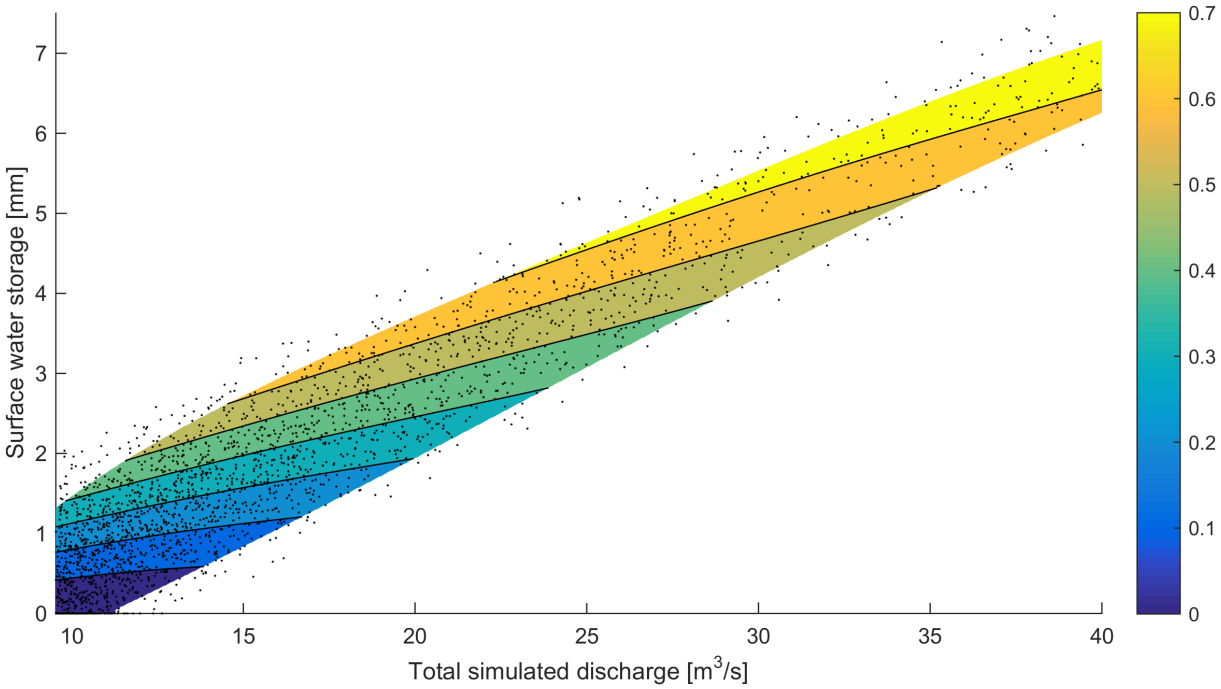


Figure 49: Contour plot of *k* (fraction of fast runoff) as a function of simulated discharge and surface water storage for the medium category of discharge ($Q_{25} < Q_0 \leq 40$ m^3/s). The contours are bounded by the 5% and 95% confidence lines as explained in section 3.4.3

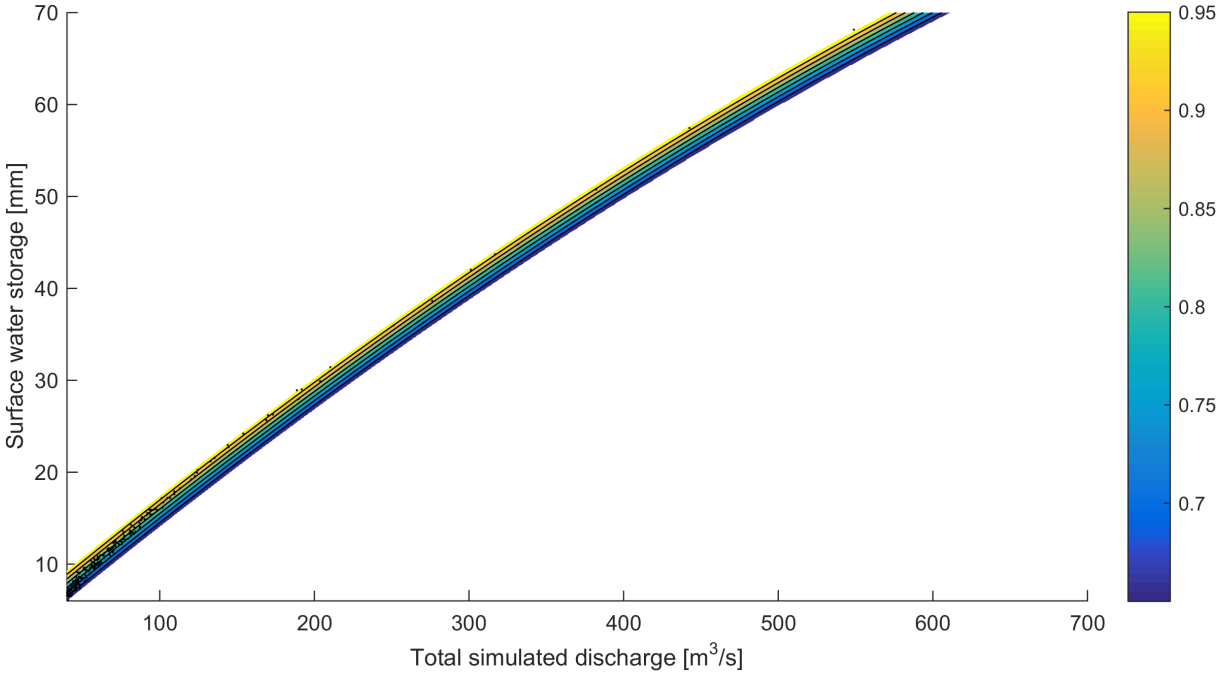
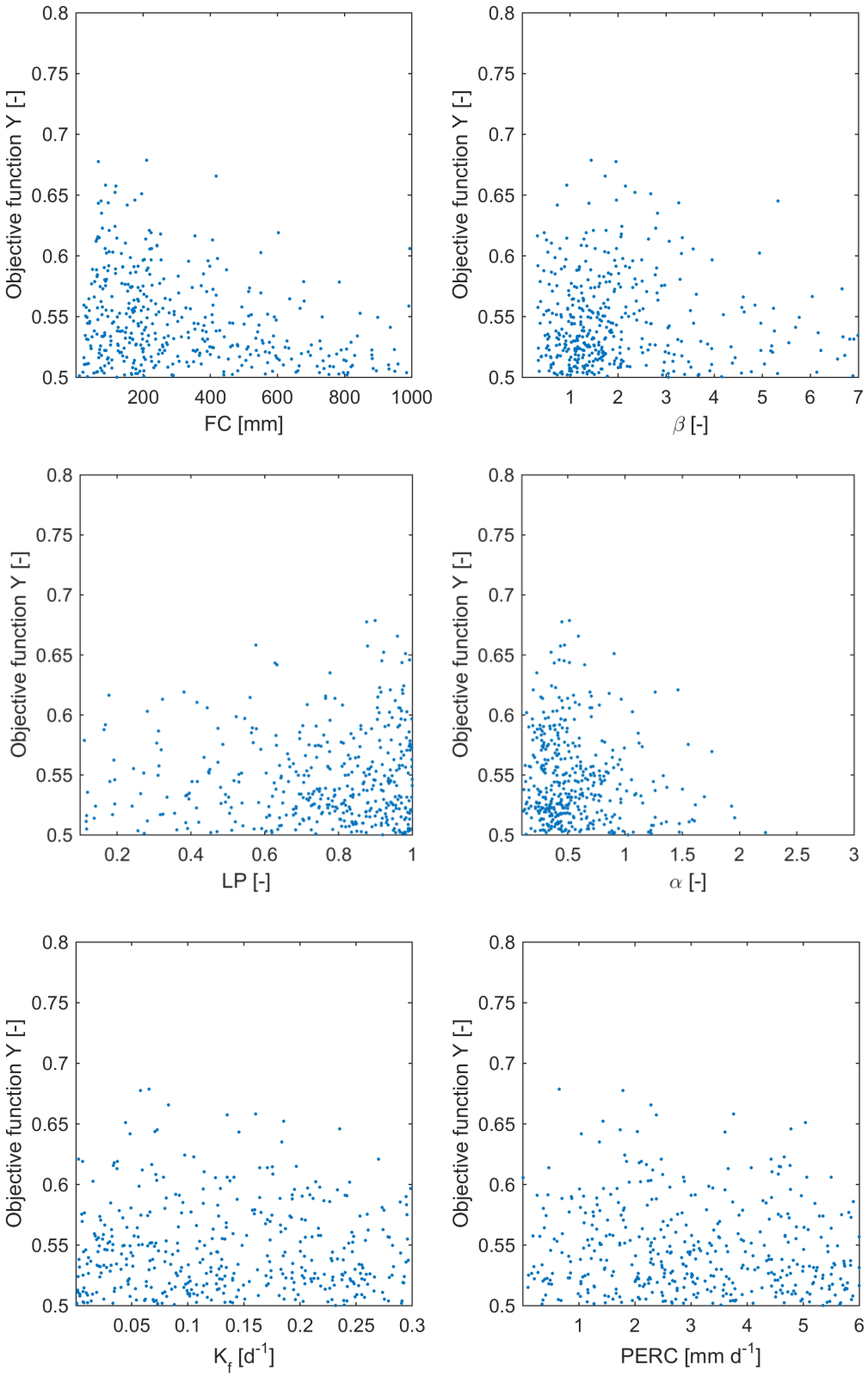


