Performance of multi-model ensemble combinations for flood forecasting

Master thesis

Loek Zomerdijk

Enschede, november 2015

UNIVERSITY OF TWENTE.

ZHEJIANG UNIVERSITY

Frontpage picture: Grand Canal in Hangzhou, China taken by Loek Zomerdijk

Performance of multi-model ensemble combinations for flood forecasting

Master thesis in Civil Engineering and Management

Loek Zomerdijk
Enschede, november 2015

Msc thesis committee:

Dr. ir. M.J. Booij
University of Twente, Department of Water Engineering and Management

Dr. M.S. Krol
University of Twente, Department of Water Engineering and Management

Dr. Y. Xu
Zhejiang University, Institute of Hydrology and Water Resources

UNIVERSITY OF TWENTE.

ZHEJIANG UNIVERSITY

Summary

Flooding is becoming a serious issue in recent decades due to urban expansion and climate change. As a consequence of floods international interest in flood forecasts has increased in the last decades. Accurate forecasting in small mountainous catchment areas is often difficult due to the short lead times of precipitation forecasts. More accurate forecasting can be obtained with the use of ensemble flood forecasts instead of deterministic forecasts. Recently research has been done on multi-model ensemble (grand ensemble) forecasts. In grand ensemble forecasts the ensembles of different EPSs are combined to improve the performance of the forecast in comparison with a single EPS. However, techniques to combine the different EPSs need to be developed. This study has the aim to develop an ensemble flood forecasting system for Quzhou (East-China) for lead times of 1 to 10 days and to evaluate different combined Grand Ensemble flood forecasts.

The lumped hydrological GR4J model is used to forecast flow with ensemble precipitation forecasts of 4 different weather centres (European Centre for Medium-Range Weather Forecasts (ECMWF); Chinese Meteorological Administration (CMA); UK Met Office (UKMO) and US National Centers for Environmental Prediction (NCEP)) as input. The EPSs of these centres have different ensemble sizes and each consists of 1 control forecast from where the other perturbed ensemble members are derived. The ensemble forecasts are bias corrected with the Quantile Mapping method and that resulted in an improvement of the forecasts.

After bias correction the precipitation forecasts are used as input to the hydrological model. The GR4J model was already calibrated for the Quzhou river basin with the Nash Sutcliffe efficiency coefficient (NS). Since the NS is more sensitive to high flows the calibrated values from this previous study are used. To further improve the forecasts an updating procedure is used for the hydrological model that updates the initial conditions of the routing storage with discharge observations at one day before the forecast day. This resulted in an improvement of the NS value for all lead times especially for short lead times of 1-3 days.

The flood forecasts are evaluated on three important components of skill: reliability, resolution and sharpness. Six different grand ensemble flood forecasts are constructed after the evaluation of the single model forecasts. There are two simple combinations used. The first is a combination of the members where the EPSs are not weighted, as a consequence EPSs with more ensembles have more influence on the grand ensemble. The second is a combination of the models where the models are weighted so that their influence on the grand ensemble is equally. Other combinations in this study are constructed with the simple grand ensembles using weighted contributions based on skills of the evaluated EPSs.

As expected, evaluation of the flood forecasts show that skill decreases with lead time and with increasing exceedance threshold. Two recognizable components of the forecast error, the meteorological error and the hydrological model error both increase with lead time, with an increasing contribution of the meteorological error compared to the hydrological error with lead time. All forecasts have relatively good performance reliability, resolution and sharpness. In general the single model forecasts of ECMWF proves to be the most skilful model and CMA the least skilful model in this study for

the Quzhou catchment area and the precipitation and hydrological forecasts. For short lead times of 1-2 days NCEP is least skilful.

All evaluations of the grand ensemble hydrological forecasts show that they are beneficial. They show lower root mean squared errors (RMSE), continuous ranked probability scores (CRPS), reliability and resolution as compared to the single model EPSs. Also the sharpness is better than that of single model forecasts. The CRPS and RMSE graphs become smoother as a result of the different biases of the single forecasts that cancel out in the grand ensemble forecasts. Simple combination methods of the grand ensembles show similar skill as combinations of ensembles forecasts using weighted contributions based on skills. This is because EPSs with less skill than other EPSs still can add skill in a grand ensemble. A model with less skill might be able to add model structure errors that's missing in other EPSs with good skill and might have good performance on days when the other models show low performance. Generally it can be concluded that there is no significant difference between the different combination methods. Previous studies showed that increasing ensemble size leads to little improvement, however models with less members can be better than models with more members. Therefore it is best to use an approach where the models are weighted with the method of equal probability of selection so that the influence is not dependent on ensemble size.

Preface

This report is the final product of my master study Civil Engineering and Management, with the specialization in Water Engineering and Management, at the University of Twente. In this study I have developed a system to forecast river discharges and created grand ensemble forecasts and evaluated this system and the grand ensemble forecasts for high flows and different lead times. This study was partially done at the Zhejiang University in Hangzhou in China and partially at the University of Twente to finalize this thesis.

I would like to thank the students and the professors of the Hydrology and Water Resources department of the Zhejiang University for the warm welcome in China, for the help with my thesis if I had questions, for the lunch breaks and for the football matches we played. Special thanks are for Yue-Ping Xu, my supervisor at the Zhejiang University, for ideas on my research, help with my research and the feedback from you.

I would also like to thank my supervisors Martijn Booij and Maarten Krol at the University of Twente, who gave me very helpful advice and feedback to finalize this MSc thesis.

Finally I would like to thank the students at the graduation room, for the support during the last couple of months and my friends and family for the support during my study at the University of Twente.

Table of Contents

1. Introduction

1.1. Motivation

Flooding is becoming a serious issue in China and worldwide due to urban expansion and climate change (Du et al., 2010). In China, the urban expansion caused remarkable spatial stress to various wetlands, which therefore have been decreased in size resulting in more frequent flood hazards (He et al., 2011). Also the frequency of extreme rainfall events increases due to climate change which results in flooding. In addition, many urban areas are developing quickly with population and asset growth, which further increases the vulnerability of cities to floods (Yang et al., 2015). In recent years cities like Beijing, Hangzhou and Guangzhou have experienced large floods already. The Qiantang River basin, as the most important river basin of Zhejiang Province in East China, also has a large population and suffers from extreme weather (Tian et al., 2014). This study will therefore focus on the Quzhou river basin, which is part of the Qiantang river basin.

As a consequence of floods, the international interest in flood protection and awareness has been growing over the last decade together with the improvement of flood forecasts (Cloke & Pappenberger, 2009). Operational flood forecasting systems play a major role in preparation strategies for disastrous flood events by providing early warnings several days ahead. In this case emergency responders have preparation time to reduce the impact of flooding. Accurate forecasting of floods in the cities is often difficult due to the short lead times of precipitation forecasts. However, more accurate forecasting can be obtained with the use of ensemble flood forecasting (Demeritt et al., 2013). Ensemble Prediction Systems (EPSs) have two significant advantages over conventional deterministic forecasting techniques. First, EPSs have shown evidence of greater skill in medium term (with lead times of 3-10 days ahead) rainfall and flood forecasts. Second, EPSs can also provide quantitative probability forecasts (QPFs) for different future system states and estimates the inherent uncertainty (Demeritt et al., 2013). Therefore the most likely and the most extreme scenarios can be identified and presented to emergency responders to get better prepared and allow them to optimize risk management responses by balancing the losses against the costs of measures been taken to reduce the impact of flooding. Consequently ensemble flood forecasting is widely used in recent years.

1.2. State of the art on ensemble flood forecasting

Many flood forecasting systems rely on precipitation inputs, which initially come from observation networks (rain gauges) and radar (Cloke & Pappenberger, 2009). However, lead times are very short when using precipitation observations, especially in small and medium sized catchments where the catchment response times are short (Nester et al., 2012). Often more time is required for flood response actions. Hence one of the main challenges in flood forecasting and warning is to extend forecast lead times beyond the catchment response time. For medium term forecasts (~3-10 days ahead), Numerical Weather Prediction (NWP) models have to be used (Cloke & Pappenberger, 2009). Single deterministic weather forecasts from NWP models cannot take uncertainties and systematic biases into consideration and thus often fail to replicate weather variables correctly (Bao et al., 2011). Therefore flood forecasting systems around the world are recently increasingly moving towards using ensembles of NWPs known as EPSs instead of using single deterministic forecasts. An EPS is then usually used as input to a hydrological

and/or hydraulic model to produce river discharge predictions (Cloke & Pappenberger, 2009). Several different hydrological and flood forecasting centres now use EPSs and it is expected that many others will follow. Over the last 20 years EPSs have already often been used in weather forecasts. It is an attractive method, because with EPSs it is possible to make multiple weather predictions for the same location and time. This is a better method than a single deterministic forecast, because it is not possible to predict the exact state of the atmosphere and therefore the weather. Hence EPS weather forecasts are an attractive product for flood forecasting systems because it has the potential to extend lead times and better quantify the uncertainty.

The EPSs change and continue to improve, since EPS forecasting is relatively new the (Cloke & Pappenberger, 2009). These improvements are required for predictions from EPSs. However, the impact of these improvements on hydrological models is uncertain. A good strategy to improve the EPS forecasts is to use a 'grand ensemble', which means using several EPSs from different weather centres together. This is explained by the fact that EPS forecasts from a single weather centre only account for part of the uncertainties originating from initial conditions and the forecast model. When a grand ensemble of EPS from different weather centres combined is used also other sources of uncertainties, including numerical implementations and/or data assimilation, can be assessed (He et al., 2010), because different analyses, perturbation generation methods and forecast models are combined (Johnson & Swinbank, 2009). Bao et al., (2011) also state that the aggregation of various models producing EPSs from different weather centres results in a better retaining of and accounting for the probabilistic nature of the ensemble precipitation forecasts. Various studies applied the principle of equal probability of selection (Bao et al., 2011; He et al., 2010; Park et al., 2008). This means that every ensemble model has the same weight in the multi-model forecast. Further improvements might be made by giving the models different weights, because some models might be better than others (Johnson & Swinbank, 2009). Other studies have shown that model-dependent weights can give improvement, but that care should be taken in how the weights are calculated and used for the combination of the models (Raftery et al., 2005; Stefanova & Krishnamurti, 2002). Raftery et al. (2005) used a Bayesian Model Averaging approach to derive weights. Johnson and Swinbank (2009) also used some weighting methods in their multi-model mean sea level pressure (mslp) and 500 hPa height forecasts; they concluded that a simple RMSE skill based method to derive weights improves the multi-model forecasts.

THORPEX Interactive Grand Global Ensemble (TIGGE) network gives a platform to use the strategy of multi-models in order to capture the uncertainties in initial conditions and parameterisations of individual NWP models together with the uncertainties in structure and data assimilation (Cloke & Pappenberger, 2009). The TIGGE network provides a collaboration platform to improve development and understanding of ensemble weather predictions from around the world (Bougeault et al., 2010). The TIGGE network covers large parts of the globe and is detailed enough to use for flood forecasting (Cloke & Pappenberger, 2009). The TIGGE network thus has great potential for global scale forecasting and has been used in many hydrological and meteorological forecasting studies (Ye et al., 2014). Several studies showed already that the TIGGE database can produce an improved early flood warning of up to 10 days ahead (He et al., 2010).

Using EPSs in flood forecasting systems usually requires some kind of meteorological post-processing (Cloke & Pappenberger, 2009). This means that the meteorological input used by the hydrological model is not equivalent to the original EPS forecasts. Scale corrections are required and also the ensembles may need to have some kind of correction applied for under-dispersivity or bias. Under-dispersivity means that there is not enough spread, and thus under-representation of uncertainty. If an ensemble is biased this means that there is a difference between climatic statistics of ensemble predictions and corresponding statistics of related observations. Scale corrections are often required if the time/space scale of the hydrological model does not match the scale of the meteorological model. Therefore the EPS forecasts are usually downscaled or disaggregated in some way.

Generally, literature agrees that EPS flood forecasting is a useful activity and has the potential to inform early flood warning (Cloke & Pappenberger, 2009). Published literature gives encouraging indications that such activity brings added value to medium-range flood forecasts, especially in the ability to issue flood alerts earlier and with more confidence. However, there is a lack of evidence and many more case studies are needed.

1.3. Research gap

Since more frequent floods have been experienced by regional communities in recent decades in catchments, flood forecasting is becoming more important. As described before, NWP forecasts can extend lead times in comparison with forecasts based on observed data forcing a hydrological model. EPSs of NWPs are even more attractive for flood forecasting systems, because they have both the potential to extend lead time and better quantify the predictability. Up to now, this method has not been used in flood forecasts in the Quzhou River basin. In addition, more research on hydrological ensemble prediction systems is required (Cloke & Pappenberger, 2009).

Techniques to deal with multi-model forecasts need to be developed. Various studies applied the principle of equal probability of selection (Bao et al., 2011; He et al., 2010; Park et al., 2008). This means that every ensemble model has the same weight in the multi-model forecast. However, different weather forecasts may be assigned a different weight coefficient depending on their skill. This might improve the performance of the grand ensembles, because with equal weights large ensemble models have more influence than small ensemble models and with the weighting based on skill the better performing models have more influence in the multi-model (Park et al., 2008). In the state of the art is described that various studies have shown that model-dependent weights result in improvements, but that care should be taken in how the weights are calculated and used for the combination of the models. Raftery et al. (2005) used a Bayesian Model Averaging approach to derive weights with the result of improved multi-model forecasts. Johnson and Swinbank (2009) also used some weighting methods in their multi-model forecasts, they concluded that simple skill based methods to derive weights also improves the multi-model forecasts. However they only used a deterministic RMSE skill score for the weights and the forecasts used were mean sea level pressure (mslp) and 500 hPa height. Therefore it is interesting to investigate if multi-model ensemble flood forecast based on a probabilistic weighting will lead to higher improvements compared to weighting based on the deterministic RMSE.

1.4. Research objective and questions

1.4.1. Research objective

The purpose of this study is to develop an ensemble flood forecasting system for Quzhou (East-China) for lead times of 1 to 10 days and to evaluate different combined Grand Ensemble flood forecasts.

1.4.2. Research questions

In this paragraph the research questions are described to achieve the purpose of this study.

1. What is the performance of the meteorological forecasts and the hydrological model and how does this improve with the implementation of a bias correction method and a hydrological updating procedure?

2. What are the performances of the ensemble flood forecasting system for the different TIGGE ensemble prediction models in the study area?

3. What are the performances of grand ensemble flood forecasts with different weighting methods?

1.5. Report outline

Chapter 2 describes the study area, the data and hydrological model used in this study. Chapter 3 describes the research methodology. The results of the implementation of the methods, the single EPS forecasts and the grand ensemble forecasts are given in chapter 4. Chapter 5 presents a discussion about the study. Finally, conclusions and recommendations are presented in chapter 6.

2. Study area, data and hydrological model

This chapter is about the study area and the data used as input for the bias correction method, the hydrological model (GR4J) and for the evaluation of the forecasts. Also the GR4J model and the calibration and validation of the model are described in this chapter. In this study daily observed precipitation, daily observed discharge, daily potential evapotranspiration and raw ensemble precipitation forecast data of NWP models from the TIGGE database are used.

2.1. Study area

The study area is located in the upper reaches of the Qiantang river basin, located in the Zhejiang Province in East China. Quzhou, the city wherefore flow forecasts will be derived, is located in the Lanjiang river basin, which is one of the two important sub-basins of the Qiantang river basin (Xu et al., 2013). The basin Lanjiang is in the southern region of the Qiantang river basin. Quzhou is downstream of a sub-basin of the Lanjiang river basin called the Quzhou river basin (Tian et al., 2014). This basin is therefore relevant in this study (see Figure 1). The Quzhou river basin has a catchment area of 5,290 km^2 and is dominated by mountains and hills. The climate in the basin is semi-humid with an annual mean precipitation and temperature of 1500 mm and 15-18 °C respectively. Maximum temperature is about 40 °C. Characteristic for the climate are the hot and rainy summers and cold and dry winters. More than 50 % of the annual precipitation occurs from May to July.

There are three meteorological stations in the study area and one discharge station (Quzhou). The station in Quzhou observes the discharge, precipitation and evaporation. The other two stations in Misai and Changshan only observe precipitation.

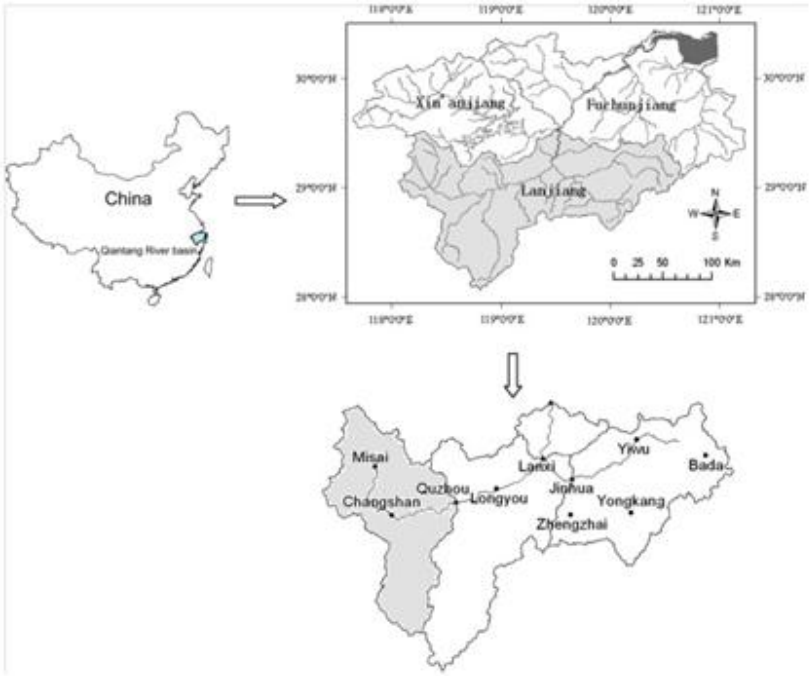

Figure 1 Location of the Quzhou river basin and the meteorological stations. The grey area is the Quzhou river basin. The meteorological stations are also showed. (Xu et al., 2013)

2.2. Observed data

Observed precipitation is used for the validation of the GR4J model, for the bias correction of the raw TIGGE ensemble precipitation data and for the perfect forecast simulations for the evaluation of the flood forecasts. Observed discharge is used for the validation of the GR4J model; for the evaluation of the ensemble forecasts and for the hydrological updating procedure used in this study to update the model states every time step during the forecast period. Temperature data is used to calculate the climatological potential evapotranspiration. The climatological value for the potential evapotranspiration will be used, which is a seasonally variable evapotranspiration, because the TIGGE archive does not have forecasts of evapotranspiration. In addition, previous studies have shown that there were no systematic improvements in the rainfall-runoff model efficiencies when using temporally varying evapotranspiration for the GR4J model and the other GR models (Oudin et al., 2005).

Observed precipitation data come from three meteorological stations in the Quzhou river basin: Quzhou, Misai and Changshan. Observed discharge data comes from the Quzhou meteorological station (see section 2.1). The data are available for the period 01/01/2009-31/12/2013 and are issued at the time step 00:00 UTC. Figure 2 show the timeseries of the areally averaged observed daily precipitation and Figure 3 the timeseries of the observed daily discharge. Figure 3 shows that there are two periods with errors. Missing values of the observed daily discharge are interpolated and are therefore not similar to the historic discharges (see Figure 4). These periods with interpolated values will therefore not be used in this study.