Figure 50: Contour plot of *k* (fraction of fast runoff) as a function of simulated discharge and surface water storage for the highest category of discharge (40 m^3/s < Q_0). For this flow spectrum it was not possible to establish 5% and 95% confidence lines like in Figure 48 and Figure 49 (see section 3.4.3), so *k* values are bounded by 0.65 and 1.

Appendix 4: Parameter uncertainty results

In this appendix the results of the Monte Carlo simulations (described in section 3.3.5) are presented. In Figure 51 the scatterplots of behavioural parameter sets are given and in Figure 52 the distributions of behavioural parameter sets over the parameter range are presented.



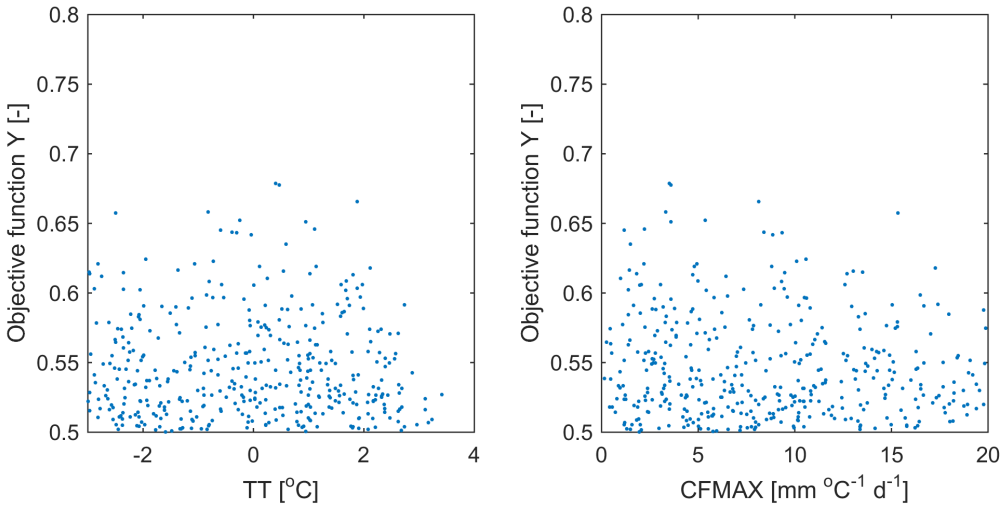


Figure 51: Monte Carlo simulations of behavioural parameter sets (*Y* > 0.5) of the 8 most sensitive parameters against objective function *Y*