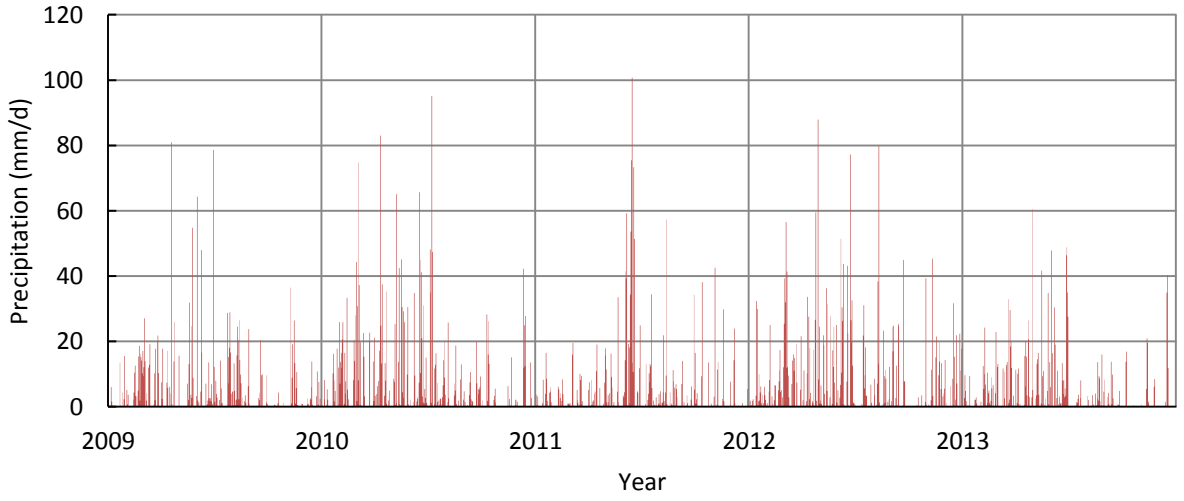

Figure 2 Pluviograph of observed daily areally averaged precipitation for the period 2009-2013. Data is retrieved from the meteorological stations Quzhou, Misai and Changshan.

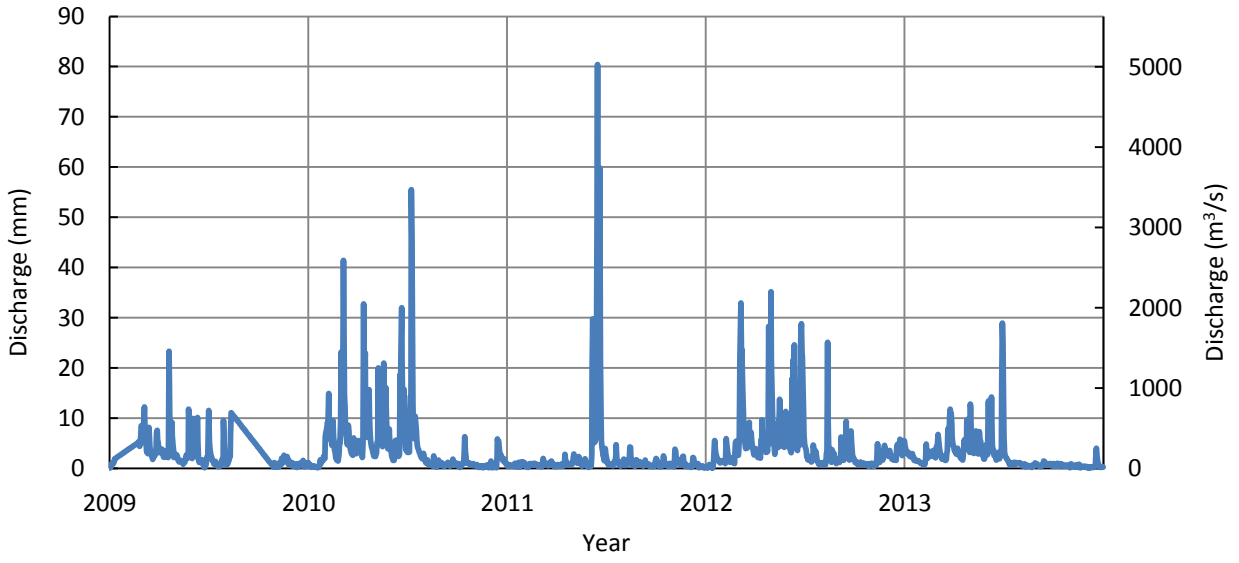

Figure 3 Hydrograph of the observed daily discharge for the period 2009-2014. Data is retrieved from the Quzhou meteorological station.

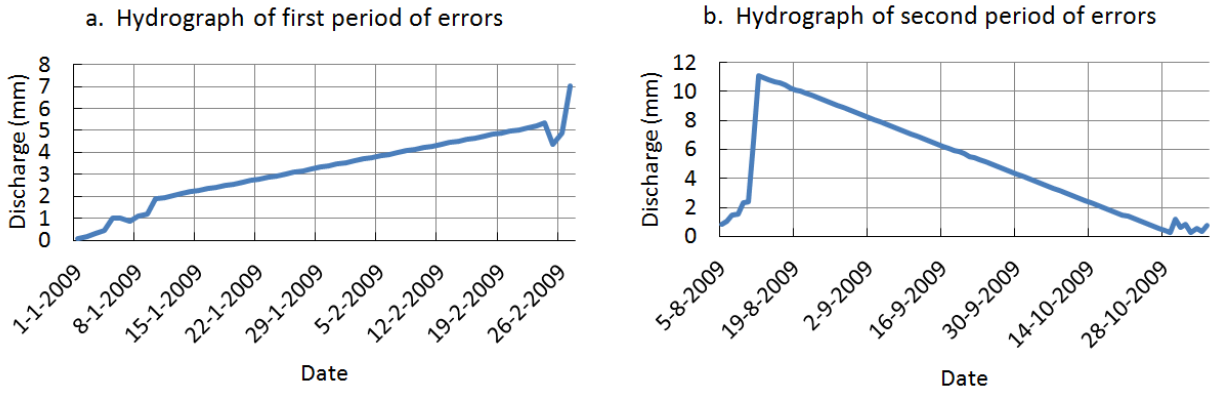

Figure 4 Errors in the timeserie of the observed daily discharge.

2.3. TIGGE ensemble precipitation forecast data

The TIGGE network consists of several NWP centres which generate ensemble forecasts and covers large parts of the globe and is detailed enough to use for flood forecasting. Therefore, the TIGGE network has great potential for global scale forecasting and has been used in many hydrometeorological forecasting studies (Ye et al., 2014). TIGGE is a component of THORPEX. THORPEX is the World Weather Research Programme project with the aim to accelerate improvements in the accuracy of 1-day to 2-week high-impact weather forecasts (Bougeault et al., 2010). TIGGE is a key component to achieve this aim and was initiated in 2005. Several studies showed already that the TIGGE database can produce an improved early flood warning of up to 10 days ahead (He et al., 2010). TIGGE develops a deeper understanding of the contribution of observation, initial and model uncertainties to forecast error and investigates new methods of combining ensembles from different sources to correct systematic errors (Bougeault et al., 2010).

Ten centres supply daily forecasts to the TIGGE archive (Park et al., 2008). Nine of these centres are running a medium-range global ensemble prediction system: European Centre for Medium-Range Weather Forecasts (ECMWF); US National Centers for Environmental Prediction (NCEP); Meteorological Service of Canada (MSC); the Australian Bureau of Meteorology (BoM); the Chinese Meteorological Administration (CMA); the Brazilian Centre for Weather Prediction and Climate Studies (Centro de Previsao de Tempo e Estudos Climáticos, CPTEC); the Japanese Meteorological Administration (JMA); the Korean Meteorological Administration (KMA); and the UK Met Office (UKMO) . Météo-France has a short forecast range. In Park et al. (2008) a medium-range global ensemble system is formulated as an ensemble system designed to provide probabilistic forecasts for at least up to 7 days and for the whole globe.

Ensemble prediction systems are designed to represent the effect on weather forecasts of observation uncertainties, imperfect boundary conditions and data assimilation assumptions and model uncertainties (Park et al., 2008). Model uncertainties may occur due to a lack of resolution, simplified parameterization of physical processes and the effect of unresolved processes. Data-assimilation assumptions may occur due to the data-assimilation methods and underlying statistical assumptions. When a grand ensemble of EPS from different weather centres combined is used also other sources of uncertainties, including numerical implementations and/or data assimilation, can be assessed (He et al., 2010). The aggregation of various models that produce EPS from different weather centres also results in a better retaining of and accounting for the probabilistic nature of the ensemble precipitation forecasts (Bao et al., 2011).

The TIGGE ensemble prediction systems are based on several time integrations of a numerical weather prediction model, with the control forecast starting from a 'central' analysis, this is the unperturbed analysis generated by a data-assimilation procedure, and the other perturbed forecasts starting from perturbed initial conditions defined to simulate the effect of initial condition uncertainties (Park et al., 2008).

Buizza et al. (2005) studied the three global ensemble systems ECMWF, MSC and NCEP and concluded that for these systems the spread of ensemble forecasts is insufficient to systematically capture reality and suggested that none of them is able to simulate all sources of forecast uncertainty. Therefore MSC and NCEP have decided to combine their operational ensemble systems in the North American Ensemble Forecasting System (NAEFS) to address the suboptimal simulation of model uncertainties and the limited ensemble size (Park et al., 2008). The other centres also investigated the potential of combining ensemble forecasts generated by different centres and established TIGGE. Since then three centres (CMA, ECMWF and NCAR (US National Centre for Atmospheric Research)) became TIGGE Data Centres, and have started collecting the TIGGE ensemble data of the different NWP centres (see Figure 5). The three TIGGE Data Centres made the data accessible to the scientific community for research and education with a 2 day time delay (Bougeault et al., 2010).

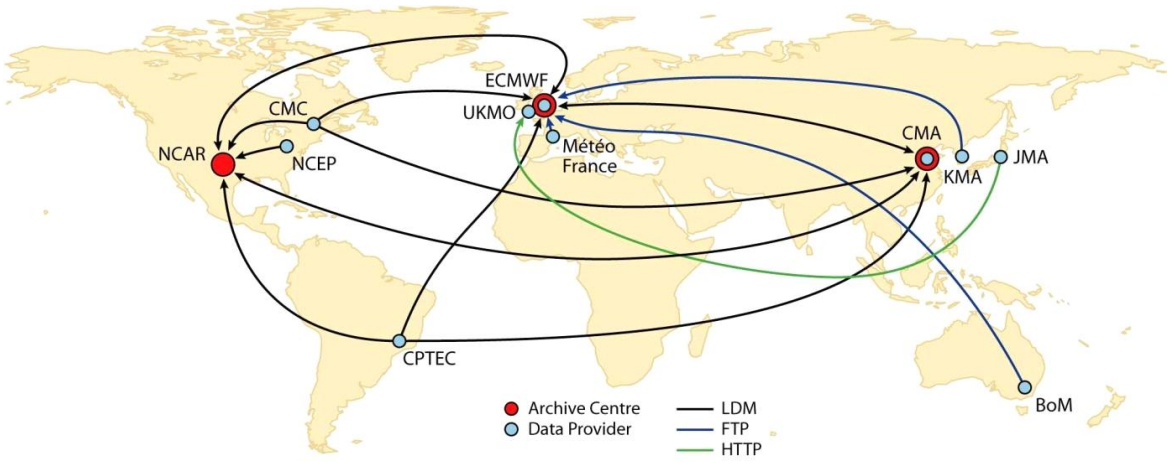

Figure 5 The TIGGE network with its data providers and archive centres (Orientplus, 2015).

2.4. Content and format of the TIGGE archive

Bougeault et al. (2010) described the content and format of the archive. Fields in the TIGGE dataset are described by the following attributes: analysis date, analysis time, forecast time step, origin centre, ensemble member number, level, and parameter. The parameter in the TIGGE dataset refers to the physical quantity represented by the field and is in this research precipitation only, because for the other input parameter, evapotranspiration, the daily climatological value is used.

Data providers preserve their original model grids and resolutions whenever possible to guarantee the best precision (Bougeault et al., 2010). Therefore they can choose their own horizontal grid to supply their data on, which will be as close as possible to the computational grid of their model. The data are stored in the database in their original state. However, users usually want data interpolated on common regular grids of their own choice. Therefore, the archive centres offer an interpolation service. This interpolation service allows users to interpolate data to a single point or to a regular, limited-area, or global latitude-longitude grid specified to their own choice.

Data used from the TIGGE archive

The ensemble precipitation forecasts are retrieved from four weather centres: ECMWF, CMA, UKMO and NCEP. ECMWF, NCEP and UKMO are chosen because they show the highest skill in different studies(Su et al., 2014; Tao et al., 2014). The data of JMA is available from 2011 and is thus not suitable for this study (ECMWF, 2015). CMA is chosen because it is the Chinese weather centre and the study area is in China. Each of these centres provides one 'central' unperturbed analysis and a number of forecasts with perturbed initial conditions. ECMWF has 50 perturbed ensemble members, CMA 14, UKMO 23 and NCEP 20. The grid scale of the ensemble forecasts are automatically interpolated by the TIGGE interpolation service to a grid scale of 0.5°x0.5°. These meteorological grid scales are considered to be comparable with the lumped hydrological model and the observed precipitation. With Thiessen Polygons the grid forecasts are calculated to areally averaged precipitation forecasts to be used along with the observed areally averaged data. ECMWF delivers forecast lead times up to 15 days, while CMA delivers forecast lead times up to 10 days. Hence, the lead times used in this study are up to 10 days. The TIGGE data are retrieved for the period from 17/12/2008 to 14/10/2013. It should be noted that the retrieved TIGGE

data period does not cover the whole data period of available observed data. This is because there are missing data in the TIGGE archive of the CMA model from 30/10/2013 till 14/11/2013. Also noteworthy is that the retrieved data from the TIGGE archive starts at 17/12/08, because for bias correction of the data a moving window of 31 days will be used (15 days before and 15 days after the forecast issue date). The resulting validation period of the TIGGE data will be from 01/01/2009 to 14/10/2013. The TIGGE data are retrieved for the time step 00:00 UTC to get along with the other observed data which is also issued on 00:00 UTC. There is one missing forecast in the NCEP forecast data in the data period of retrieved TIGGE data. Su et al. (2014) had the same problem for the NCEP forecast dataset. They considered that replacing this small fraction of data will not influence the final results. The missing NCEP forecast data are therefore replaced with the interpolated value of precipitation values of the day before and after this missing day.

The data retrieved from TIGGE is the accumulated total precipitation. The data are processed to 24 hour accumulated precipitation values by the subtraction of the accumulated total precipitation of the lead time -1 day. However, after this process there are some negative values. Small negative values (-1 - 0 mm/d) are caused by the scaling of values during Gridded Binary (GRIB) packing or interpolation errors (ECMWF, 2013). All negative values of 24 hour precipitation forecasts are set to zero which was also done by Su et al. (2014).

Table 1 Weather centres and their properties used in this study (ECMWF, 2015)

Weather centre	Number of Members	Grid scale	Lead times	Period	Time step
ECMWF	51	0.5°x0.5°	1-10 days	17/12/08 - 29/10/13	00:00 UTC
CMA	15	0.5°x0.5°	1-10 days	17/12/08 - 29/10/13	00:00 UTC
UKMO	24	0.5°x0.5°	1-10 days	17/12/08 - 29/10/13	00:00 UTC
NCEP	21	0.5°x0.5°	1-10 days	17/12/08 - 29/10/13	00:00 UTC

2.5. GR4J model

2.5.1. Model description

The hydrological model used in this study is the GR4J model (modèle du Génie Rural à 4 paramètres Journalier). GR4J is a daily lumped four-parameter rainfall-runoff model and belongs to the family of soil moisture accounting models (Perrin et al., 2003). The GR4J model is the last modified version of the GR3J model. Figure 6 shows the model structure of the GR4J model. The model has as input P (areal catchment rainfall) and E (areal catchment potential evapotranspiration (PE)). E can also be a long-term average value, the climatological value, as used in this study. In this case the same PE series is repeated every year. The model consists of a production function and a routing function. The production function computes the net rainfall and PE, a production store S and percolation leakage from the production store S. The routing function includes unit hydrographs and a non-linear routing store which transforms the

unit hydrographs together with a calculated groundwater exchange F into a catchment outflow. The GR4J model consists of four parameters that have to be optimised for the catchment by the use of observed discharge:

x_1: maximum capacity of the production store (mm)
x_2: groundwater exchange coefficient (mm)
x_3: one day ahead maximum capacity of the routing store (mm)
x_4: time base of unit hydrograph UH1 (days)

All these four parameters are real numbers. x_1 and x_3 are positive, x_2 can be either zero, negative or positive and x_4 is greater than 0.5. The GR4J model runs at daily time steps.

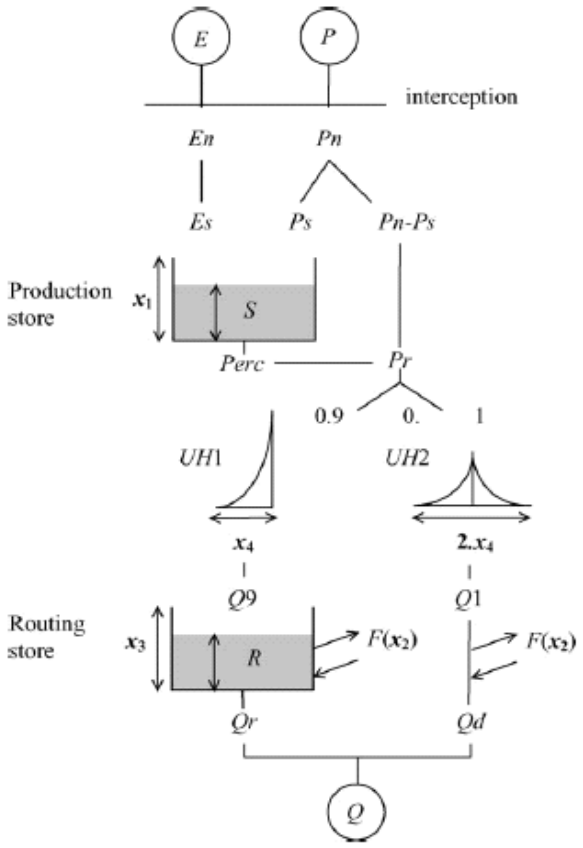

Figure 6 Model structure of the GR4J rainfall-runoff model (Perrin et al., 2003)

2.5.2. Calibration and validation of the GR4J model

The model has been calibrated by Tian et al. (2014) and observed data from 1981-1990 were used to calibrate the model using the Generalised Likelihood Uncertainty Estimation (GLUE) approach. The GLUE method is a Bayesian analysis based Monte Carlo method for model calibration and uncertainty analysis. The likelihood function chosen to calibrate the model was the Nash-Sutcliffe efficiency (NS) coefficient.

Tian et al. (2014) used 30,000 randomly generated parameter sets in the GLUE method. For each parameter set, the NS was calculated. The optimum simulation result was then obtained through the

parameter set with the maximum NS value. Also the Relative Volume Error (RVE) was calculated for the optimum parameter set.

The maximum NS value for the Quzhou River Basin in the calibration period was 0.93 and the corresponding RVE was -1.1. The obtained optimum parameter set for the Quzhou river basin is presented in table 2.

Table 2 Optimum parameter set for the Quzhou river basin (Tian et al., 2014)

Parameter	Explanation	Value
X1	Maximum capacity of the production store	141.1 mm
X2	Groundwater exchange coefficient	0.1 mm
X3	One day ahead maximum capacity of the routing store	44.7 mm
X4	Time base of unit hydrograph	2.1 days

This optimum parameter set is used for the validation of the model. The model has been validated in this study for the period 2009-2013, which is the same period as the forecasting period. The NS value for the validation of the calibrated model is 0.91 and the RVE is -3.03. The model performance for the validation period is just a little worse than the performance of the model during the calibration period, but remains high. In this study also a hydrological updating approach will be used to improve the performance of the model.

3. Methods

This chapter describes the methods used in this study to develop the ensemble flood forecasting system and the grand ensembles. Also the evaluation methods are described. In section 3.1 the bias correction approach is described. Section 3.2 describes the updating procedure used in this study. Section 3.3 describes the evaluation methods to evaluate the ensemble forecasts. The last section describes the combination method of the EPSs to construct a grand ensemble forecast. Figure 7 shows a flow chart of the forecasting system.

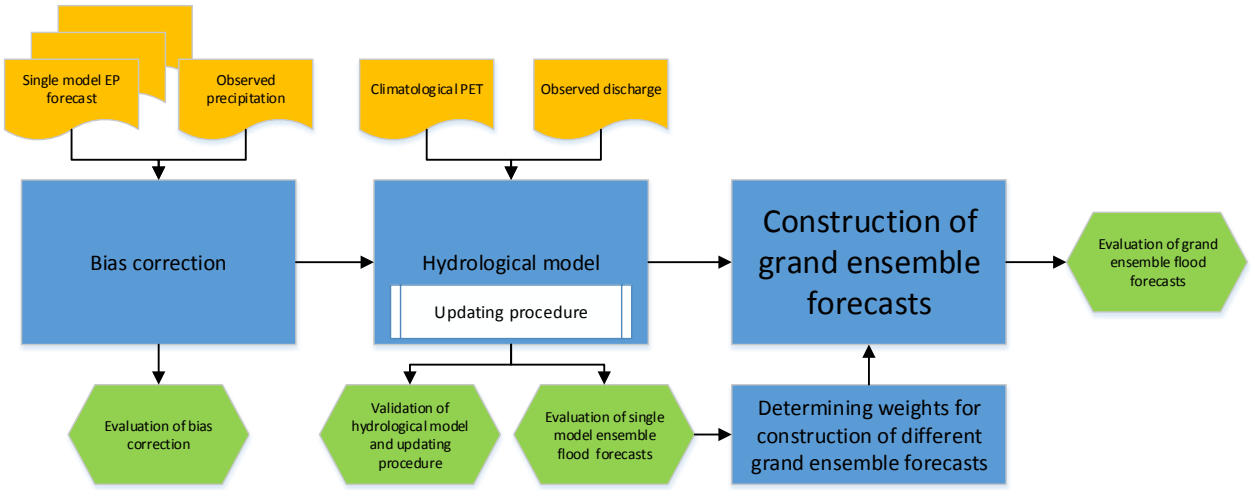

Figure 7 Flow chart of the methods used in this study (Blue objects form the Ensemble forecast system; Orange objects are the input data of the Ensemble forecast system; Green objects are various evaluations steps in this study).

3.1. Bias correction Quantile mapping method

In general the raw ensemble forecasts from NWP models are biased in the mean and spread (Wu et al., 2011). Bias is the systematic difference between the forecast and its verification which is often an observation. Even if the forecasts are not biased at the model grid scale, they may be biased at the catchment scale, due to heterogeneity in the forecasted variable within the model grids. This is depending on the size of the basin. Correcting such biases is normally referred to as post-processing or statistical calibration. There are four reasons why EPSs in flood forecasting systems usually require some kind of meteorological post-processing (Tao et al., 2014):

1. The accuracy of the raw ensemble meteorological forecasts evaluated at grid size is still limited and not suitable for direct hydrological modelling to forecast floods, even though the ensemble meteorological forecasts have been improved significantly.

2. The spread of the raw meteorological ensembles may be unreliable / underdispersed.

3. The spatial resolution of meteorological ensemble forecasts is not equivalent to those required for generating hydrological forecasts. Hydrological models are usually run over catchments, while meteorological forecasts are generally run over grids.

4. The temporal resolution of meteorological ensemble forecasts is not equivalent to those required for generating hydrological forecasts.

Bias-correction for precipitation ensemble forecasts has proven to be very challenging because of the large space-time variability of precipitation (Wu et al., 2011). Wu et al. (2011), therefore expect that significant additional efforts will be needed to produce operational ensemble forecasts that are good enough for hydrological applications, especially for large precipitation amounts and small catchments.

Statistical methods are often used for post-processing and downscaling raw meteorological ensemble forecasts. Statistical downscaling is often performed to correct for point 3 and 4 in the numeration above. However, since the TIGGE archive centres offer an interpolation service, downscaling of the raw meteorological ensemble forecasts is done with this service. In 2.4 was described that the data is downloaded at a resolution of 0.5° x 0.5° which is comparable to the observation and the GR4J model used. Therefore the focus in this section is on the bias correction method to correct for point 1 and 2.

Systematic bias is unavoidably present in the precipitation forecasts, and is usually a function of spatial location and forecast lead time (Voisin et al., 2010). For stream flow forecasting a preferred approach of bias correction is to use a bias correction transformation to correct all model-simulated ensemble time series (Hashino et al., 2007), because bias-corrected ensemble time series can be used in water resources applications. One of these bias correction methods is the quantile mapping method.

The quantile mapping method uses the cumulative distribution functions (CDF) for observed and simulated values for each lead time to remove the biases. The quantile mapping method tends to improve the skill score and tends to lead to high sharpness (the tendency of the forecast to predict extreme values (WMO, 2015)) and discrimination (the ability to discriminated among observations, meaning that forecasts have higher prediction frequencies for event occurrences and lower for nonoccurrences (WMO, 2015)) (Hashino et al., 2007) . Therefore this method is used in this study.

Forecasts have discrimination when the forecasts issued for different outcomes (event occurrences or nonoccurrences) are different. Hence, for forecasts to have good discrimination, they must both be sharp and have high potential skill

The bias correction of quantile-based mapping is achieved by replacing the forecasted value with observed values with the same percentiles (nonexceedance probabilities) (Voisin et al., 2010). Bias correction is done for each day and each lead time in the set of 10-day forecasts in the period 2009-2013. The corresponding method is as follows:

1. Derivation of the forecast daily cumulative distribution function (CDF)

The CDFs of the daily areally average forecasts were derived for a 31-day moving window (15 days before and 15 days after the issue date were included). It is a daily CDF resulting in 366 CDFs for each lead time. For each 31 days moving window, the CDFs were derived using all ensemble members of the EPSs precipitation forecasts issued over the Jan 2009 - Oct 2013 period. For the CDF of the ECMWF EPS forecasts this means that 51 members x 31 days x 4 or 5 years = 6324 or 7905 values have to be ranked for the CDF; depending on how many times the issued date is in the dataset. For the CDF of the CMA EPS

forecasts only 15 members x 31 days x 4 or 5 years = 1860 or 2325 values have to be ranked. The CDF of UKMO consists of 24 members x 31 days x 4 or 5 years = 2976 or 3720 values and of NCEP 21 members x 31 days x 4 or 5 years = 2604 or 3255 values.

2. Derivation of the observed daily CDF

The Jan 2009 - Sept 2013 daily precipitation datasets from the different weather stations in the Quzhou river basin were interpolated with the Thiessen Polygon method to daily areally averages. The CDF of the observed daily areally average precipitation is also derived for a 31 day moving window. Using the 31 day moving window, in between 31 days x 4 and 5 years = 124 and 155 values had to be ranked in each CDF; depending on how many times the issued date is in the dataset. The daily CDFs were derived for each day, so 366 CDFs were derived.

3. Quantile mapping

The quantile mapping approach is applied to each daily ensemble forecast set in the Jan 2009 - Oct 2013 period. Each ensemble member and each lead time of the different EPSs is corrected with the quantile mapping approach independently so that different biases at different lead times can be corrected and that the forecast can be corrected in the spread. The quantile (Qn) of the daily precipitation forecast member is estimated in the corresponding forecast CDF (appropriate day, centre of the 31 days moving window, and lead time. This estimated quantile is substituted for the observed value with the same quantile in the corresponding daily CDF (CDF for that day and lead time) (see Figure 8). The corresponding definition of the quantile mapping method is as follows:

$$BC_{fcst} = {CDF_{obs}}^{-1}\left(CDF_{fcst}(Fcst)\right) = {CDF_{obs}}^{-1}(Qn)$$

Where BC_{fcst} is the bias-corrected forecast value, $Fcst$ is the forecast value, $CDFobs$ is the CDF of the observed climatology, CDF_{fcst} is the forecasted CDF, and Qn is the quantile of the forecast value in the forecast CDF.

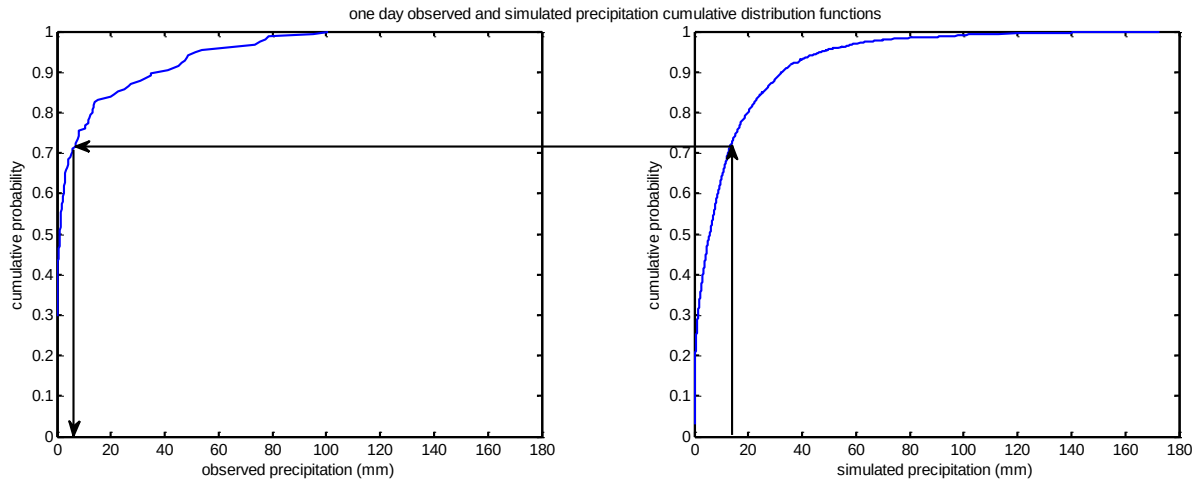

Figure 8 Quantile mapping approach with the daily observed and simulated precipitation cumulative distribution functions

4. Daily precipitation intermittency correction

The quantile mapping method is limited when dealing with intermittency. Intermittency is defined as the difference in the dry day frequency of the raw model output and the observations. The precipitation intermittency issue can occur in two ways. First there is the case that the forecasts have less dry days than the observations and second the case that the forecasts have more dry days than the observations. In the first case the intermittency is automatically corrected by the quantile mapping bias correction method. The quantiles of the forecast distribution below the no-precipitation threshold in the CDF of the observations are defined as dry (Figure 9a). However, for the second case, when the forecasts have more dry days than the observation the intermittency is not automatically corrected by the quantile mapping bias correction method. Using the quantile mapping method for bias correction in this case leads to a strong positive bias after the correction (Figure 9b). Therefore, an intermittency correction approach is implemented for the bias correction of the precipitation forecasts. In this intermittency correction approach quantiles are randomly selected between zero and the corresponding quantile of the no-precipitation threshold from the forecasted CDF. The corresponding observation quantile of this randomly selected quantile may or may not be associated with precipitation. This approach can be presented as follows:

$$BC_{fcst} = 0, \qquad if \quad Fcst = 0 \quad \& \quad Qn_{obs} \geq Qn_{fcst}$$

$$Qn = random\left(0, Qn_{fcst}\right), \qquad if \quad Fcst = 0 \quad \& \quad Qn_{obs} < Qn_{fcst}$$

Where Qn_{obs} and Qn_{fcst} are respectively the largest observed and forecast quantiles associated with a zero precipitation value.

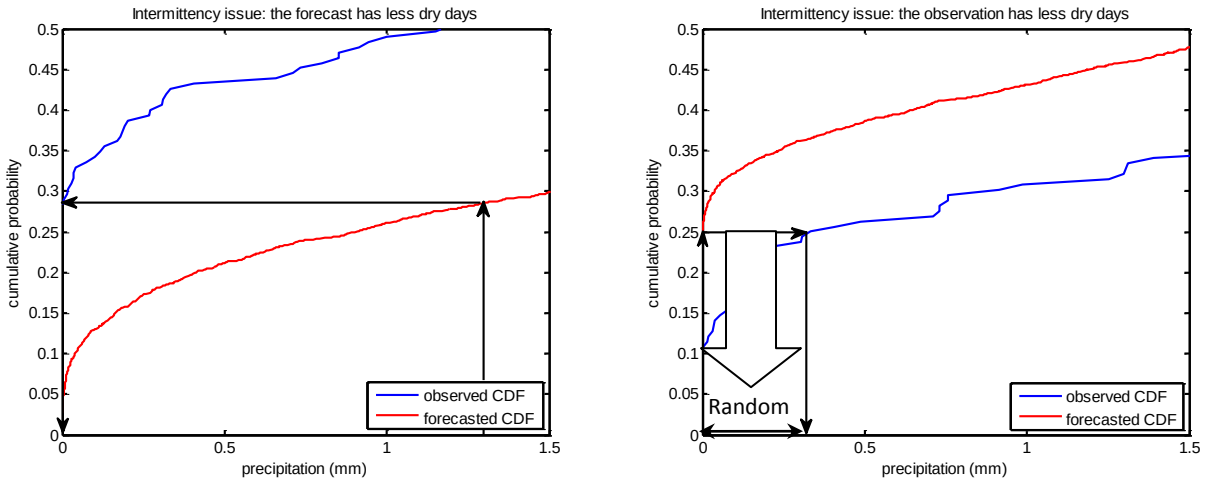

Figure 9 Intermittency issues: the forecast has less dry days then the observation (left); the observation has less dry days then the forecast (right)

3.2. Hydrological updating

3.2.1. Introduction

As described before there are many sources of errors, when forecasting runoff. In ensemble flood forecasting forecasted precipitation input data are used in hydrological models to extend lead times. This generates a major uncertainty for the hydrological forecasting system (Kahl & Nachtnebel, 2008). As a result the simulated and forecasted hydrographs will never fit perfectly to the measurements. To compensate the input and model uncertainties partially, techniques have been developed to minimize errors in simulation of the recent history and improve the forecast. Errors in the recent past can influence the forecast negatively. Therefore updating procedures are developed to update model input, state/storages and output so that the current situation in the river basin is better represented (Wöhling et al., 2006).

Popular updating procedures are Auto Regression models and Kalman filtering, however these procedures are not suitable for short forecast periods and steep flood hydrograph characteristics which are typical for small, quickly reacting mountainous catchments such as the Quzhou catchment area (Wöhling et al., 2006). For this kind of catchments it is the primary goal to extend the forecast lead time. Moreover, classical updating procedures, such as Auto Regression models, focus on the river flow itself which leads to a significant loss of forecast lead time in small, quick reacting catchments. Also more complicated procedures, such as Kalman filtering, are mathematically too complex to be easily accommodated in highly non-linear models. Therefore a simple effective updating procedure developed by Demirel et al. (2013) for the GR4J model is used to minimize errors in the initial state.

3.2.2. Updating procedure developed by Demirel et al. (2013)

The updating procedure developed by Demirel et al. (2013) is a model storage update procedure based on the observed discharge on the forecast issue day. This is an important step for medium-range flow forecasts since the model initial state determines the model outputs. The routing storage variable in the GR4J model is updated during the flow forecasts with using this approach. In GR4J the simulated runoff Q is calculated with the fast runoff component Qr and slow runoff component Qd with the following formula:

$$Q = Qr + Qd$$

The fast runoff component and slow runoff component in the GR4J model can be estimated with the use of the empirical relations between the simulated discharge and the fast runoff to divide the observed discharge between the fast and slow runoff components. With this empirical relation a fraction k of the slow runoff component compared to the simulated discharge can be calculated:

$$k = \frac{Qd}{Qr + Qd}$$

k is calculated each day of the forecast. In this updating routine the observed discharge at the forecast issue date is related to the updated Qr and Qd ($\tilde{Q}r$ and $\tilde{Q}d$ respectively) and consequently k as expressed in the following equations:

$$\tilde{Q}d = k * Q_{obs}$$

$$\tilde{Q}r = Q_{obs} - \tilde{Q}d$$

In the GR4J model the outflow Qr is calculated as:

$$Qr = R\left\{1 - \left[1 + \left(\frac{R}{X3}\right)^4\right]^{-1/4}\right\}$$

Since Qr and Qd are updated with the equations above, the routing storage (R) will be updated for a given value of the X3 parameter by inverting the latter equation. This updated routing storage is used for the calculation of the forecast of the next day.

3.2.3. Implementation in the ensemble flow forecasting system

The updating procedure in this study provides initial model storages for the forecast issue day based on the observation and the simulation of the day before the forecast issue date. This is the approach for a lead time of 1 day. If this approach would be used for longer lead times the model states would be updated with the observation value of lead time days before the forecast issue date. This approach would result in a fast decrease in the performance of the model, since the autocorrelation is decreasing fast with lead time in a small mountainous quick reacting catchment. Therefore the initial states are not updated with the approach for lead times longer than 1 day. Instead, the initial states calculated with the GR4J model of the previous lead time are used (see Figure 10).

3.2.4. Check whether the hydrological updating procedure is realistic

To ensure the hydrological updating procedure is realistic, the routing storages with and without the hydrological updating procedure are calculated over the validation period. With this comparison it becomes clear whether the use of the hydrological updating procedure results in tuning (small difference) or curve fitting (big difference) of the new simulated discharges. If the routing storages change in order of magnitude with the implementation of hydrological updating, the predictive power of the GR4J model is weakened.

Figure 11 shows the routing storage for the years 2009-2013. The blue line represents the routing storage without updating and the red line represents the routing storage with use of the updating procedure. The calibrated GR4J model has a maximum routing storage capacity of 44.7 mm. Figure 11 shows that the routing storage is often around 30 mm and is always below the maximum capacity. The differences between the routing storages for the case with and without the updating procedure are not in orders of magnitude. Hence, the hydrological updating procedure used in this study tunes the simulated discharges instead of curve fitting and therefore the updating procedure is realistic. In addition, Figure 11 shows that the updating routine has effect on the increase of the routing storage from the start of the forecast period. The increase is faster with the result that the start-up time for the model to reach realistic values is decreased. Therefore forecast values can be evaluated from the start of the forecast period. Also notable are the two periods where the updated routing storage is curved. These errors are the effect of the interpolated observed discharge data for missing values described in section 2.2. However, these two periods will be ignored in the evaluation of the forecasts.

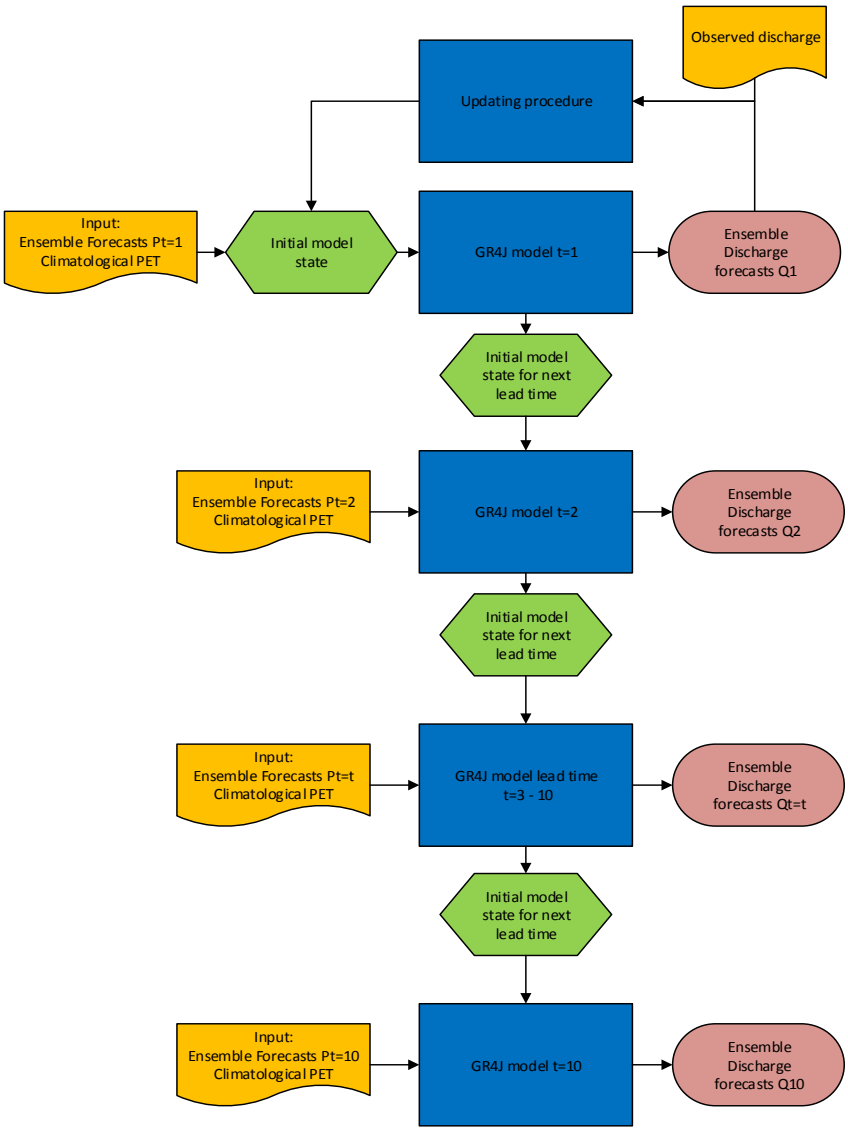

Figure 10 Ensemble flow forecast system with updating system. Blue objects are the processes, orange are the input data, green are the initial model states (routing storage) and pink are the output ensemble discharge forecasts.