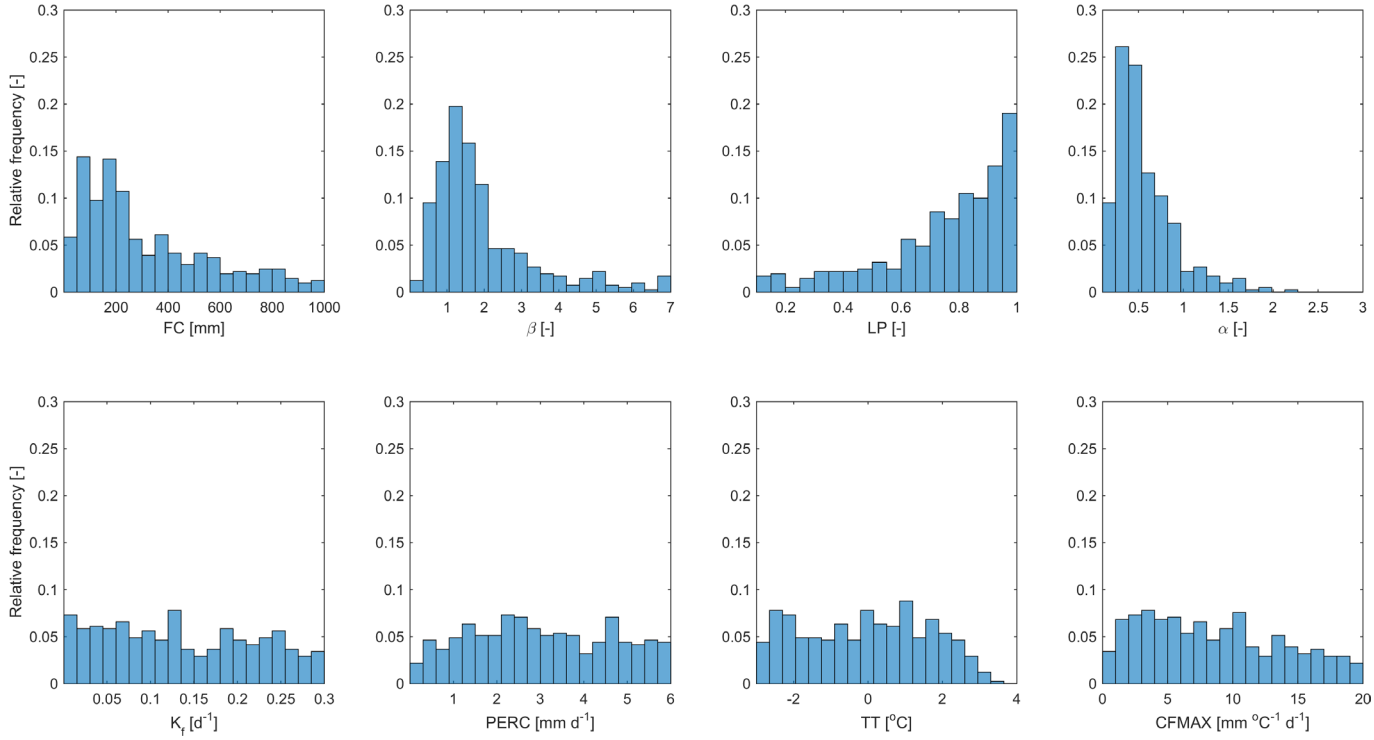


Figure 52: Frequency histograms of behavioural parameter sets from GLUE

Appendix 5: Rank histograms of precipitation and temperature

In this appendix the rank histograms of precipitation and temperature forecasts before and after pre-processing over the validation period 2008-2011 are presented.

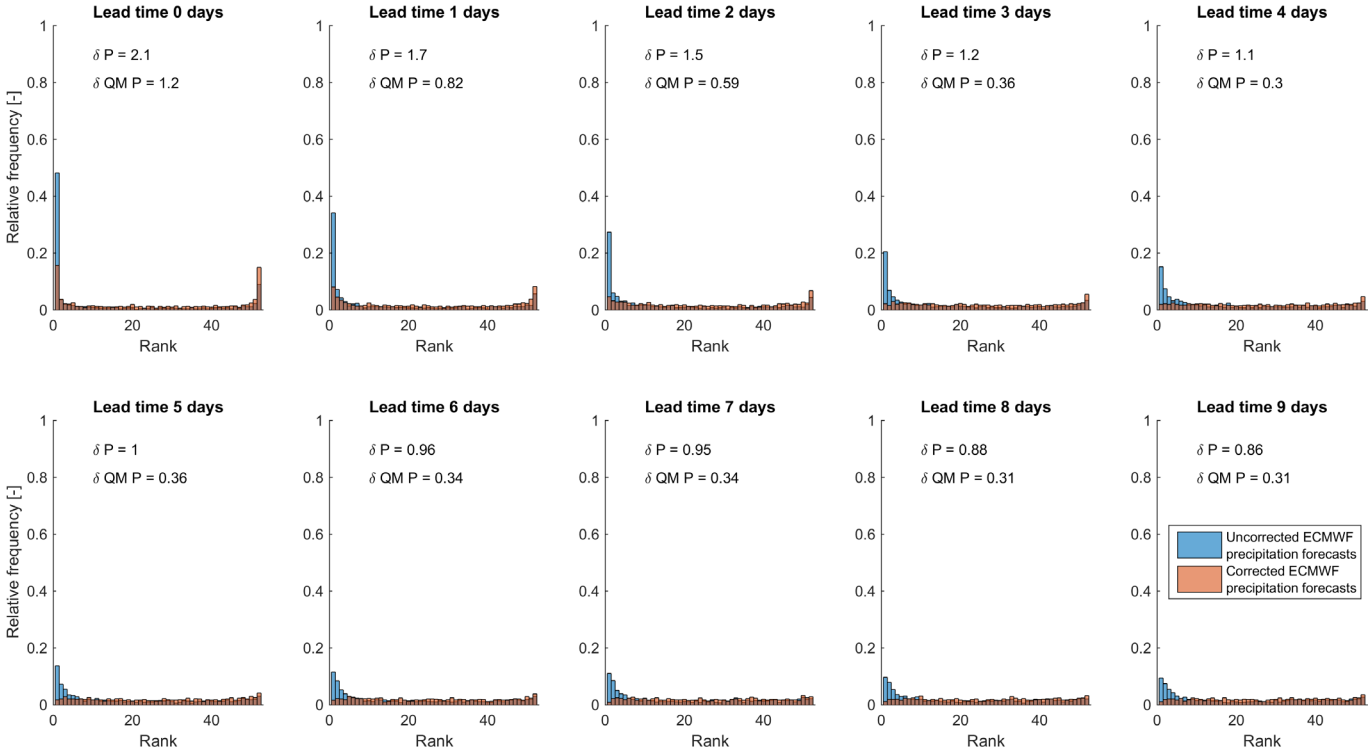


Figure 53: Rank histograms of uncorrected and corrected precipitation forecasts with QM with separate lead times, over the validation period 2008-2011

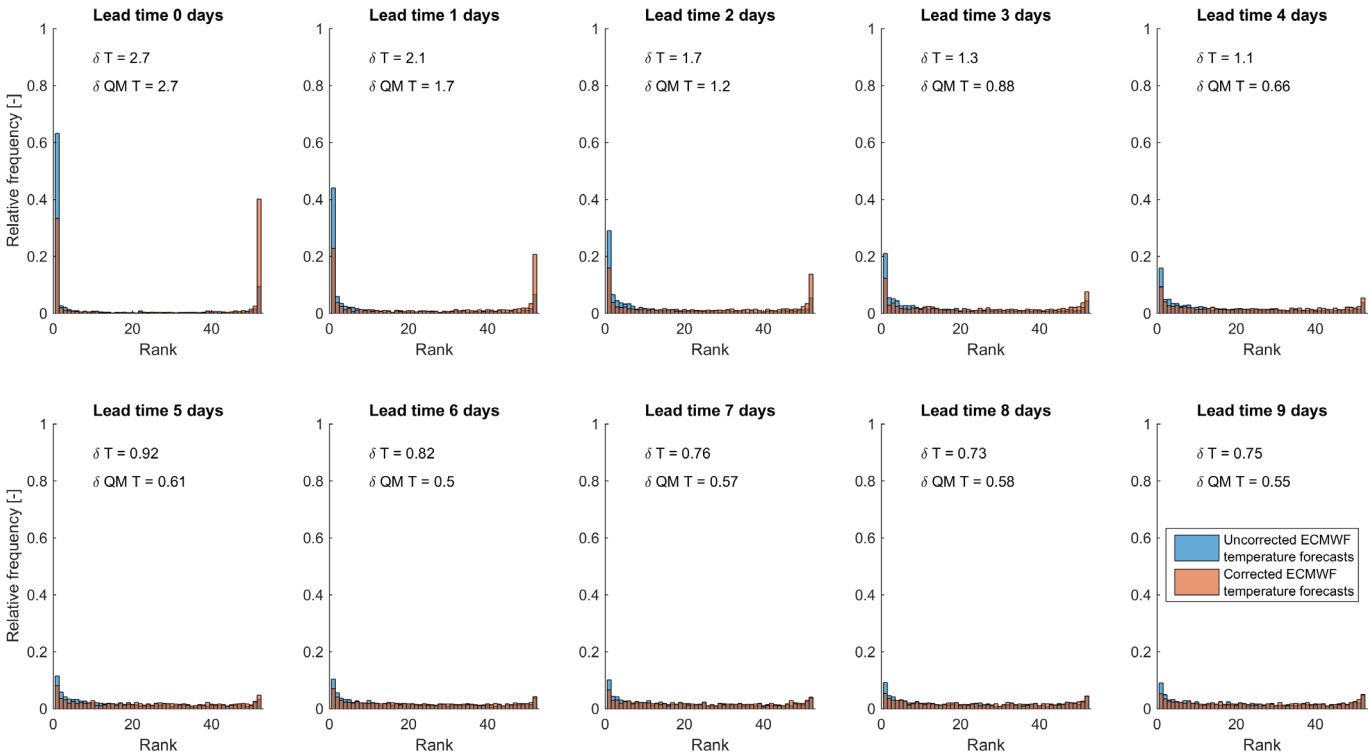


Figure 54: Rank histograms of uncorrected and corrected temperature forecasts with QM with separate lead times and two seasons (summer and winter), over the validation period 2008-2011

Appendix 6: Evaluation results post-processing strategies over the training period

This appendix presents the evaluation results of the post-processing strategies over the training period. This indicates the potential of processing with QM when a consistent bias would be present. Over the training period strategy 3 with seasonal distinction gives the best performance.

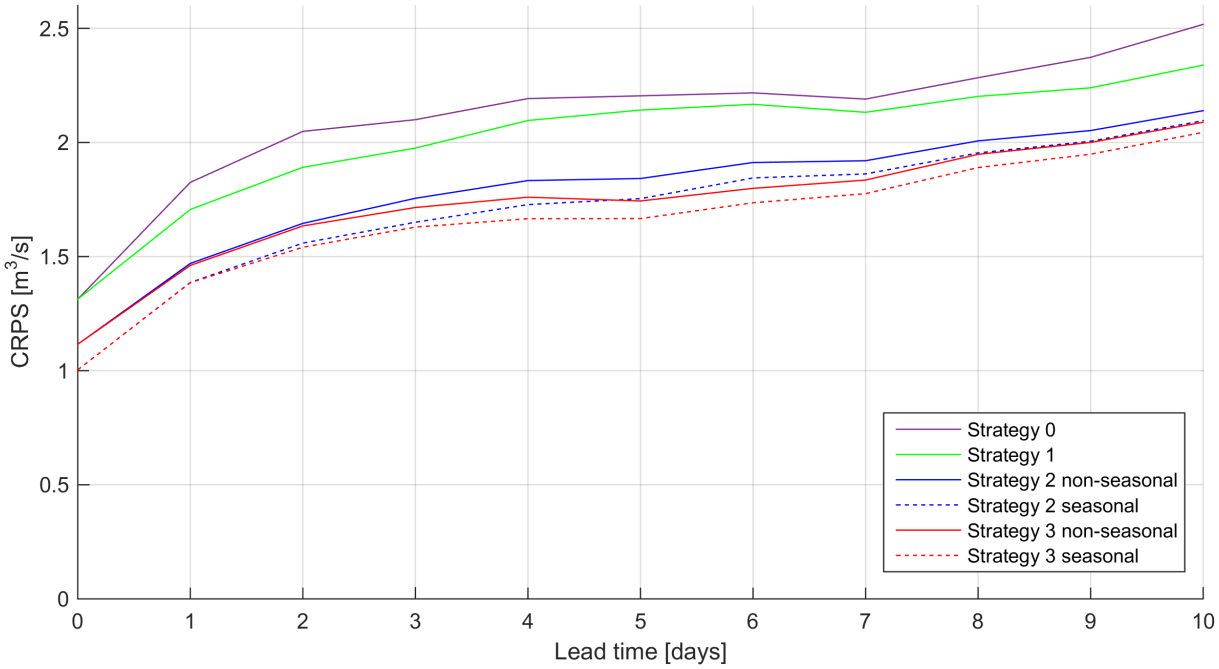


Figure 55: *CRPS* of the post-processing strategies, over the training period 2012-2013

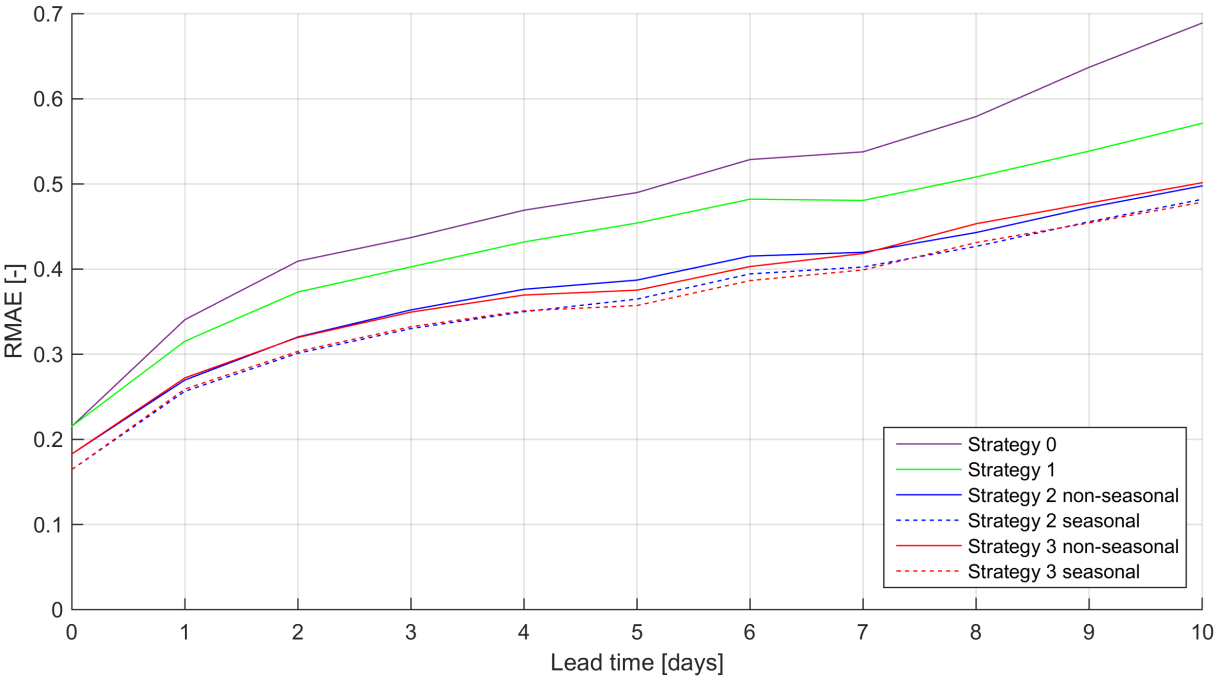


Figure 56: *RMAE* of the post-processing strategies, over the training period 2012-2013