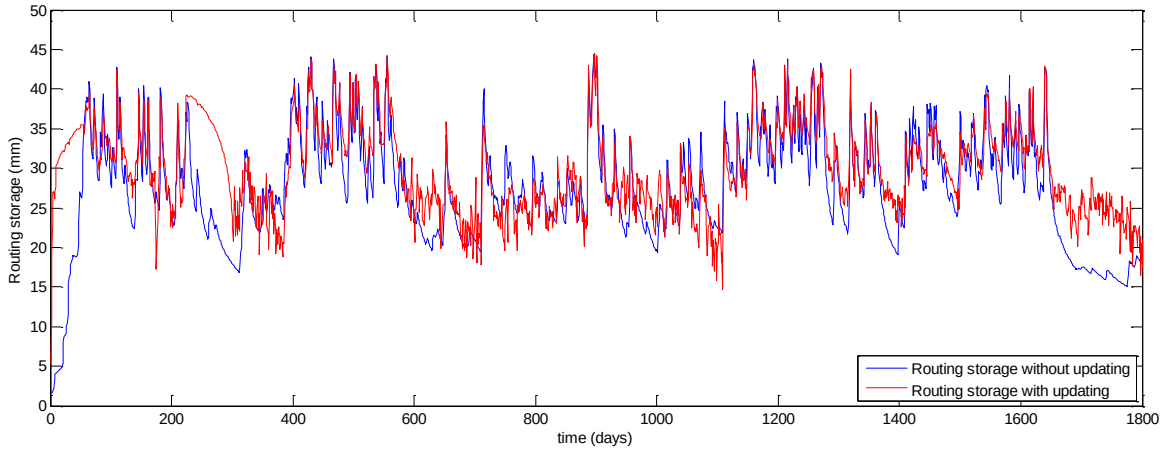

Figure 11 Routing storage (mm) with and without updating for the period 2009-2013 for lead time 1.

3.3. Evaluation methods

Evaluation methods monitor and compare the quality of forecasts and identify the strengths and weaknesses of the forecast system (Tödter & Ahrens, 2012). Evaluation possibly allows a future improvement.

Three properties of an accurate probability forecast system are defined in WMO (2015):

- Reliability: the agreement between forecast probability and mean observed frequency of an event. The closer the mean observed frequency matches the forecasted probability, the more reliable the forecasts are.

- Sharpness: the tendency to forecast probabilities of an event occurring near 0 or 1 instead of values clustering around the mean. So, the tendency of the forecast to predict extreme values.

- Resolution: the ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes. It measures how much the conditional probabilities given the different forecasts differ from the climatic average. Even if the forecasts are wrong, the forecast system has resolution if it can successfully separate one type of outcome from another.

To assess these properties, and evaluate probabilistic forecasts, several statistical evaluation measures have been proposed in the literature such as the Brier score, the ranked probability score, the continuous rank probability score, the reliability diagram and the rank histogram (Cloke & Pappenberger, 2009). The ranked probability score and the continuous ranked probability score would enable to assess the overall quality of the ensemble, or the quality on a certain range of the forecast. The BS on the other hand permits to focus on specific warnings and thresholds meaningful for studies to flood forecasts. The BS is chosen for this study as it permits to look at specific thresholds and also two out of the three properties defined above that describe an accurate forecast can be calculated, namely reliability and resolution. The other property, sharpness, can be shown by a reliability diagram with a corresponding sharpness diagram and will therefore also be used in this study. Next to these two evaluation methods the continuous ranked probability score and the root mean square error will be used for the combination of the different TIGGE ensemble forecast models and for the evaluation of the overall performance of the forecasts.

3.3.1. Thresholds

The forecasts will be evaluated at certain thresholds derived from the observed precipitation and discharge. Since this study focus on high flow forecasts the thresholds chosen in this study are quantile 75%, 85% and 95% with an exceedance of 25%, 15% and 5% respectively. Both precipitation forecasts and discharge forecasts are evaluated, so P75, P85 and P95 as well as Q75, Q85 and Q95 of the observed data will be used. Table 3 show the threshold values for the precipitation and discharge evaluation.

Table 3 Precipitation and discharge thresholds used in this study

Thresholds	75 %	85 %	95 %
Precipitation (mm)	5.31	12.31	27.75
Discharge (mm)	3.77	5.52	11.68

3.3.2. Evaluation of the forecast against reference flow

As described before hydrological forecasts used in this study contain both errors from the meteorological forecasts and the hydrological model. When evaluating these hydrological forecasts against observed discharge a third error component is present: the discharge observation measurement error (Renner et al. 2009). However, in the hydrological model used here the updating procedure also uses discharge observations to update the initial state of the routing storage of GR4J. So the discharge observation measurement error is also present in the hydrological forecasts. To give better interpretation of the forecast performance and come to better conclusions it is useful to know where the largest errors in the forecasting system are coming from. To see which errors have the largest influence some of the errors have to be removed from a reference flow. This can be done by using perfect flow forecasts as reference flows. In this case the hydrological model error component and discharge observations measurement error component are removed from the evaluation with the hydrological forecast, because both forecasts contain these two errors. With the elimination of two errors it is still not possible to know if the meteorological error contributes more to the hydrological error than the hydrological model error. However, if it is assumed that observation errors can be neglected, evaluation of the forecasts against perfect flow forecasts contain error components from the meteorological forecasts, while evaluation against observed discharge contain both error components from the meteorological forecasts and the hydrological model. For the general scores of the deterministic evaluation measure (RMSE) and probabilistic evaluation measure (CRPS) this comparison of evaluation against observation and perfect flow forecast will be made. It is not possible to do a simpler comparison with the evaluation scores of flow forecasts and meteorological precipitation forecasts, because the evaluation scores of RMSE and CRPS depend on the magnitude of the investigated parameter (discharge and precipitation respectively).

3.3.3. Deterministic evaluation

The root mean square error (RMSE) is common used in forecast evaluation. The RMSE is not a probabilistic evaluation measure but a deterministic evaluation measure (WMO, 2015). So when calculating the RMSE for a probabilistic forecast like the TIGGE ensemble forecasts first the ensemble mean have to be calculated. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$

where F_i is the ensemble mean, O_i is the observed value and N is the number of forecasts.

The RMSE is used in this study as extra evaluation of the forecasts next to the probabilistic evaluation and as a measure where weights for Grand Ensembles will be based on. Since other weighting methods are based on the CRPS it is interesting to see if the RMSE as well as the CRPS will improve when using a Grand Ensemble based on the RMSE or CRPS of single models.

3.3.4. Probabilistic evaluation

3.3.4.1. Continuous Ranked Probability Score (CRPS)

The CRPS is also often used as an evaluation method for probabilistic forecasts. The CRPS is an evaluation tool that evaluates the accuracy of a probabilistic forecast distribution by comparing the distribution of an ensemble of forecasts to the observed value (Liu & Xie, 2014). The CRPS has some interesting properties. First, it is sensitive to the entire permissible range of the parameter of interest, discharge in this case. Second, its definition does not require the introduction of a number of predefined classes on which results may depend as is the case with the ranked probability score. The CRPS takes instead of the ranked probability score the limit of an infinite number of classes, each with zero width (Hersbach, 2000). This avoids loss of information and sensitivity of the score to the number and choice of the thresholds values (Tödter & Ahrens, 2012). Last, the CRPS is equal to the mean absolute error for a deterministic forecast; it therefore has a clear interpretation. The CRPS is an evaluation method that is sensitive to the overall (with respect to a certain variable) performance of a forecast system. The CRPS of a probability forecast is defined as:

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}[F_i(x) - O_i(x)]^2\,dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\left[F_i(x) - H\left(x - x_{o,i}\right)\right]^2\,dx$$

where *n* is the number of forecasts, *F(x)* is the forecast cumulative distribution function (CDF), *O(x)* is the observed CDF. $H(x-x_0)$ is the Heaviside function that jumps from 0 to 1 at the observed value, if $x < x_{o,i}$, $H(x-x_o) = 0$, otherwise, $H(x-x_{o,i}) = 1$. An example for the cumulative distribution for an ensemble of five members and the cumulative distribution of the observation is given in Figure 12. Note that the shaded area, representing the CRPS, is not similar to the area under or above the forecast CDF given the observation. This is because the CRPS takes the integral of the squared difference between the forecast CDF and the observed CDF. Hence the shaded bars are the squared difference between the blue and the red line in Figure 12. The CRPS has the dimension of the parameter x and is like the Brier score also negatively oriented. So, worse forecasts receive higher values. The minimal value of zero is only achieved when *F(x)* is the same as *O(x).*

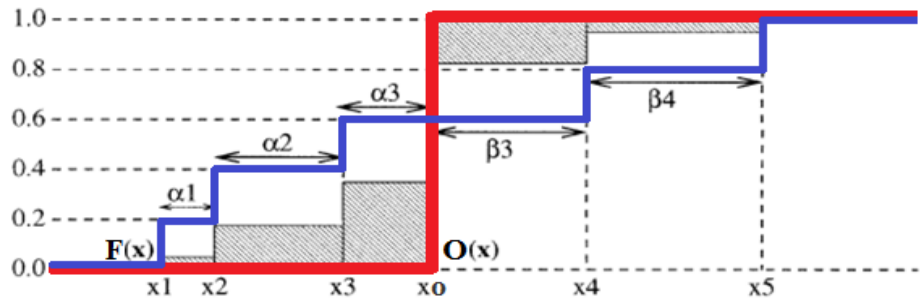

Figure 12 Example of a cumulative distribution for an ensemble {x_1, ... , x_5} of five members (blue line) and for the verifying analysis of the observation x_o (red line), the CRPS is represented by the shaded area. (Hersbach, 2000)

The calculation of the CRPS will be done for the cases where the observation is above the Q75 value, for the evaluation of the forecasts and the calculation of the weights for the combination of the forecasts.

3.3.4.2 Continuous Ranked Probability Skill Score (CRPSS)

The CRPSS is a normalized form of the CRPS. The CRPS is normalized by a reference which is usually the climatology or persistence of a study area. The Continuous Ranked Probability Skill Score (CRPSS) is used to evaluate the quantitative skills of the ESP to study the added value of the probabilistic forecasts.
CRPSS is defined as (e.g. Hersbach, 2000):

$$CRPSS = 1 - \frac{CRPS_{forecast}}{CRPS_{reference}}$$

CRPSS measures the improvement of an ensemble forecasting system over the reference forecast. The reference forecast is an alternative forecast and is in this study compared to the hydrological forecast. The value of CRPSS ranges from -∞ to 1 (Alfieri et al., 2014). CRPSS gives a score of 1 when the forecast has perfect skill, while negative values indicate worse performance than its reference forecast. So EPSs are only valuable when CRPSS > 0. In that case the forecast is performing better than the reference forecast.

For the calculation of the CRPSS different reference forecasts can be used. The better the reference forecast the more difficult it is to beat this reference forecast. Therefore it is important to make a good choice which reference forecast will be used. The most useful and honest reference forecast to use for evaluation of the forecasts is one that is difficult to beat (Pappenberger et al., 2015). Often a reference of hydrological climatology or persistence is used to analyse the added value of a hydrological forecast(Pappenberger et al., 2015; Bennett et al., 2013). Hydrological climatology forecasts use average observed discharge values over a reference period for a specific day, while hydrologic persistency forecasts use the most recent observations. Pappenberger et al. (2015) conclude that other reference forecasts might be more difficult to beat with the consequence that a reference forecast based on hydrological climatology or persistence results in an overestimation of forecast skill. Hydrological climatology has high CRPS values because discharge is strongly auto correlated at short lead times, while hydrological persistency forecasts results in low CRPS values of the reference forecast because persistency forecasts usually are very weak for longer lead times of 1 day (Bennett et al., 2013). Bennet

et al. (2013) and Pappenberger et al. (2015) both advise to use a reference forecast of an ensemble of hydrological forecasts calculated by the hydrological model using historical meteorological observations at the same calendar day as input to the model, because this leads to the lowest CRPS values of the reference forecast. Thus the reference forecast has the same number of ensembles as the number of years in the reference period. An ensemble of precipitation and potential evapotranspiration is used here to run the hydrological model with updating and compute the reference forecast and the corresponding $CRPS_{reference}$ for the same validation period as for the forecast from 1 January 2009 to 15 October 2013. The reference forecast consists of 10 ensemble members, because the period of reference is from 1981 to 1990.

The CRPSS will be used in this study to evaluate the added value of the forecasts over the reference forecast.

3.3.4.3 Brier Score

The Brier score is a relatively simple and often used evaluation method for probabilistic forecasts and is one of the oldest evaluation tools in use (Hersbach, 2000). The Brier score takes instead of the CRPS with an infinite number of classes, each with zero width, only two classes into account. The Brier score is essentially the mean squared error of the probabilistic forecast and evaluates the accuracy of the probabilistic forecast at a chosen threshold (Wilks, 2006). Usually, Brier scores are evaluated for different threshold levels. The Brier score has the attractive property that it can be decomposed into a reliability, a resolution, and an uncertainty part (Tödter & Ahrens, 2012). The reliability is the mean conditional bias and quantifies the calibration of the forecast system. The resolution measures the average distance of all climatological probability of observed occurrence to the unconditional probability of occurrence. Here large differences are preferable. The last term, uncertainty, measures the variance of the observations. It is only dependent on observed data and therefore for each lead time the same. Brier scores thus should not be compared on different samples. With the decomposition of the Brier score you can obtain a detailed insight into the performance of the forecast system with respect to the event under consideration (Hersbach, 2000). The Brier score has the advantage that it can be applied to both deterministic and probabilistic forecasts, without the need to transform the ensemble forecast into a deterministic one (e.g. by considering the median only) (Addor et al., 2011).The Brier score is defined as:

$$BS = \frac{1}{n}\sum_{k=1}^{n}(f_k - o_k)^2$$

where the index k denotes a numbering of n forecasts, f_k is the forecasting probability that was forecasted and o_k is the actual outcome of the event at instance k ($o_k=1$ if the event occurs, otherwise $o_k=0$) (Wilks, 2006).

When decomposed into the three components: reliability, resolution and uncertainty the Brier score is defined as:

$$BS = REL - RES + UNC = \frac{1}{n}\sum_{k=1}^{n}(f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{n} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o})$$

where $\bar{o} = \sum_{k=1}^{n} o/n$ is the observed climatological base rate for the event to occur, $\bar{o}_k$ is the observed frequency of occurrence of events at instance *k*, given the forecasting probability f_k.

The Brier score is negatively oriented, so it assigns better forecasts lower scores. The score can only take on values in the range 0 ≤ BS ≤ 1, since individual forecasts and observations are both bounded by zero and one (Wilks, 2006).

3.3.4.4. Reliability diagram

The reliability diagram and sharpness diagram are commonly used for the evaluation of probabilistic forecasts. They measure reliability, resolution and sharpness (three important attributes of a good probability forecast system).

Reliability diagram

The reliability diagram consists of a plot of the observed relative frequency against the predicted probability of the forecasts. In other words, it measures how closely the forecast probability of an event corresponds to the actual chance of observing the event. The reliability diagram provides a quick visual inter-comparison between for example different thresholds or different forecasts (Bröcker & Smith, 2007) and can evaluate the reliability and the resolution of a forecast.

In the evaluation of the reliability diagram the observed relative frequency of exceeding a threshold is compared with the predicted probability of exceeding a threshold over a set of probability bins. The diagonal line indicates perfect reliability (in this case the average observed frequency is equal to the predicted probability for each category). The horizontal deviation of a forecast from this line indicates the "conditional bias". If the curve of the forecast lies below the diagonal, the probabilities are too high and this indicates overforecasting; if the curve of the forecast lies above the diagonal this indicates underforecasting and the probabilities are too low (Figure 13 left). The vertical deviation from the horizontal climatology line in the reliability diagram indicates the resolution of the forecast (Figure 13 right). So the flatter the curve in the reliability diagram the less resolution a forecast has.

The set of probability bins that are represented in the reliability diagrams are chosen as: 0%-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100%. In many studies, the forecast probabilities are plotted at the center of each bin in the reliability diagram (e.g. Olsson & Lindström, 2008) However, Bröcker & Smith (2007) showed that the reliability of the reliability diagram increases when the forecast probabilities in a probability bin are plotted at the average probability of all forecasts in that bin. Plotting it in the usual way may lead to deviations from the diagonal that do not relate to simulation errors or meteorological errors. In this study therefore the approach of Bröcker & Smith (2007) is used. The events that are considered for the reliability diagrams are the same as for the Brier Score, which are Q75, Q85 and Q95.

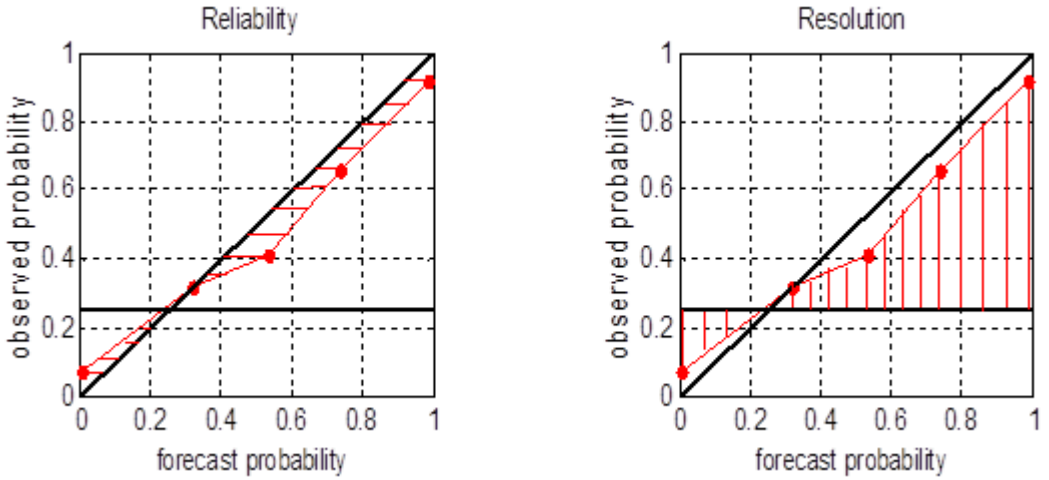

Figure 13: Example of a reliability diagram with the deviation from the reliability line to measure the reliability (left) and the deviation from the climatology line to measure the resolution (right)

Sharpness diagram

A reliability diagram is often seen in combination with a sharpness diagram of the forecast, to which it is closely connected. The sharpness diagram shows the relative frequency of the forecasts exceeding a certain threshold over the reference period in each probability bin of the set of probability classes of the reliability diagram (0%-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100%). As defined before, forecast systems that are able to predict events with probabilities different from the observed event frequency are said to have 'sharpness'. Forecast systems with little sharpness would have a frequency peak near the climatological frequency and therefore indicate that the most of the forecasts predict the event with a probability near the climatological frequency. These forecast systems therefore offer little value for planning purposes over simple use of the observed climatology.

The reliability diagram and sharpness diagram are applied in this study next to the other evaluation measures to compare the forecasts qualitatively and to give insight in the sharpness of the forecasts, as sharpness is described as one of the three important attributes of a good probability forecast system.

3.4. Combination methods for grand ensembles

As described before ensemble forecasts aim to give a better representation of the uncertainty arising from the initial conditions and the forecast model. A multi-model ensemble (grand ensemble) has the aim to account for the errors in both the initial conditions and the forecast model by combining ensembles from different centres (Johnson & Swinbank, 2009). This results in an increase in ensemble size and a combination of different analyses, perturbation generation methods to construct the ensemble members and forecast models. It is expected that the different biases of the different single model ensembles would partly cancel each other out in the multi-model ensemble mean. The multi-model ensemble is expected to cancel out the random, unpredictable components of the errors in the model that cannot be predicted using bias correction. Increasing the ensemble size generally improves the forecast skill a little (Khan et al., 2014). However, grand control ensembles combining the control forecasts from all the models results in corresponding performance as a grand ensemble even though

being much smaller in size. This implies that the combination is more important than the increase in ensemble size.

Even though a multi-model ensemble combines the strengths from different models, some model might perform better than others at different times and for different catchments (Johnson & Swinbank, 2009). Therefore further improvements of the multi-model ensemble might be realized by giving the models different weights. In the introduction is described that previous research has shown that model-dependent weights can give improvement, but that caution should be taken in how they are calculated and used (e.g. Raftery et al., 2005; Stefanova & Krishnamurti, 2002). For example, Raftery et al. (2005) used Bayesian Model Averaging (BMA) model-dependent weights to construct their grand ensemble and concluded that they created a better deterministic forecast. Section 3.4.1 describes the combination procedure and in section 3.4.2 the weighting methods used in this study are described and supported.

3.4.1. Combination procedure

The combination procedure can be described with the use of the law of total probability (Raftery et al., 2005). Following Raftery et al. (2005) the multi-model probability density function (PDF) of the variable x is then given by an average of the PDFs from the single-models:

$$p(x) = \sum_{k=1}^{M} p(x|M_k)p(M_k)$$

where $p(x|M_k)$ is the pdf based on model $\mathcal{M}_k$ and $p(M_k)$ is the probability of M_k being the best model and can be viewed as model-dependent weights ω_k. Thus, $(M_k) = \frac{\omega_k}{M}$, where $0 \leq \omega_k \leq M$ and $\sum_{k=1}^{M} \omega_k = M$.

In more detail this combination procedure can be described as follows. The multimodel ensemble is given by the union of the ensemble members from the single model ensembles for each time step and lead time. Johnson and Swinbank (2009) derived mean probabilities and ensemble mean from the multi-model ensemble combination procedure of Raftery et al. (2005) which can be used for the calculation of the Brier score, CRPS, RMSE and reliability diagrams.

The multi-model ensemble mean probability derived is:

$$p_{MM} = \frac{1}{M}\sum_{k} \omega_k p_k$$

where M is the number of models, ω_k is the weight for single model k and p_k is the single model probability given by the following formula:

$$p_k = \frac{1}{N_k}\sum_{k} o_k^i$$

N_k is the number of ensembles in single model k and $o_k^i$is the binary forecast of whether or not an event occurs, based on ensembles member i from single model k.

The formula for the single model probability is used for the calculation of the Brier score, CRPS and the reliability score for the single model performance. While, the formula of the multi-model ensemble probability is used for the calculation of the Brier score, CRPS and the reliability diagram for the different multi-model combinations.

The multi-model ensemble mean is given by:

$$\bar{x}_{MM} = \frac{1}{M}\sum_{k=1}^{M} \omega_k \bar{x}_k$$

wherein $\bar{x}_k$ is the single-model mean which is given by the following formula:

$$\bar{x}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_k^i$$

The formula for the single-model means is used for the calculation of the RMSE of the single model performance. The formula of the multi-model ensemble mean can be used for the calculation of the RMSE of the different multi-model combinations.

3.4.2. Combinations based on weights

Since the models used in this study have different numbers of ensemble members there are two basic combination methods for the single model forecasts to construct the grand ensemble (see Figure 14). The first alternative is to combine all ensemble members in the grand ensemble in a single distribution where the multi-model mean forecast and the multi-model mean probability is given by the average forecast value and the average ensemble probability of the ensemble members from all models. In this firs alternative the model with more members will get more weight. The second alternative is to combine the models in the grand ensemble. In this case the members of the multi-model are sampled from the individual distributions of the models with the result that the multi-model mean is given by the average of the single model means and the multi-model mean probability is given by the average of the single model mean probability. In this case each single model has the same influence on the multi-model ensemble independent of their ensemble member size.

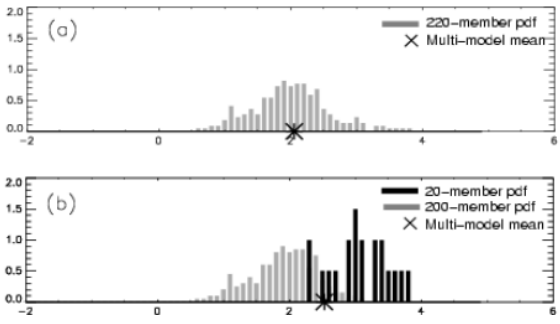

Figure 14 Example of two alternative combination methods of two models. Model 1 has 20 members and model 2 has 200 members. In (a) all members are combined in one distribution and in (b) the multi-model is constructed by the single model means where the members of the multi-model are considered to be sampled from the single model distributions separately (Johnson & Swinbank, 2009).

As described before the weights can be viewed as the probability of a particular model being the best model. It is the aim to give more weight to the more skilful models. Johnson and Swinbank (2009) described that there are various methods to define these weights. One method to compute the weights is Bayesian model averaging (BMA), where the weights and variances are computed simultaneously with the use of maximum likelihood estimation and aims to fit the PDFs to the calibration data (Raftery et al., 2005). Another method is the much more simple skill-based method (Johnson & Swinbank, 2009). In this method the weights are dependent on a measure of forecast skill. Despite the methods simplicity, the derived weights are surprisingly similar to those derived from the more complicated multiple-regression and BMA methods, as seen in studies by Raftery et al. (2005) and Johnson (2006). Therefore the skill-based method is used in this study.

Johnson and Swinbank (2009) used the mean squared error to calculate the weights. The mean squared error is a measure of skill based on the mean of the forecast ensembles and does not take the advantages of the ensembles in account. Since CRPS is a measure of skill of the probability of the ensembles this is also an interesting measure of skill to use for the calculation of the weights. Therefore both measures of skill (RMSE and CRPS) are used in this study to calculate the weights and to compare the results of these different weighting methods for the grand ensemble.

In 3.4.1 different formulas are described for the procedure to construct a multi-model ensemble. In the formulas to calculate the ensemble means there is a weighting parameter present. The given weight to each ensemble member of the single models is defined as:

$$\omega_k = \frac{\mathrm{A_k}}{\mathrm{B_k}} * \gamma$$

with $A_k = 1 \; or \; A_k = skill\; measure_k$ and $B_k = 1 \; or \; B_k = N_k$ depending on the weighting method in this study. N_k is the number of members in model *k*, skill measure$_k$ is the used skill measure to make a skill based combination. γ is a normalization factor to ensure that $\frac{1}{M}\sum \omega_k = 1$, where M is the number of models used in the combination.

With the weighting formula it is possible to construct different set ups to create grand ensembles. In this study 6 different grand ensembles are constructed and evaluated. Table 4 shows the different grand ensemble set ups and their weighting formula used to calculate the weights given to the single model ensemble members. The first two set ups are the two simple combination methods: combination weighted by members (GE1) and combination weighted by the models (GE2). In GE1 the models with more ensemble members are more weighted in the grand ensemble, while in GE2 all models in the grand ensemble get the same weight and therefore influence. The third and fourth grand ensemble, GE3 and GE4, are a weighted combination based on the differences in the CRPS and RMSE values of the single models respectively. The fifth and sixth grand ensemble, GE5 and GE6, are a weighted combination based on the differences in both ensemble member size and the differences in the CRPS and RMSE values of the single models respectively.

Table 4 Grand ensemble set ups used in this study together with their weighting formula used to calculate the weights given to the single model ensembles.

Grand Ensemble abbreviation	Grand ensemble combination method	Weight given to single model ensemble members
GE1	Combination of all members	$\omega_k = 1$
GE2	Combination of the models	$\omega_k = \frac{1}{\mathrm{N_k}} * \gamma$
GE3	Combination of the models (weights based on CRPS)	$\omega_k = \frac{\mathrm{CRPS_k}}{\mathrm{N_k}} * \gamma$
GE4	Combination of the models (weights based on RMSE)	$\omega_k = \frac{\mathrm{RMSE_k}}{\mathrm{N_k}} * \gamma$
GE5	Combination of the members (weights based on CRPS)	$\omega_k = \mathrm{CRPS_k} * \gamma$
GE6	Combination of the members (weights based on RMSE)	$\omega_k = \mathrm{RMSE_k} * \gamma$

4. Results

The results of this study are divided in five sections. First an example of a simple graphical visualization of the ensemble precipitation forecasts and ensemble discharge forecasts of ECMWF for a typical heavy precipitation period and precipitation peak and a typical high runoff period and runoff peak is given. The second section describes the results of the bias correction of the precipitation forecasts of the EPSs of ECMWF, CMA, UKMO and NCEP. The third section describes results from the hydrological updating procedure used. The fourth section presents the evaluation results of the single model hydrological forecasts. The last section presents the evaluation results of the grand ensemble hydrological forecasts.

4.1. Pluviographs and hydrographs

Figure 15 shows the observed precipitation and observed discharge from 4-5-2010 to 23-7-2010. It shows that the catchment area is quickly reacting to a rainfall event with a lag time of approximately one day between the rainfall peak (middle of the bar) and the discharge peak.

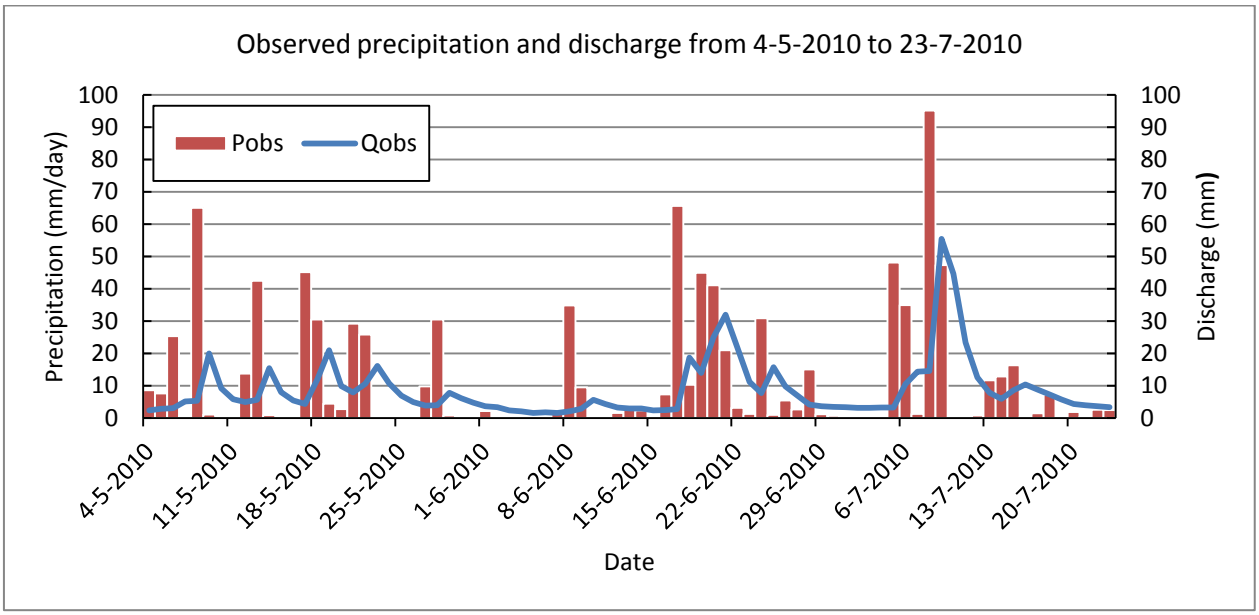

Figure 15 Observed precipitation and discharge from 4-5-2010 to 23-7-2010 showing the time lag between rainfall and discharge.

Figure 16 and Figure 17 show examples of pluviographs and hydrographs of a typical heavy rainfall and high runoff period and a typical peak precipitation and discharge peak for the observed precipitation and discharge respectively and the precipitation forecast of the ECMWF ensemble and the hydrologic forecast of the ECMWF ensemble respectively. Both figures show that the ensembles sometimes overpredict and sometimes underpredict the precipitation and the discharge. However, the observation almost always falls in the range of the ensembles. The ensemble forecast of the typical peak is an example of a good forecast where the observation falls in the range of the ensembles. The time difference between the precipitation peak and the peak discharge of one day is also visible here when Figure 16 and Figure 17 are compared.

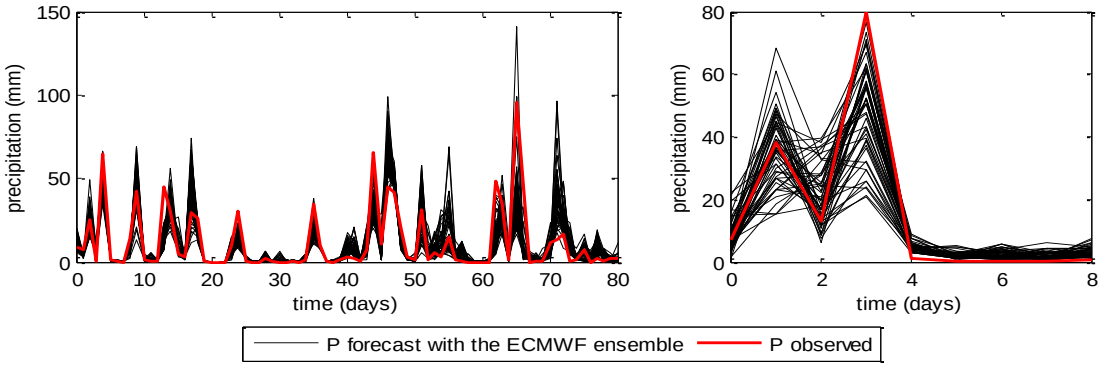

Figure 16 Pluviograph of a typical heavy rainfall period and a typical precipitation peak for a lead time of 1 day. The black lines are the 51 ensembles from the ECMWF precipitation forecast and the red line the observation.

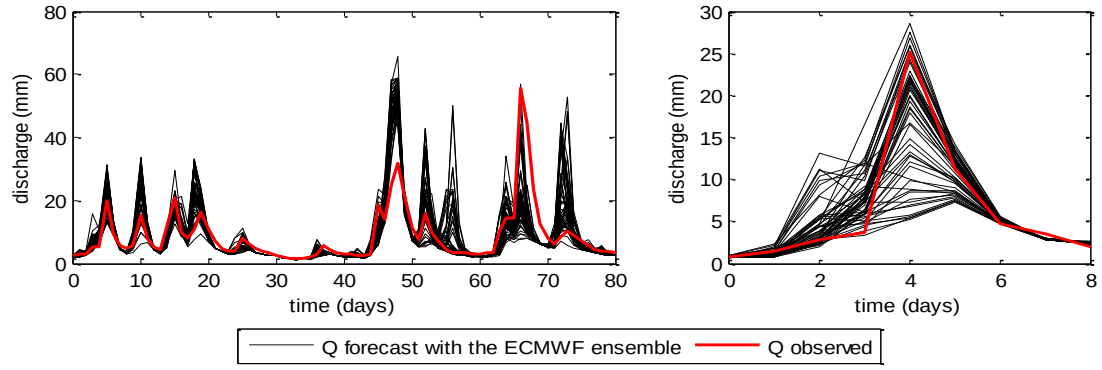

Figure 17 Hydrograph of a typical high runoff period and a typical runoff peak for a lead time of 1 day. The black lines are the 51 ensembles from the ECMWF flow forecast and the red line the observation.

4.2. Results bias correction

This section shows the results of bias correction of the single model ensemble forecasts of ECMWF, CMA, UKMO and NCEP. The bias correction is evaluated with the mean bias, correlation, CDF diagrams and the Brier score to evaluate reliability and resolution of the ensemble forecasts.

4.2.1. Bias

Figure 18 shows the biases of the raw mean ensemble areally averaged precipitation forecasts for the four models relative to the areally averaged daily observations for the Quzhou catchment for different forecast lead times. The areally averaged daily precipitation is 5.5 mm. The biases of the raw mean ensemble of all EPSs are quite significant and do not have a strong relation to lead time. The biases of CMA are the largest and for NCEP the smallest. The bias of UKMO is large for small lead times. The biases of the bias corrected mean ensemble forecasts are not shown here, since they are very close to zero. Normally the quantile mapping bias correction method would lead to biases of zero relative to the reference of the bias method, in this case the areally averaged daily observations. However, the bias is close to zero. This is the result of the quantile mapping method used here. The corrected values are an approximation of the corresponding empirical CDFs, because the forecasted empirical CDF have more values than the observed empirical CDF. The steps of the quantiles of the observed empirical CDF are smaller than the steps of the forecasted empirical CDF. As a result the corrected values can deviate a little from the value that it should have been corrected to.

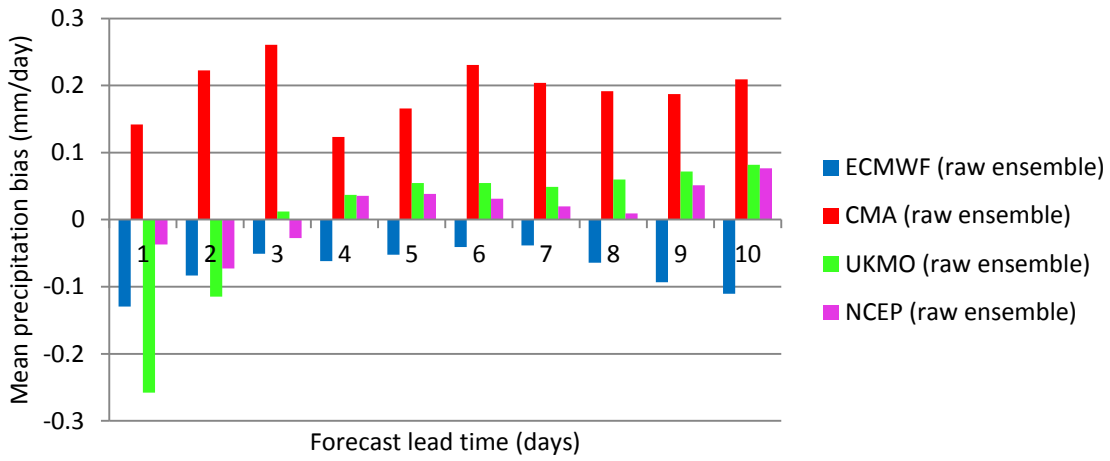

Figure 18 Mean daily bias of the raw ensemble precipitation forecast of ECMWF, CMA, UKMO and NCEP over the validation period compared to the average precipitation of 5.5 mm.