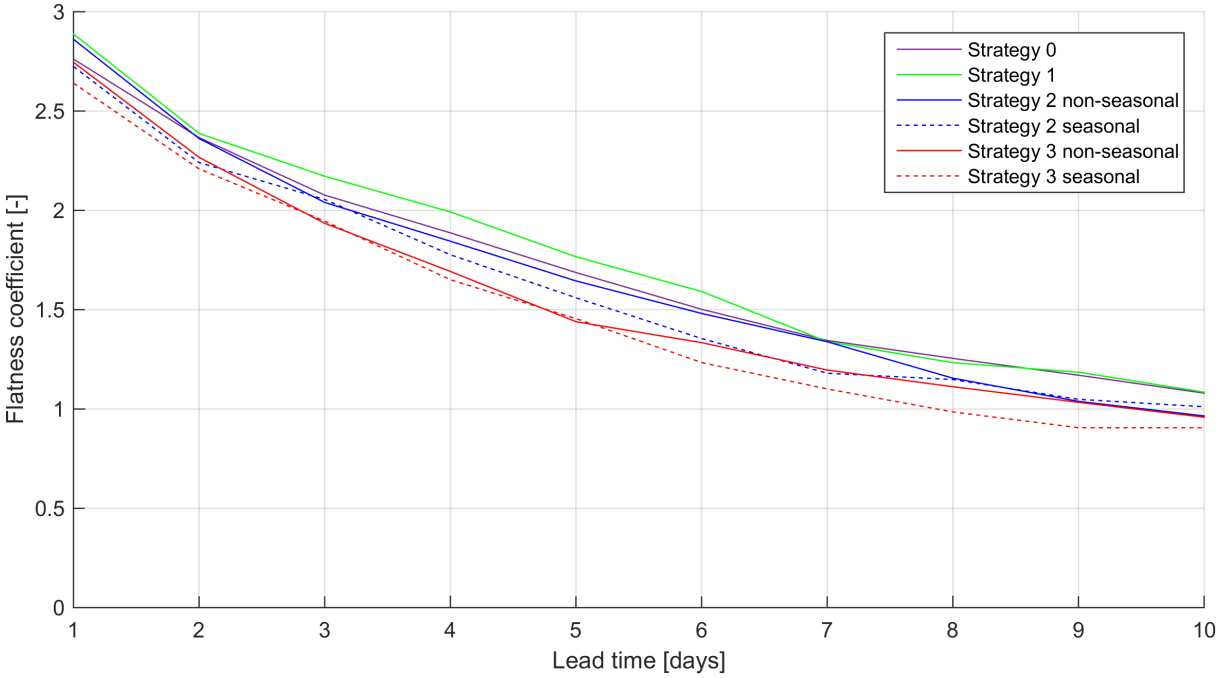


Figure 57: Rank histogram flatness coefficients of the post-processing strategies, over the training period 2012-2013

Appendix 7: Rank histograms of flow forecasts

This appendix presents the rank histograms of all flow forecasts (Figure 58) and the rank histograms for low, medium and high flow forecasts (Figure 59), established over the evaluation period 2008-2013.