4.2.2. Correlation

Figure 19 shows the correlation coefficients between the mean ensemble forecasts of both raw forecasts and bias corrected forecasts and the corresponding areally averaged daily observations for the time series above the Q75 threshold. The figure shows that the correlation coefficients of the raw forecasts and the bias corrected mean ensemble forecasts are very similar. Though, the correlation coefficient is a little improved for the bias corrected results. ECMWF has the best correlation with the observations, while CMA has the worst correlation. The correlation, however, is significant with values above 0.6 in some cases. Figure 19 also shows that the correlation coefficient is strongly related to the lead time. As expected, shorter lead times show a higher correlation. Tao et al. (2014) consider a correlation of 0.3 still meaningful. Lead times of 8 days thus still show meaningful correlation for ECMWF, UKMO and NCEP and for ECMWF even 10 days with values close to 0.4.

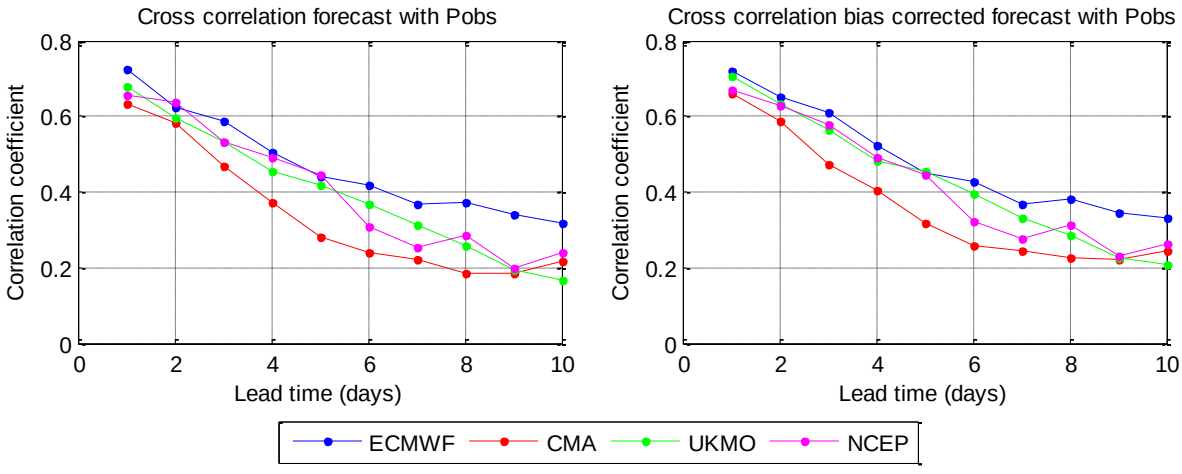

Figure 19 Cross correlation of the single model ensemble forecasts of ECMWF, CMA, UKMO and NCEP with the observed precipitation (raw forecasts left and the bias corrected forecasts right).

4.2.3. CDFs of forecasts

Figure 20 shows the empirical CDFs for lead times 1, 3, 5 and 10 days of both the observed precipitation and the EPSs precipitation forecast of the four EPSs over the whole evaluation period. The left part of the figure shows the CDFs based on the ensembles of the raw precipitation forecasts of the models and the right part of the figure shows the CDFs based on the ensembles of the bias corrected precipitation forecasts.

From the CDFs in the figures can be seen whether the quantile mapping bias correction method leads to improvements of the forecasts and if there is an improvement in the frequency distribution of precipitation. Figure 20 shows that the CDFs of the raw forecasts differ significantly from the observed precipitation CDFs. All EPSs forecast too high extreme values for all lead times, so the forecasts tend to overpredict the most extreme events. The figures show that the observation has a higher non-exceedence frequency of zero precipitation. This is the drizzle of the meteorological forecasts. Drizzling is the effect that really low values of precipitations are present. The quantile mapping approach automatically corrects the drizzle, because when the zero precipitation probability of the observation is higher than that of the forecast, the forecasts quantiles below the zero precipitation probability of the observation are corrected to zero. This is also visible in the figures of the bias corrected values where the drizzle has been corrected. After bias correction the dry day frequency of the forecasts is almost similar compared to the observations. As expected the bias corrected forecast CDFs to the right of Figure 20 are almost similar to the CDFs of the observations. The maximum precipitation values are bias corrected to a value close to the maximum precipitation value of the observed precipitation, however this is not good visible in these graphs.

4.2.4. Brier score

Figure 21 shows the Brier score and the decomposition of the Brier score in a reliability score and a resolution score component for both the raw precipitation forecasts and the bias corrected precipitation forecasts of the different EPSs exceeding the P75, P85 and P95 thresholds.

The left side of the figure shows the graphs of the Brier score. In general all Brier scores increase with increasing lead time, with the exception of lead time 7 and 8 in some cases. So the quality of the forecast decreases with increasing lead time since the Brier score assigns lower scores to better forecasts. Also notable is that the Brier score decreases with increasing threshold. The match between the forecast and the reference is better in this case because there is more zero probability of exceeding the threshold for both forecast and reference.

Figure 21 shows that the Brier scores for all EPSs and the three thresholds for almost every lead time have improved (closer to zero) if the quantile mapping bias correction is applied. Only the Brier scores of the raw EPS of NCEP at lead times of 5, 6 and 7 days with threshold P95 are not improved by the bias correction method and even worsened. However this is just a minor difference. Overall can be said that the improvement due to bias correction of the EPSs are highest for threshold P75 and lowest for threshold P95. CMA shows the highest improvement, NCEP second highest and ECMWF shows the lowest improvement in the Brier score.

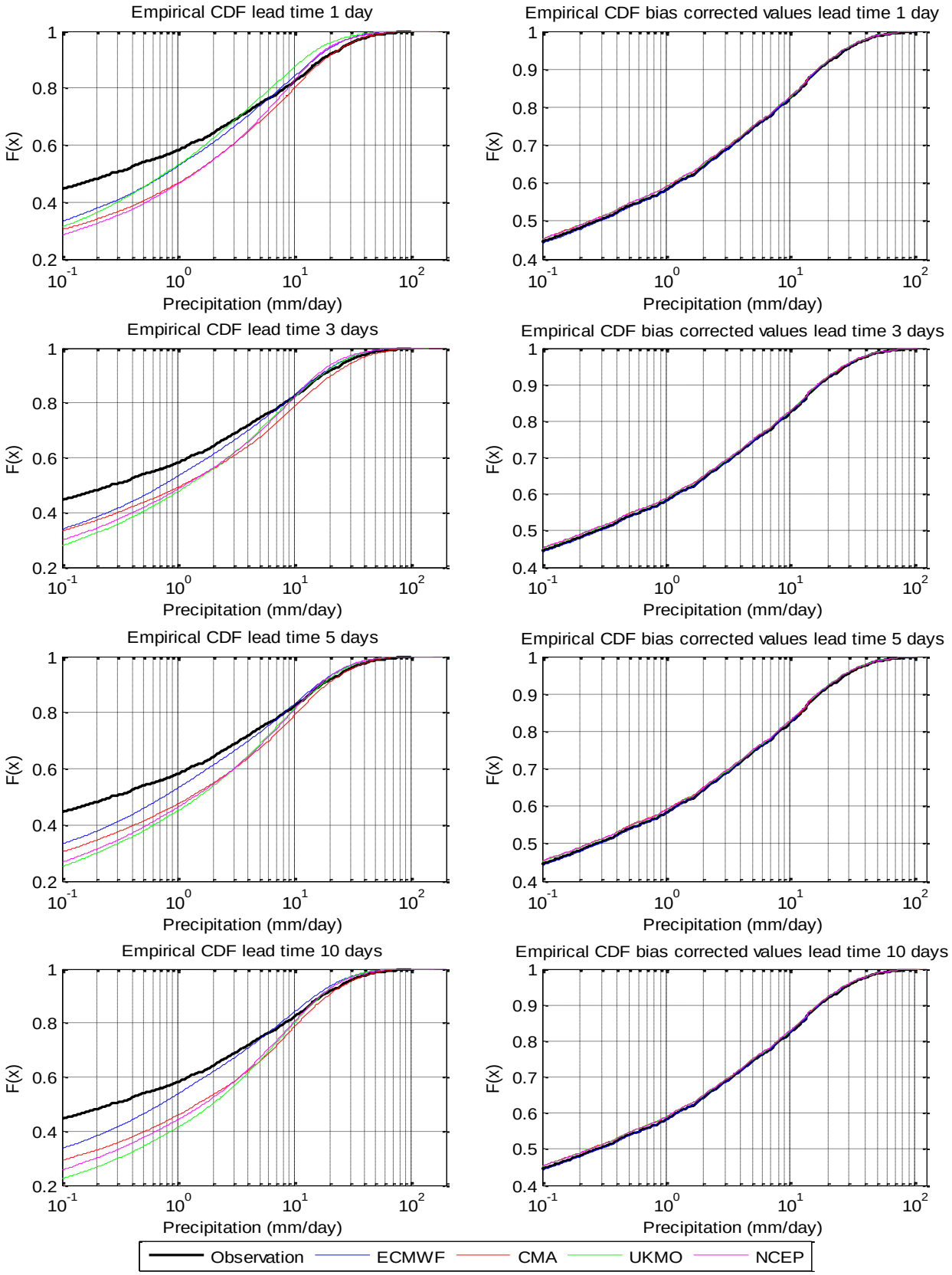

Figure 20 Empirical CDFs for lead times 1, 3, 5 and 10 days of the observed precipitation and the ensemble precipitation forecast of ECMWF, CMA and UKMO of raw forecasts (left) and bias corrected forecasts (right).

The middle of Figure 21 shows the reliability (conditional bias) for the raw and bias corrected EPSs. Here we can see that the ECMWF EPS is best performing while CMA is the worst performing. The reliability scores also show that the reliability for all three thresholds, all EPSs and for almost every lead time have improved when the bias correction is applied. The reliability score has not improved for some lead times of the forecasts of ECMWF with a threshold of P85 and P95 and for the forecasts of UKMO and NCEP with a threshold of P95. However, this is just a minor difference in comparison with the improvements of the reliability score in general. Because the reliability score is related to the Brier score the same tendency as for the Brier score is visible: the improvements due to bias correction of the EPSs are highest for threshold P75 and lowest for threshold P95, where the bias correction even has some negative influence on the reliability for some EPSs and lead times. Here CMA also shows the highest improvement, NCEP second highest and ECMWF shows the lowest improvement in the reliability score.

The right side of the figure shows the resolution (measure of difference from climatology) for the raw and bias corrected EPSs. In this case is a high score is better. The graphs for P75, P85 and P95 also show that ECMWF is best performing and CMA worst. The performance of UKMO is close to ECMWF, while the performance of NCEP is close to that of CMA. The tendency visible here is that the bias correction of the EPSs results in only a minor improvement of the resolution score in comparison with the improvement of reliability.

It is difficult to say something about the performance in general based on the Brier score, because higher thresholds are assigned better scores.

4.2.5. Conclusions

The quantile mapping bias correction used in this study results in improved forecasts. Cross correlation of the forecasts with observations results in relatively high correlation coefficients and shows that forecasts with lead times over 8 days can still be meaningful. The frequency distribution has improved as well. After bias correction there is a good match between the CDFs of the forecasts and the observations. Differences in dry day frequencies are corrected and the most extreme events are closer to the observations. From the evaluation of the ensemble models with the Brier score can be concluded that the reliability and resolution has improved for most lead times and threshold values.

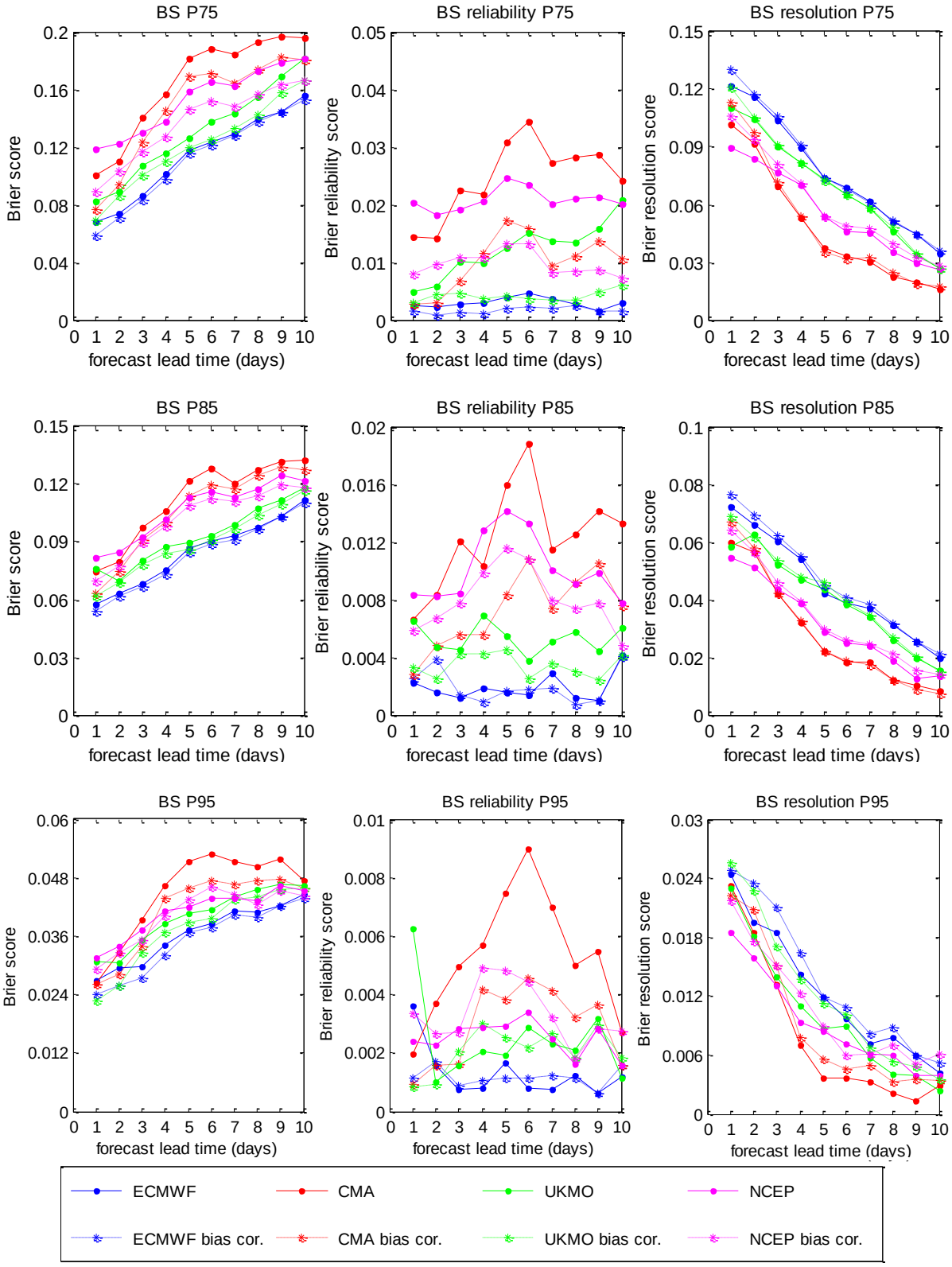

Figure 21 Brier score and decomposition into reliability and resolution of the raw and bias corrected EPSs of ECMWF, CMA, UKMO and NCEP for lead times 1-10 and thresholds Q75, Q85 and Q95.

4.3. Hydrological updating procedure

The updating procedure is tested with a so called 'perfect forecast'. Measured daily precipitation data and discharge data of the period 2009-2013 for the Quzhou river basin are used as input of the model with updating procedure. In general this resulted in an improvement of the Nash-Sutcliffe efficiency (NS) coefficient and the Relative Volume Error (RVE). These evaluation measures are calculated for a 'perfect forecast' with the implementation of the updating procedure described in chapter 3. For other lead times these values differ. Table 5 shows the results for different lead times of a 'perfect forecast' with and without the implementation of hydrological updating. Note that without updating the performance is not shown for different lead times, because without updating the lead times use their own initial states calculated from the output of the day before the issue date.

The NS value without the use of the updating procedure is 0.91. With the use of the updating procedure the NS value improves to a value of 0.94 for a lead time of 1 day. For high flows above the Q75 threshold the improvement of the NS value is similar (from 0.89 to 0.92). The RVE improves from a percentage of -3.03 % without hydrological updating to a percentage of -0.37 % with the implementation of hydrological updating for a lead time of 1 day. The improvement for high flows is from a percentage of 7.72 % without updating to a percentage of 3.23 % for a lead time of 1 day with updating.

Table 5 NS and RVE for simulated discharges based on observed meteorology and various lead times using updated initial states for a lead time of 1 day and the case where the initial states are not updated.

leadtime	*All flows*		*High flows > Q75*	
	Without updating			
	NS (-)	**RVE (%)**	**NS (-)**	**RVE (%)**
	0.91	-3.03	0.89	7.72
	With updating			
	NS (-)	**RVE (%)**	**NS (-)**	**RVE (%)**
1	0.94	-0.37	0.92	3.23
2	0.94	-0.39	0.92	3.24
3	0.93	-0.02	0.90	7.40
4	0.92	-0.51	0.90	8.37
5	0.92	-1.08	0.89	8.55
6	0.92	-1.55	0.89	8.35
7	0.92	-1.97	0.89	7.95
8	0.92	-2.30	0.89	7.81
9	0.92	-2.44	0.89	7.80
10	0.91	-2.52	0.89	7.77

The NS and RVE for the lead times longer than 1 day are improved compared to the forecast without using the updating procedure for the case where all flows are included. For high flows above the Q75 threshold the NS of the forecasts with updating is improved to the forecast without updating for lead times 1-4 days. Longer lead times have similar NS values as the forecast without updating. The RVE of high flows is improved for lead times 1-3 days for forecasts with updating and a little worsened for longer lead times compared to the forecast without updating.

4.4. Results single model forecast

This section presents the results of the single model hydrological forecasts derived with the GR4J model and the four bias corrected precipitation forecasts of the EPSs ECMWF, CMA, UKMO and NCEP as input. The evaluation period is from the 1st of January 2009 to the 29th of October 2013. The single model hydrological forecasts are first evaluated on the mean ensemble forecasts with the root mean square error and after that evaluated on the ensemble forecasts with respectively the CRPSS, contribution of meteorological and hydrological model error, Brier score and reliability diagrams. Weights are derived for the grand ensemble ensembles based on the RMSE and CRPS of the forecasts.

4.4.1. Mean ensemble evaluation

4.4.1.1. Root mean square error

Figure 22 shows the RMSE for the four EPSs ECMWF, CMA, UKMO and NCEP for lead times 1-10 days. The RMSE is calculated for the mean of the forecasts for the days when the observation exceeds the Q75 threshold. The mean discharge of these high flows exceeding Q75 is 9.47 mm. The mean ensemble of ECMWF shows the best results and CMA the worst. However, for lead times above 8 days UKMO is performing worst. Generally UKMO is performing better than NCEP for lead times up to six days, while for longer lead times NCEP is performing better. However, the RMSE of the ensemble forecast is still high, because of 3 days in a window of 5 days in the forecast have an average error of 36.5 mm. Without these 3 errors the RMSE improves significantly with approximately 1.5 to 2 mm. The dotted graphs in Figure 22 show the forecasts without the 3 errors. It is expected that these errors are not caused by observation errors, because in this period both precipitation and discharge show a peak. Generally there is an expected trend that the RMSE increases with increasing lead time and a high increase in errors for small lead times. This event with 3 large errors is underpredicted by all EPSs.