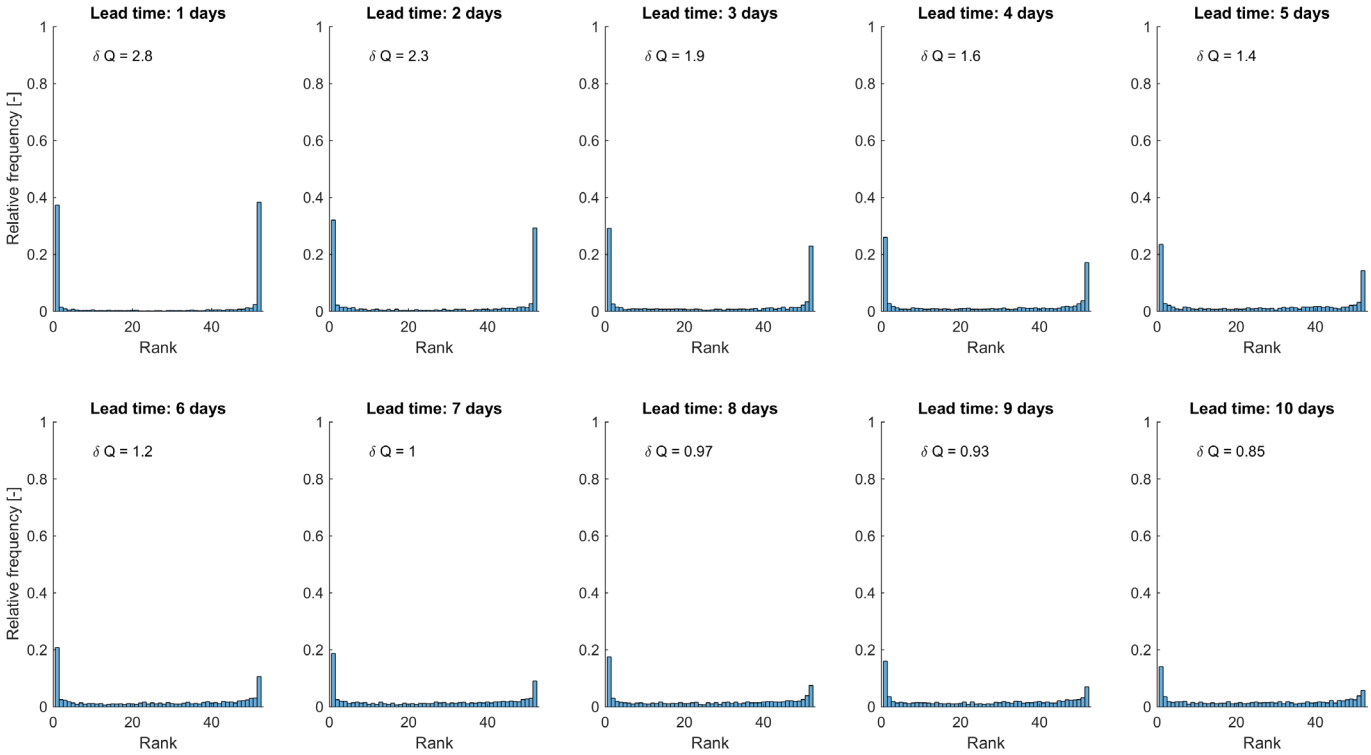


Figure 58: Rank histograms of the flow forecasts

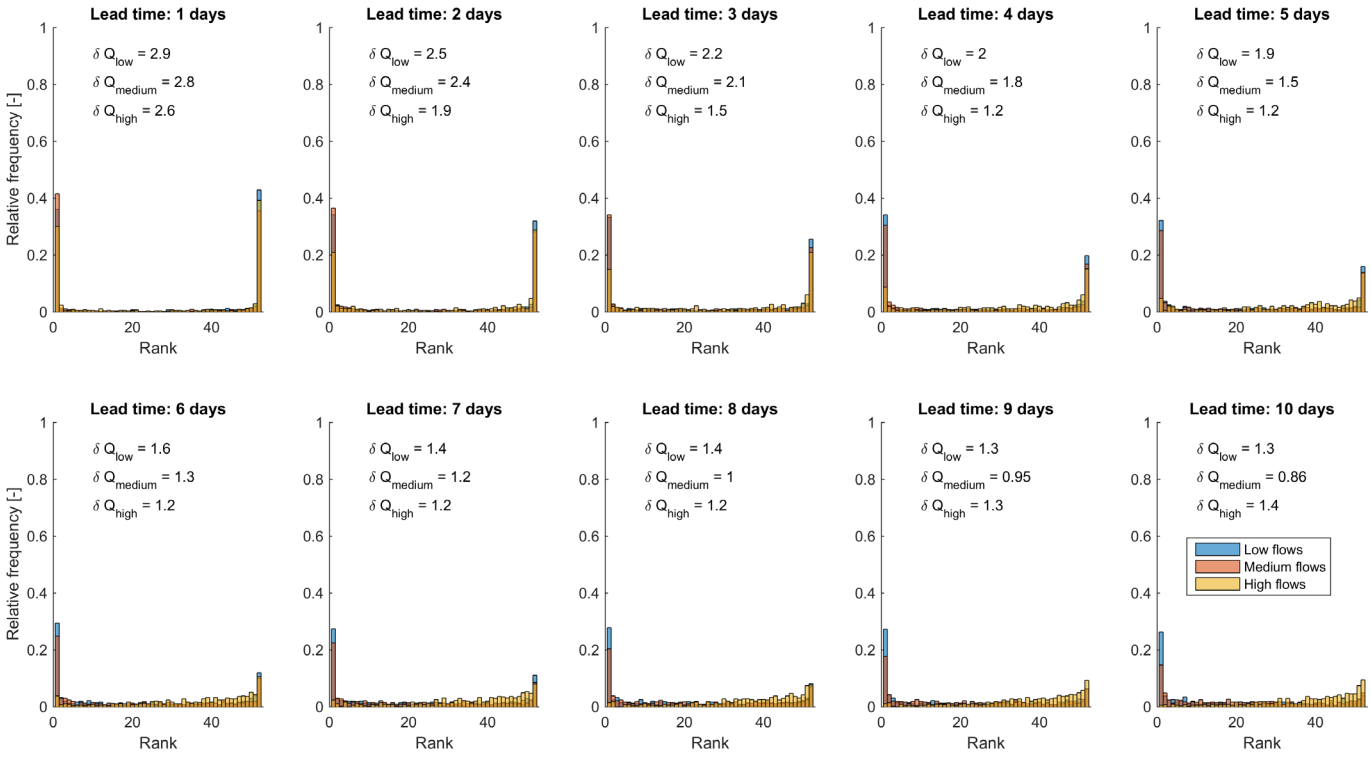


Figure 59: Rank histograms of the different flow forecast categories

Appendix 8: Relative Operating Characteristic curves

This appendix presents the ROC curves of the low flow forecasts and high flow forecasts, established over the evaluation period 2008-2013.

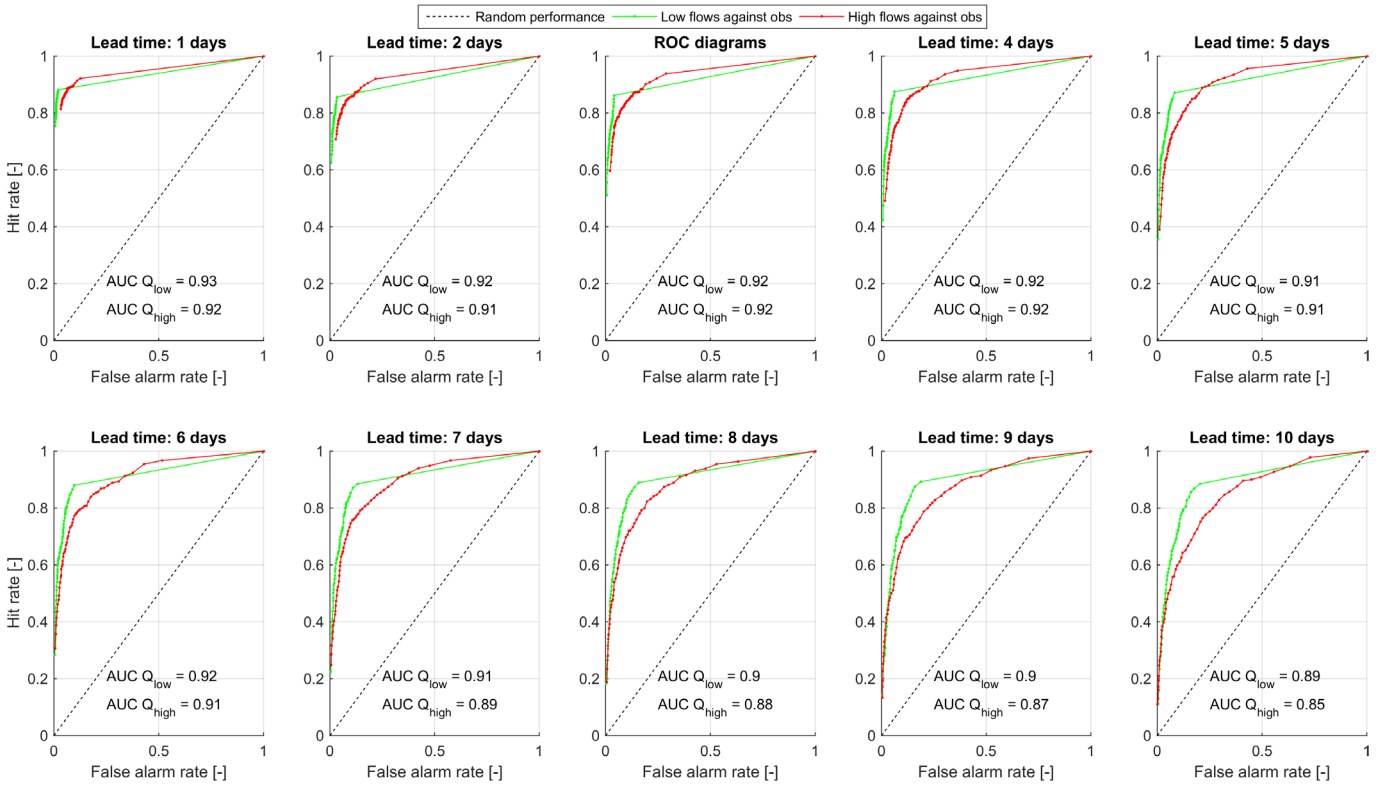


Figure 60: ROC curves and *AUC*s for low and high flow forecasts for lead times from 1 to 10 days

Appendix 9: Flow forecasts with parameter uncertainty

Only uncertainty in the meteorological input data has been included in the developed ensemble flow forecasting system. Hydrological model parameter uncertainty, initial condition uncertainty and model structure uncertainty are not included, which has been motivated in section 1.6. However the rank histograms (section 4.3.2.1) show a large under-dispersion of the flow forecasts, especially at small lead times. This indicates that the forecasts are over-confident. This could be expected, because not all sources of uncertainty has been included in the forecasting system (Bennett et al., 2014; Pagano et al., 2013). Precipitation forecasts are the dominant source of uncertainty at longer lead times and hydrological processes are the dominant source of uncertainty at shorter lead times (Bennett et al., 2014). Reliability of the flow forecasts might be improved by also incorporating hydrological model parameter uncertainty and initial condition uncertainty, which will be demonstrated with a simple approach in this appendix. Model structure uncertainty will not be included, because additional hydrological models should be introduced to do this. In section A9.1 it is explained how model parameter uncertainty is included in the flow forecasts and in section A9.2 the general performance, reliability and relative uncertainty of the flow forecasts will be investigated when meteorological input uncertainty and model parameter uncertainty are included.

A9.1 Include model parameter uncertainty with GLUE

This section explains how hydrological model parameter uncertainty is included. In section 4.1.3 the parameter uncertainty of the 8 most sensitive parameters has been described with a GLUE analysis. 0.82% of the 50000 tested parameter sets has been assigned behavioural and these parameter sets will be used to incorporate model parameter uncertainty. To calculate the likelihood of the behavioural parameter sets they are weighted linearly between the threshold for behavioural model parameter selection (Y = 0.5) and the maximum Y that has been found in the Monte Carlo analysis (Y = 0.68). Next to that the likelihoods are scaled, so that the sum of the likelihoods is 1 (Beven, 1993). A run of the ensemble model over the period from 1-11-2006 to 31-10-2013 with 1 parameter set and 51 meteorological ensembles takes about 6 minutes. To limit the computational time not all behavioural parameter sets are used, but they are sampled (with replacement) according to their likelihood. It has been chosen to select the same number of parameter sets as the number of meteorological ensembles, so 50 sampled parameter sets and the optimum calibrated parameter set are used. Instead of the 51 ensembles originating from the meteorological ensemble forecasts, now 2601 ensembles are generated.

Initial condition updating is done with the sampled parameter set, so the updated initial conditions are also different with different parameter sets. This follows the approach of Demirel et al. (2013a) to include initial condition uncertainty. However, it should be researched whether this is a valid approach, because the relationship between the fraction of fast runoff and observed discharge has been based on the optimum parameter set.

A9.2 Evaluation of flow forecasts including parameter uncertainty

Figure 61 presents the skill of the forecasts that include parameter uncertainty relative to the original flow forecasts, in terms of the *CRPS*. For all flows together the *CRPS* of the forecasts that include model parameter uncertainty is lower (better) for lead times until 3 days and for lead times of 4 days and more the forecasts that do not include parameter uncertainty are slightly better. This pattern is mainly caused by the high flow forecasts, although also for the low and medium flow forecasts the

skill decreases for increasing lead times. It has been expected that parameter uncertainty has the largest effect on the performance at small lead times. That the lowest skill values are found for the low flow forecasts can be explained by the fact that also the behavioural parameter sets from the Monte Carlo analysis are mainly calibrated on high flows. Apparently the behavioural parameter sets that should represent model parameter uncertainty perform even worse for low flow situations than the optimum parameter set. The improvement of reliability (see Figure 62) does not compensate for this.

In Figure 62 the rank histograms of the flow forecasts including parameter uncertainty are presented. The relative frequencies of the highest and lowest possible ranks are less than half of the original relative frequency, so the flatness of the rank histograms has improved by including model parameter uncertainty. However, the relative frequencies of the most extreme ranks are still much higher than the expected relative frequency in a system with perfect reliability, which indicates that the ensemble flow forecasts are still under-dispersive. The under-dispersion in the meteorological forecasts (see appendix 5) will also play a role in this. Figure 62 also shows the flatness coefficients of the flow forecasts that include parameter uncertainty. However these flatness coefficients cannot be compared to the original flatness coefficients, because the number of possible ranks differs.

In Figure 63 it can be seen that in general the relative uncertainty in the forecasts increases as a result of including hydrological model parameter uncertainty in the flow forecasts. This could be expected because with model parameter uncertainty an additional source of uncertainty is incorporated. In this case the *RCI* for a lead time of 0 days is not equal to 0, because the different model parameter sets already induce a spread in the ensemble flow forecasts at a lead time of 0 days. Especially at small lead times the *RCI* becomes larger. At larger lead times the difference between the *RCI* with and without including parameter uncertainty is less, because at larger lead times the uncertainty from the meteorological forecasts is dominant (Bennett et al., 2014).

It turned out that with another threshold for behavioural model parameter selection the results of the *CRPS* are very different. With a lower threshold for behavioural model parameter selection (Y = 0.3) the CRPS values of low and medium flow forecasts that include model parameter uncertainty are much better and the CRPS values of high flow forecasts are worse. Possibly with a lower threshold also parameter sets that perform well for low and medium flows are included. In the rank histograms the relative frequencies of the most extreme ranks become slightly lower (better) and the *RCI* increases. This makes sense, because with a lower threshold the range of model parameter uncertainty that is included is larger.

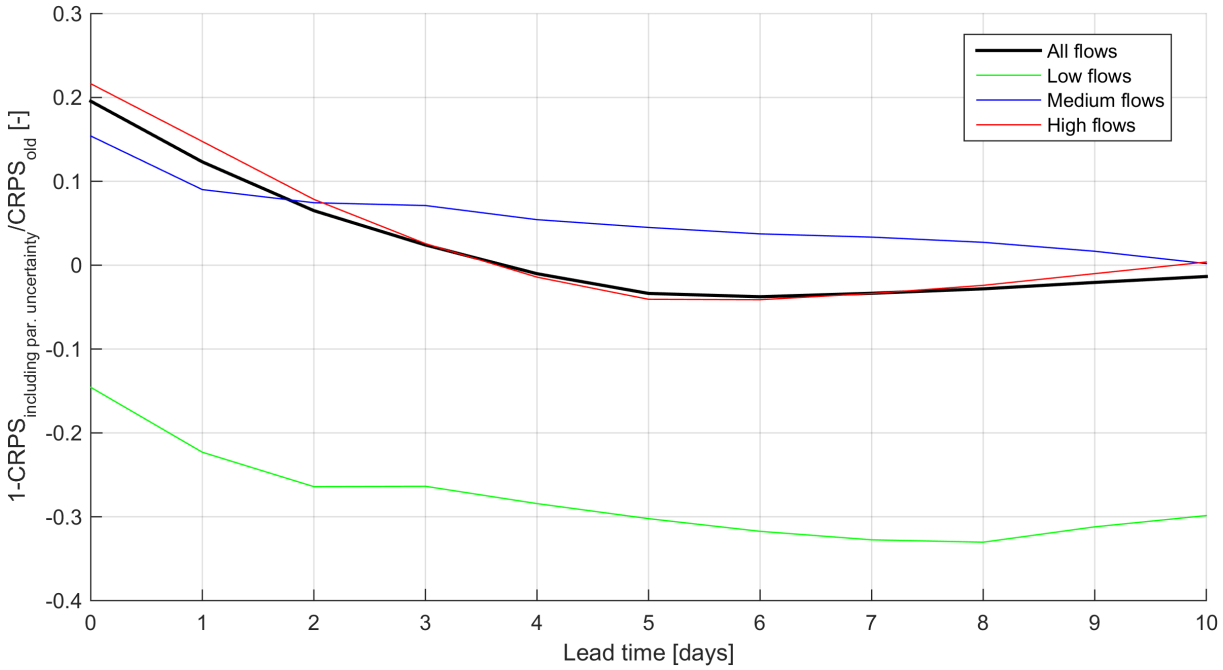


Figure 61: Skill of the flow forecasts including hydrological model parameter uncertainty relative to flow forecasts without hydrological model parameter uncertainty