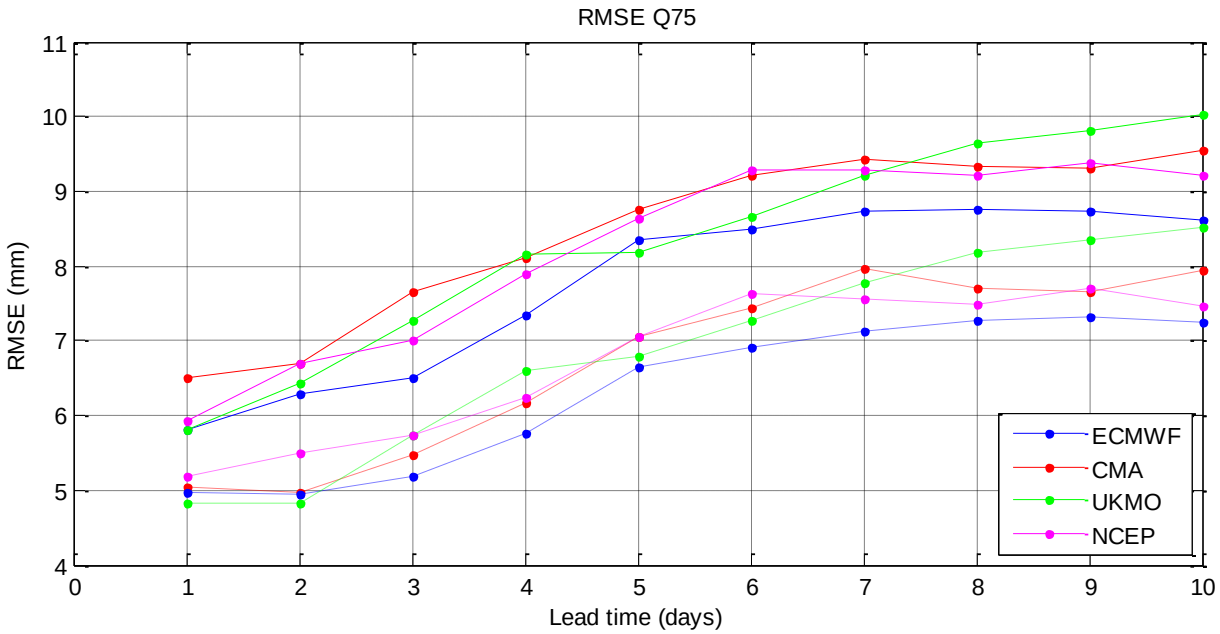

Figure 22 RMSE of the hydrological forecasts using ECMWF, CMA, UKMO and NCEP and exceeding the Q75 threshold for lead times 1-10 days. Dotted graphs show the RMSE of the forecasts with the exclusion of the 3 largest errors.

4.4.2. Evaluation of ensemble forecasts

4.4.2.1. Relative contribution of meteorological and hydrological model errors

Figure 23a show the CRPSs of the hydrologic ensemble forecasts of the 4 EPSs evaluated to the deterministic observed discharge exceeding the Q75 threshold. It shows that the CRPS increases with lead time for all EPSs and that the score flattens with increasing lead time comparable with the RMSE. However there are some minor differences visible. CMA is performing worst here for lead times longer than 5 days, while UKMO is performing worst for the RMSE for the longest lead times. These differences between the RMSE and CRPS occur due to the resolution component which is not taken into account in the RMSE. When taking resolution into account, CMA and NCEP clearly show less performance among the 4 EPSs. The CRPS scores presented here are used to calculate the weights for the grand ensembles based on the CRPS score.

Figure 23b presents the relative contribution of the meteorological forecast errors and hydrological model errors. A high percentage suggests that meteorological errors are more important while low percentages suggest that hydrological model errors are more dominant in the forecast. For all EPSs it is shown that with increasing lead time the meteorological forecast errors increase relative to the hydrological model error despite the increasing model error (presented in 4.3 Table 5). This is caused by an increasing error of the meteorological forecast with increasing lead time. Figure 23 also shows that the graphs flatten with increasing lead time especially for CMA and NCEP. This is caused by the effect that both the meteorological forecast and hydrological model errors are more or less stable for long lead times of 5-10 days (see Figure 21 and Table 5).

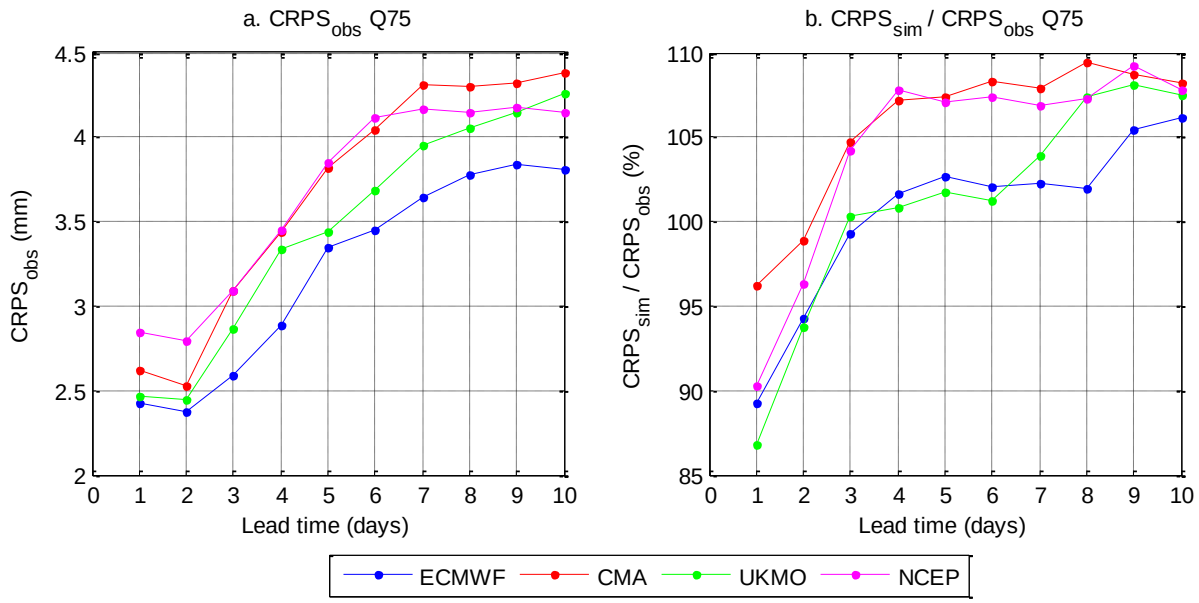

Figure 23 CRPS of the hydrological forecasts against discharge observations (a) and the relative contribution of meteorological errors determined by CRPS$_{sim}$ compared to meteorological + model errors determined by CRPS$_{obs}$ (b).

4.4.2.2. Continuous ranked probability skill score

The mean performance of the hydrological ensemble forecasts exceeding the Q75 threshold from the four meteorological centres is evaluated by the CRPSS. 10 years precipitation climatology is used to force the hydrological model and construct a 10 member hydrological ensemble to use as the reference forecast to calculate the CRPS$_{ref}$ from where the CRPSS is derived for the hydrological forecasts. Figure 24 shows the CRPSS of the hydrological forecasts for the four different EPSs for exceeding the Q75 threshold for lead times 1-10 days. It shows the relative performance of the different forecast compared to the performance of the reference forecast. The results are similar to the results before. The CRPSS also shows that the ECMWF EPS has the best skill, UKMO the second best skill and CMA and NCEP the worst. Since the CRPSS is a relative score of the CRPS of the hydrologic forecast to the CRPS of the reference forecast this figure also shows whether the forecasts show skill in comparison with the reference forecast. Figure 24 indicates that all EPSs have skill over the reference forecast which is an ensemble of discharges simulated with past observations of precipitation for all lead times. The CRPSSs are clearly above a value of zero. However, it should be noted that it is difficult to capture high events from historical observations, because high flow periods are often short and doesn't have to occur on specific days.

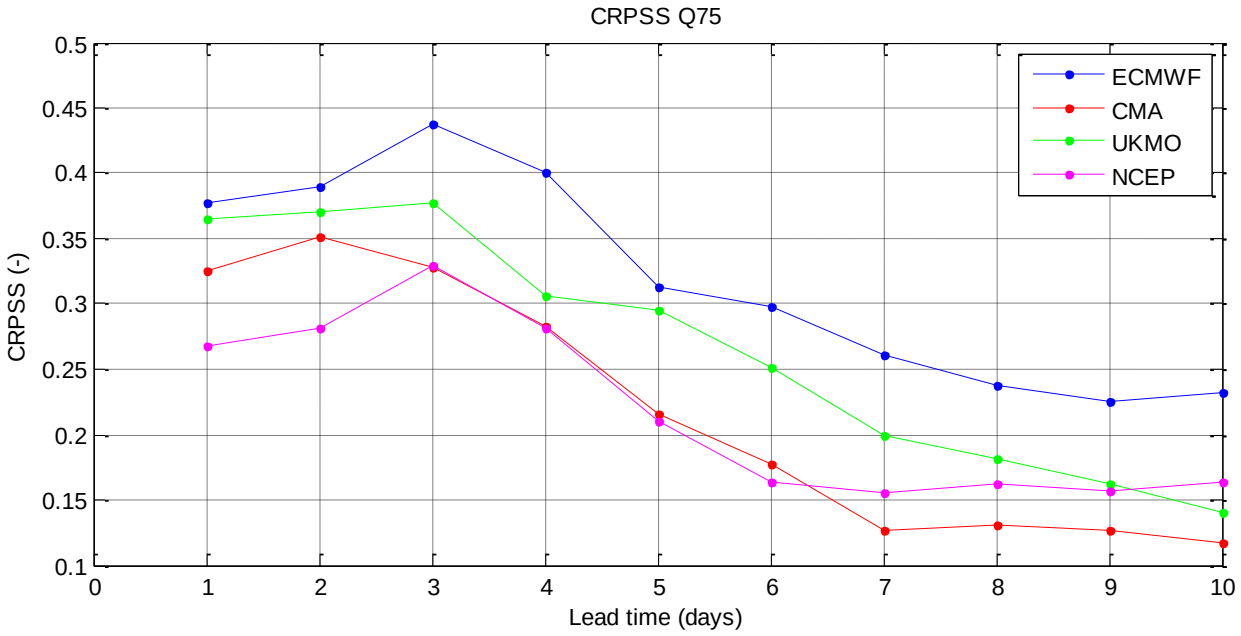

Figure 24 CRPSS, calculated by the CRPS of the hydrological forecasts of ECMWF, CMA, UKMO and NCEP compared to the CRPS of the reference forecast exceeding the Q75 threshold for lead times 1-10 days.

4.4.2.3. Brier Score

Figure 25 shows the Brier score and the decomposition of the Brier score into a reliability score and a resolution score component for the hydrologic forecasts, calculated with the GR4J model including updating procedure and with the bias corrected precipitation forecast of the different EPSs as input, exceeding the P75, P85 and P95 thresholds for lead times 1-10 days.

Overall, Figure 25 shows that the skill of CMA and NCEP is worst. The skill of NCEP is worst for small lead times of 1-2 days for thresholds Q85 and Q95 and 1-6 days for threshold Q75. This is both due to reliability and resolution as seen from the other graphs. ECMWF and UKMO have the best skill as a result of both good reliability and resolution. UKMO is best performing for flows above the Q75 threshold, while ECMWF is best performing for flows above the Q95 threshold. For flows above the Q85 threshold the Brier score of ECMWF is best for lead times up to 5 days, while UKMO has the best performance for lead times 6-10 days. This is the result from a high decrease in the reliability of the ECMWF for forecasts with lead times over 5 days. UKMO is for most lead times best or close to the best performing EPSs on reliability, except for the threshold Q95. For the resolution component of the Brier score the same tendency as for the Brier score is visible. This is because the reliability component is really small in comparison with the resolution component. Overall ECMWF and UKMO are performing best on resolution.

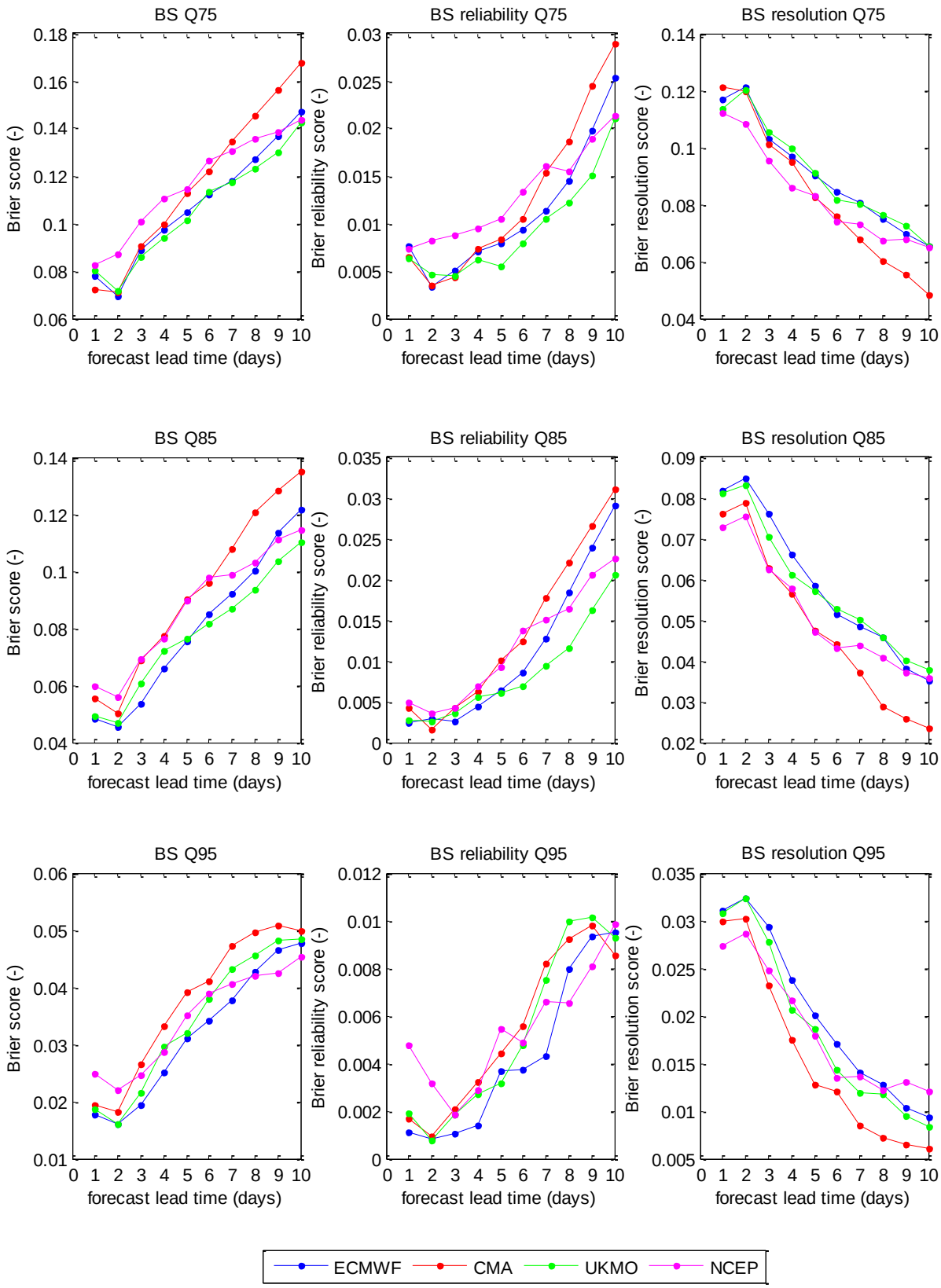

Figure 25 Brier score and decomposition into reliability and resolution of the hydrological forecasts of ECMWF, CMA, UKMO and NCEP for lead times 1-10 and thresholds Q75, Q85 and Q95.

The previous precipitation forecast Brier score results from Figure 21 are to a certain extent similar to the results of the hydrologic forecasts presented in Figure 25. The most skilful EPSs of Figure 21 are almost always the most skilful here as well and the values are in the same range. For short lead times the Brier score is often lower in comparison with the precipitation forecasts. This is probably the result from the updating procedure which improved the skill of the hydrological model especially for the short lead times (Table 5). However there is only a minor difference for lead times longer than 5 days. The Brier score of the precipitation forecast flattens out, while the Brier score of the hydrologic forecasts does not flatten out. This is probably the result of the increased forecast error for longer lead times due to the hydrological model uncertainty and the decreasing skill of the updating method for longer lead times. The graphs of the reliability score for the precipitation forecast show a decreased reliability score for lead times 5-10, while the reliability score for the hydrologic forecasts show an increased tendency for the reliability score for lead times 5-10 days. This is the result of the increasing uncertainty of the hydrological model with lead time.

4.4.2.4 Reliability diagrams

Figure 26 and Figure 27 present the reliability diagrams and the sharpness diagrams of the hydrologic forecasts of different lead times (lead time 1, 3, 5 and 10 respectively) of the 4 EPSs. Lead times between 5 and 10 days are omitted, because there is a trend visible that the deviation with the diagonal grows with increasing lead time. The reliability diagrams give insight in the reliability as well as the resolution of the hydrologic forecasts. The sharpness diagrams give insight in the sharpness (the ability to forecast extremes) of the forecasts. Next to reliability and resolution this was defined as the third important attribute of a good forecast in chapter 3.

Generally there is a trend visible that reliability and resolution decrease with increasing lead time. In the figures this can be seen by the deviation from the diagonal (reliability) and the deviation from the climatology (resolution). With increasing lead time, the graphs of the models get more deviated from the diagonal (decreasing reliability) and get closer to the climatology line (decreasing resolution). The figures also show that the lines of the models lie more often under than above the diagonal, which means that the high flows of the forecasts are mostly underpredicted. The result of the best performing model is according to the other evaluation scores. ECMWF is here also performing best. The other three models have more or less a similar overall performance, but differ in performance for different lead times and different thresholds.

The sharpness diagrams below the reliability diagrams show the relative frequency of use of the probability categories of the reliability diagrams (0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, 0.8-1.0) and give insight in the sharpness of the forecasts. In short, it shows the distribution of the 1599 forecast probabilities over the probability categories. The sharpness diagrams of the different EPSs are generally the same depending on the lead time and threshold. The frequency of use is best distributed over the different probability categories for the Q75 threshold, thereafter Q85 and worst for Q95. This is as expected, because for the Q95 threshold a value is more often forecasted below the threshold than above the threshold, with the consequence that the forecast probability of zero is more frequent than higher probabilities. For lower thresholds forecast probabilities higher than zero become more frequent, because there are more forecasts with a value above the threshold value. So it is normal that the

majority of the forecasts predict low probabilities for the threshold used here. These low probabilities are lower than the climatological probability of 25%, 15% and 5% respectively for Q75, Q85 and Q95 as can be seen by the starting point of the reliability graphs which is close to zero. The forecasts are also able to predict relatively high probabilities greater than 40% of the event in comparison with the climatologies, although such forecasts are less frequent. The forecasts here thus have sharpness. The 4 EPSs show comparable sharpness for all lead times and thresholds.

4.4.3. Conclusions

ECMWF has the most skilful hydrological ensemble forecasts with good reliability and resolution. CMA is the least skilful hydrological ensemble forecast for the evaluation period and the Quzhou river basin. When comparing the Brier scores of the precipitation forecasts with the hydrological forecasts it becomes clear that the hydrological model leads to increased errors because of the model errors. This leads to less skilful forecasts especially for longer lead times where the hydrological updating procedure has less effect. The sharpness diagrams show that the forecasts have sharpness and that ECMWF and CMA show the highest sharpness.