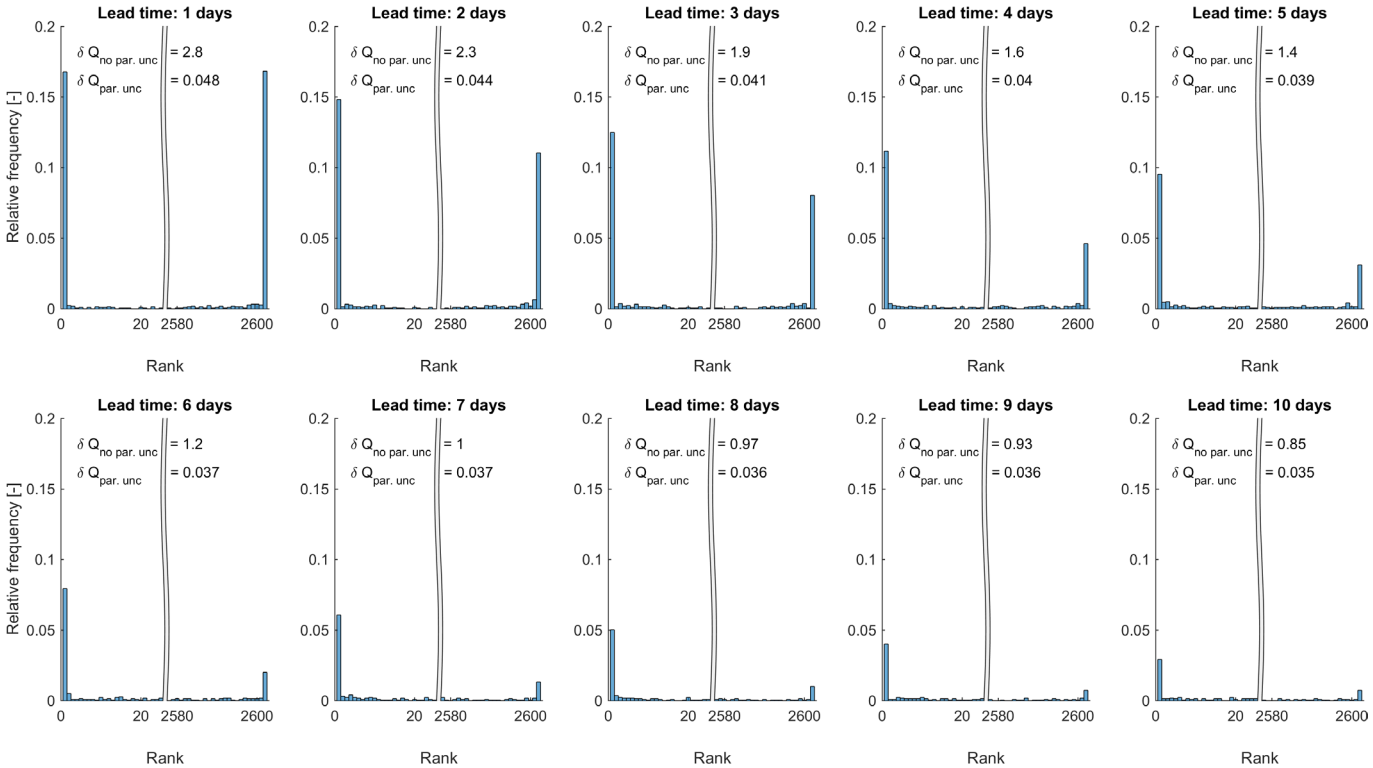


Figure 62: Rank histograms of flow forecasts with and without hydrological model parameter uncertainty

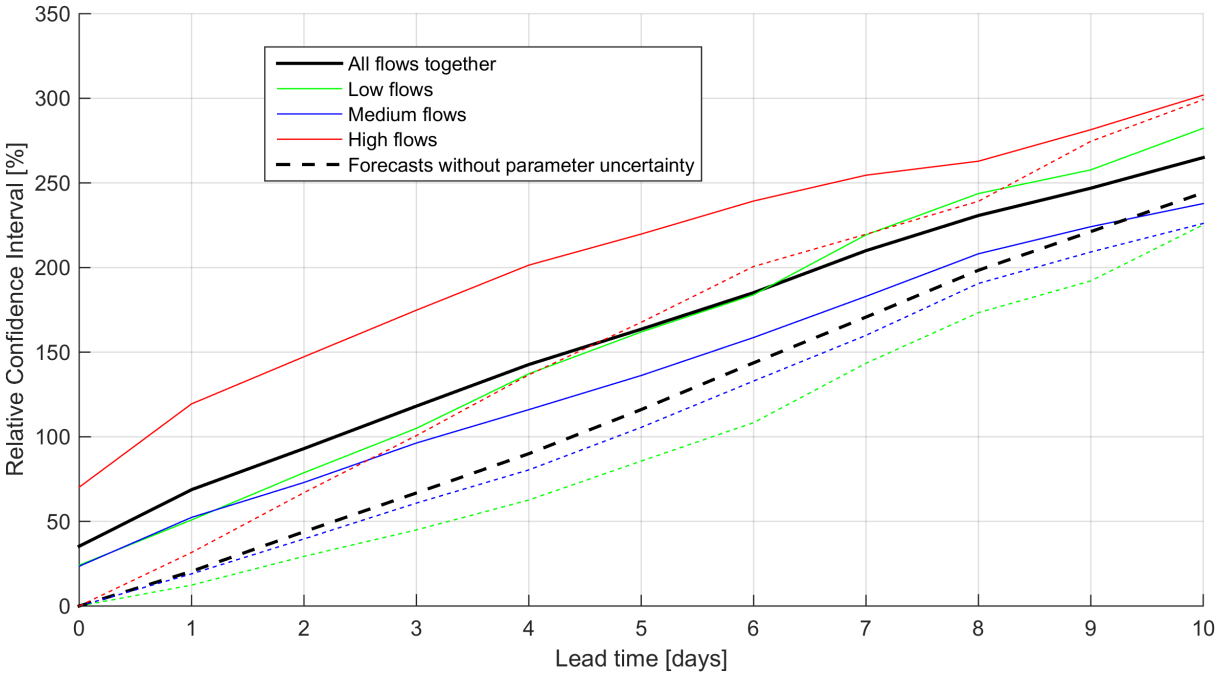


Figure 63: *RCI* of flow forecasts with and without hydrological model parameter uncertainty