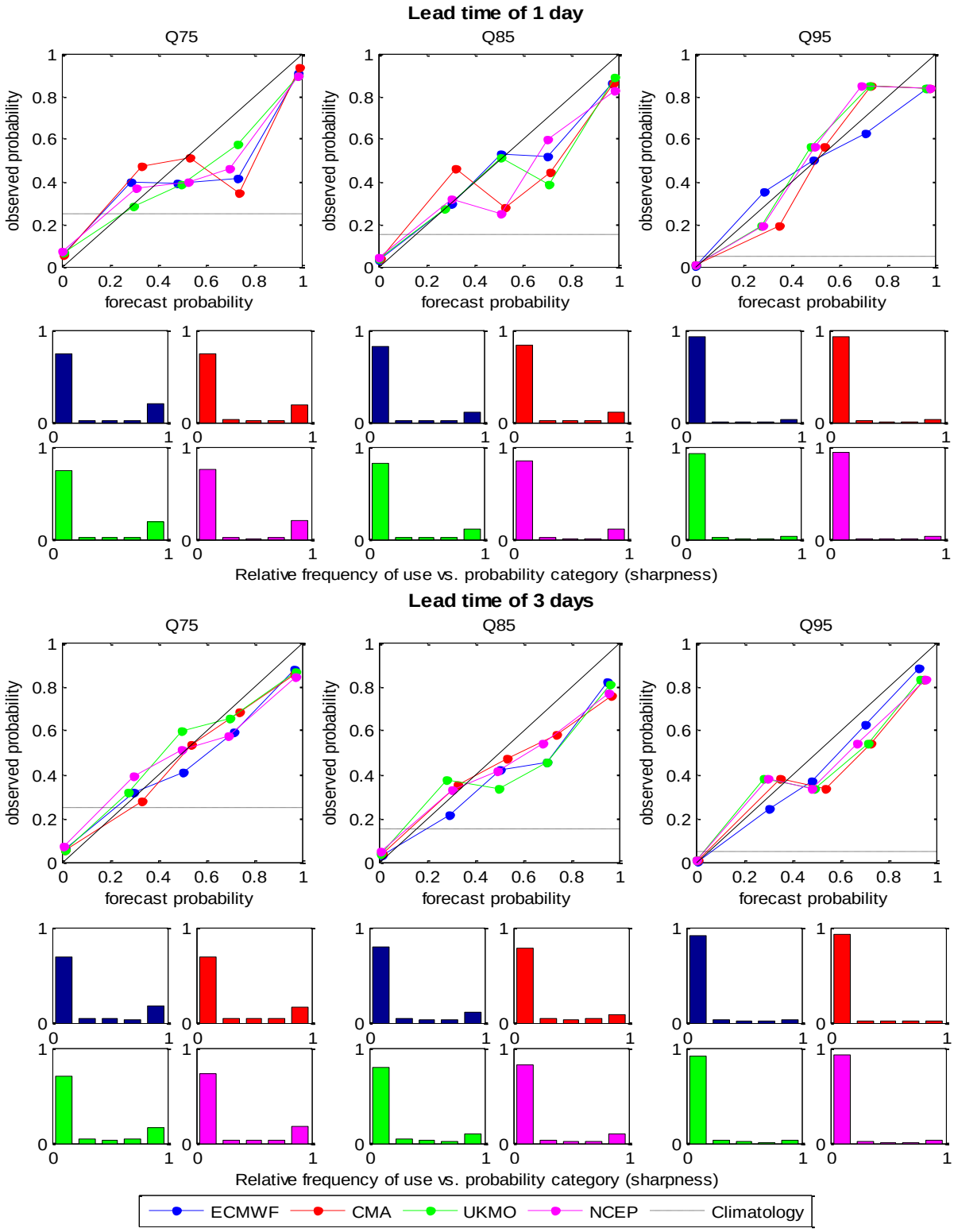

Figure 26 Reliability diagram and sharpness diagrams of the hydrologic forecasts of ECMWF, CMA, UKMO and NCEP for a lead time of 1 and 3 days and thresholds Q75, Q85 and Q95.

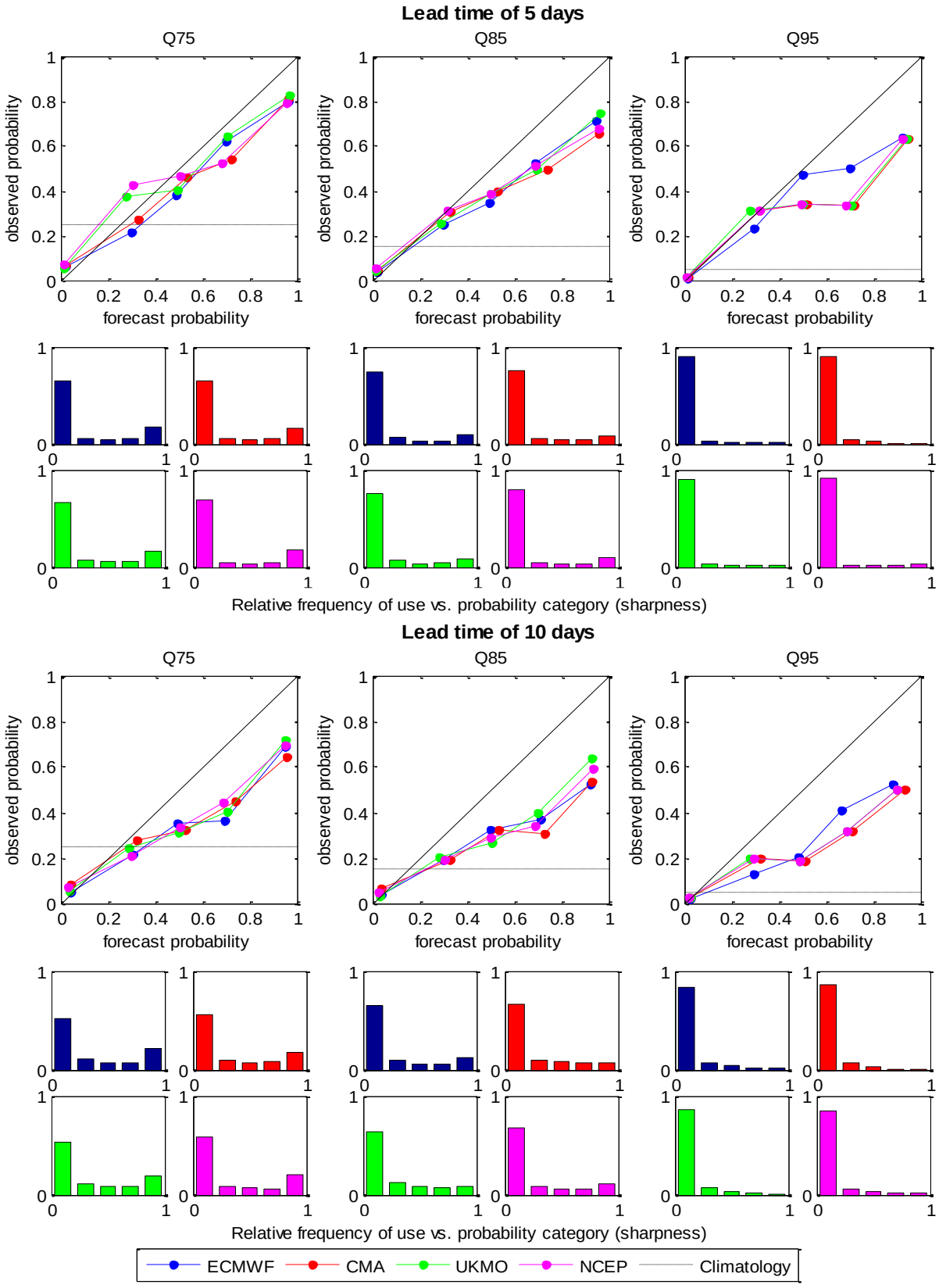

Figure 27 Reliability diagram and sharpness diagrams of the hydrologic forecasts of ECMWF, CMA, UKMO and NCEP for a lead time of 5 and 10 days and thresholds Q75, Q85 and Q95.

4.5. Results grand ensemble forecasts

This section presents the results of the grand ensemble hydrological forecasts. Six different combinations of the single model forecasts are used that result in six different grand ensemble hydrological forecasts. The first is a simple combination of the single model members. The second combination is a combination where the ensembles get weighted in such a way that the models have the same weights in a grand ensemble forecast independent of member size. The other four combination methods are a combination of the first two methods with weights based on the CRPS and the RMSE of the single model forecasts. The evaluation period is the same as for the single-model forecasts from the 1st of January 2009 to the 29th of October 2013. The grand ensemble hydrological forecasts are first evaluated on the mean ensemble forecasts with the root mean square error and after that evaluated on the ensemble forecasts with respectively the CRPS and the relative contribution of meteorological errors and hydrological errors, CRPSS, Brier score and reliability diagrams.

4.5.1. Mean ensemble evaluation

4.5.1.1 Root mean square error

Figure 28 shows the RMSE for the grand ensemble hydrological forecasts of the six different combinations and the best performing single model ECMWF for lead times 1-10 days. The RMSE is calculated for the mean of the grand ensemble forecasts of the days when the observation exceeds the Q75 threshold. The mean discharge of these high flows exceeding Q75 is 9.47 mm. As expected the RMSE decreases when using a grand ensemble forecast. All grand ensemble combinations lead to a decreased RMSE for all lead times in comparison with the EPSs. The RMSE of a lead time of 5 days is most decreased with approximately 0.7 mm. Nevertheless, the RMSE of the grand ensemble forecast is still high, because of the two days with extremely high errors. Noteworthy is the smoother line of the grand ensemble forecasts, this is because the single forecasts have different biases which cancel each other out in the multi model mean.

The RMSE of the different grand ensemble ensembles are almost similar. However there are some differences. Figure 28 shows that the combination methods where the members are combined (member size of the forecasts is included) results in a minor improvement of the RMSE in comparison with the combination method which is not affected by member size (combination models). The other combinations based on an evaluation score and the member combination or model combination do not result in a significant improvement compared to the grand ensembles without using evaluation scores to calculate weights.

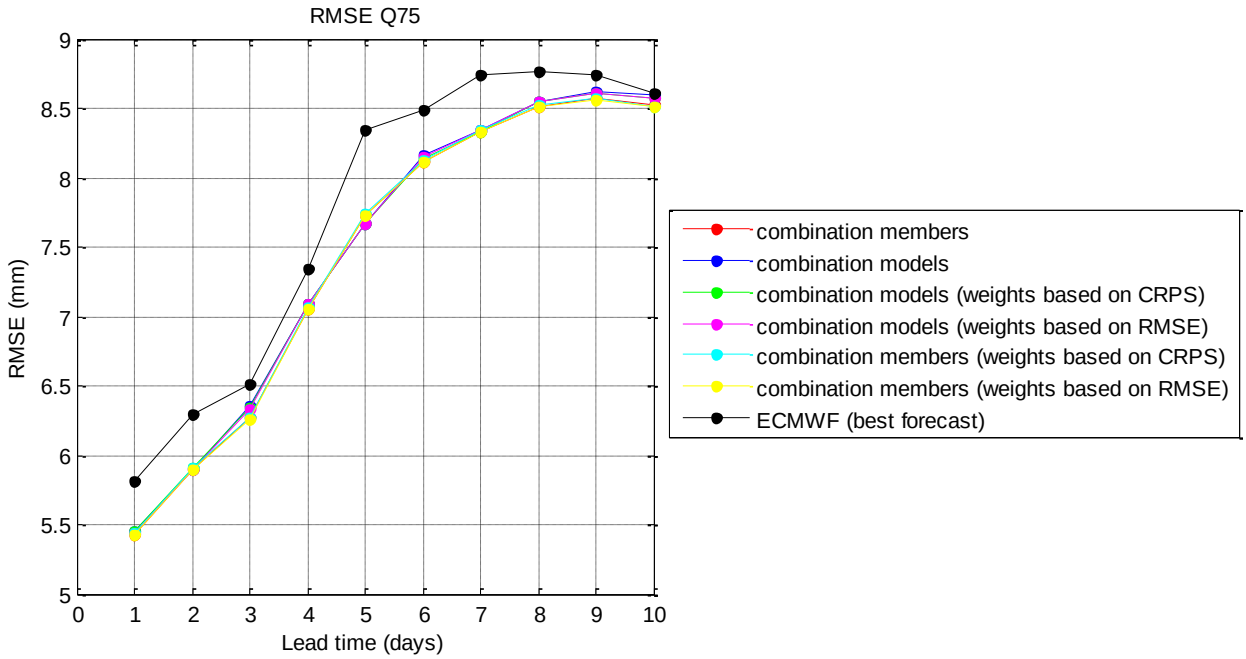

Figure 28 RMSE of the grand ensemble hydrological forecasts of the six different combinations and the best performing single model ECMWF exceeding the Q75 threshold for lead times 1-10.

4.5.2. Evaluation of ensemble forecasts

4.5.2.1 Relative contribution of meteorological and hydrological model errors

Figure 29a shows the CRPS of the hydrologic forecasts of the grand ensembles and ECMWF evaluated to the observed discharge exceeding the Q75 threshold. It shows that the CRPS results are comparable with the RMSE results and that the grand ensembles perform much better than the EPSs and that there is not much difference between the performance of the different grand ensembles.

Figure 29b presents the relative contribution of the meteorological forecast errors and hydrological model errors. As for the EPSs for all GEs it is shown that with increasing lead time the meteorological forecast errors increases relative to the hydrological model error despite the increasing model error (presented in 4.3 Table 5). In this case there is a difference visible between the grand ensembles with a combination of the models and a combination of the members. The graphs of the grand ensembles based on a combination of the members (GE1, GE5 and GE6) are below the graphs of the grand ensembles bases on a combination of the models (GE2, GE3 and GE4). For grand ensembles based on a combination of the members, the meteorological errors have less influence than for the grand ensemble combinations based on the models. This is the effect of the combination of the members where EPSs with more ensemble members are more weighted. In this study the EPSs with the most ensemble members, ECMWF and UKMO, have the highest skill and the lowest meteorological error, while the meteorological error + model errors are almost similar for all grand ensembles (Figure 29a). When comparing Figure 29b with Figure 23b it becomes visible that the grand ensemble graphs are in between all the EPSs and that they are combinations of the EPSs.

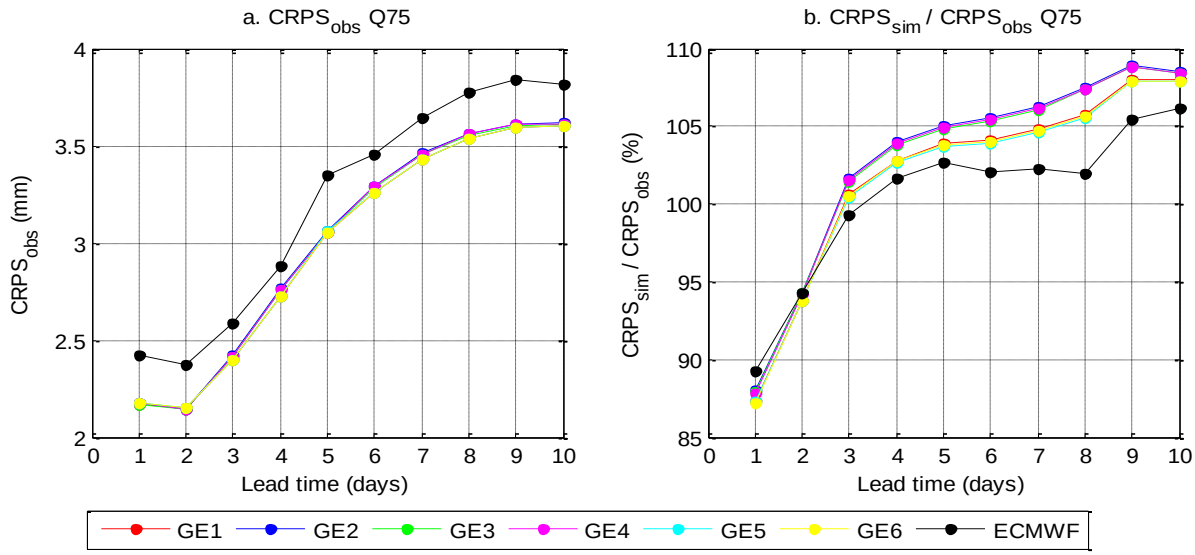

Figure 29 CRPS of the hydrological forecasts against discharge observations (a) and the relative contribution of meteorological errors determined by $CRPS_{sim}$ compared to meteorological + model errors determined by $CRPS_{obs}$ (b).

4.5.2.2 Continuous ranked probability skill score

Figure 30 shows the CRPSS of the hydrological forecasts for the four different EPSs for exceeding the Q75 threshold for lead times 1-10 days. It shows the relative performance of the different hydrologic ensemble forecasts to the performance of the reference forecast. The CRPSS of the grand ensemble forecasts has improved for all lead times and is better than the best forecasting ECMWF. There are no negative skill scores. This means that the CRPS of the grand ensemble forecasts is better than an ensemble forecast of discharges simulated with past observations of precipitation for all lead times. The results are almost similar to the results of the RMSE. The improvement of the CRPSS is highest for lead times of 1 and 2 days. Here, also a smoother line of the CRPSS of the grand ensemble forecasts is visible as a consequence of the different biases of the single forecasts that cancel out.

The CRPSSs of the different grand ensemble forecasts are also almost similar and show the same difference as described above for the RMSE. The combination methods where the members are combined (member size of the forecasts is included) results in a better CRPSS than the combination method which is not affected by member size (combination of models). The other combinations based on an evaluation score and the member combination or model combination do not result in a significant improvement in CRPSS compared to the combinations without using evaluation scores to calculate the weights.

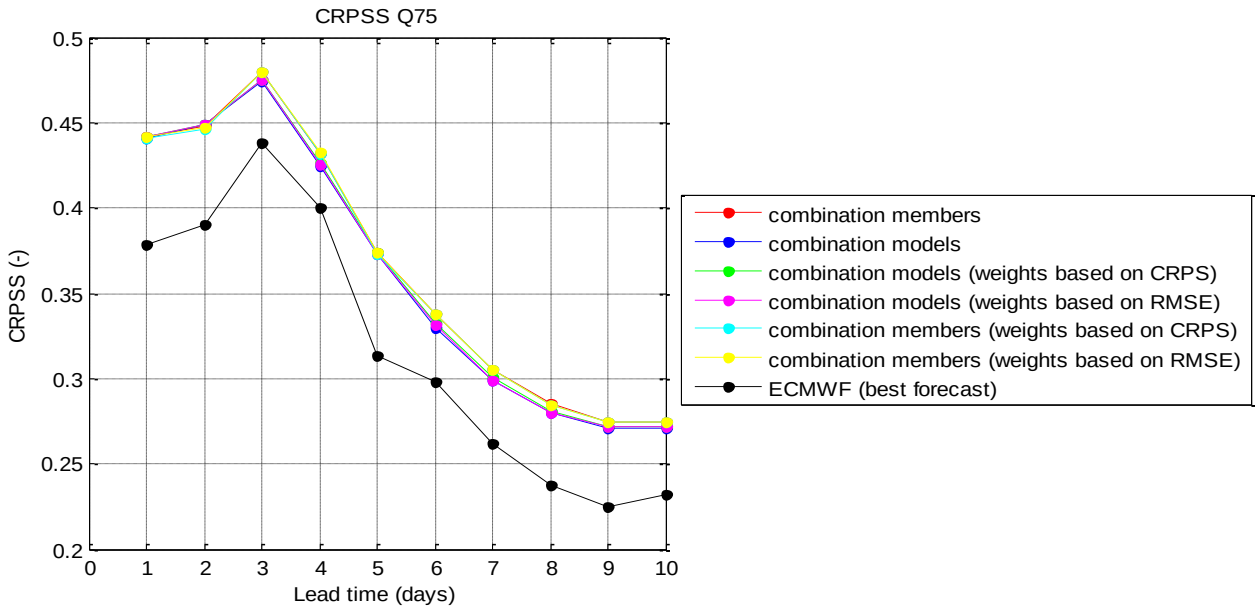

Figure 30 CRPSS of the grand ensemble hydrological forecasts of the six different combinations and the best performing single model ECMWF exceeding the Q75 threshold for lead times 1-10 and thresholds Q75.

4.5.2.3. Brier Score

Figure 31 shows the Brier score and the decomposition of the Brier score into a reliability score and a resolution score component for the hydrologic grand ensemble forecasts and ECMWF, calculated with the GR4J model including updating procedure and with the bias corrected precipitation forecast of the different EPSs as input, exceeding the P75, P85 and P95 thresholds for lead times 1-10 days.

Overall, Figure 31 shows the same tendency as the figures for CRPSS and RMSE. The skills of the grand ensembles are similar and better than the single model EPSs. Especially for long lead times there is a high improvement in the skill of the forecasts when using a grand ensemble.

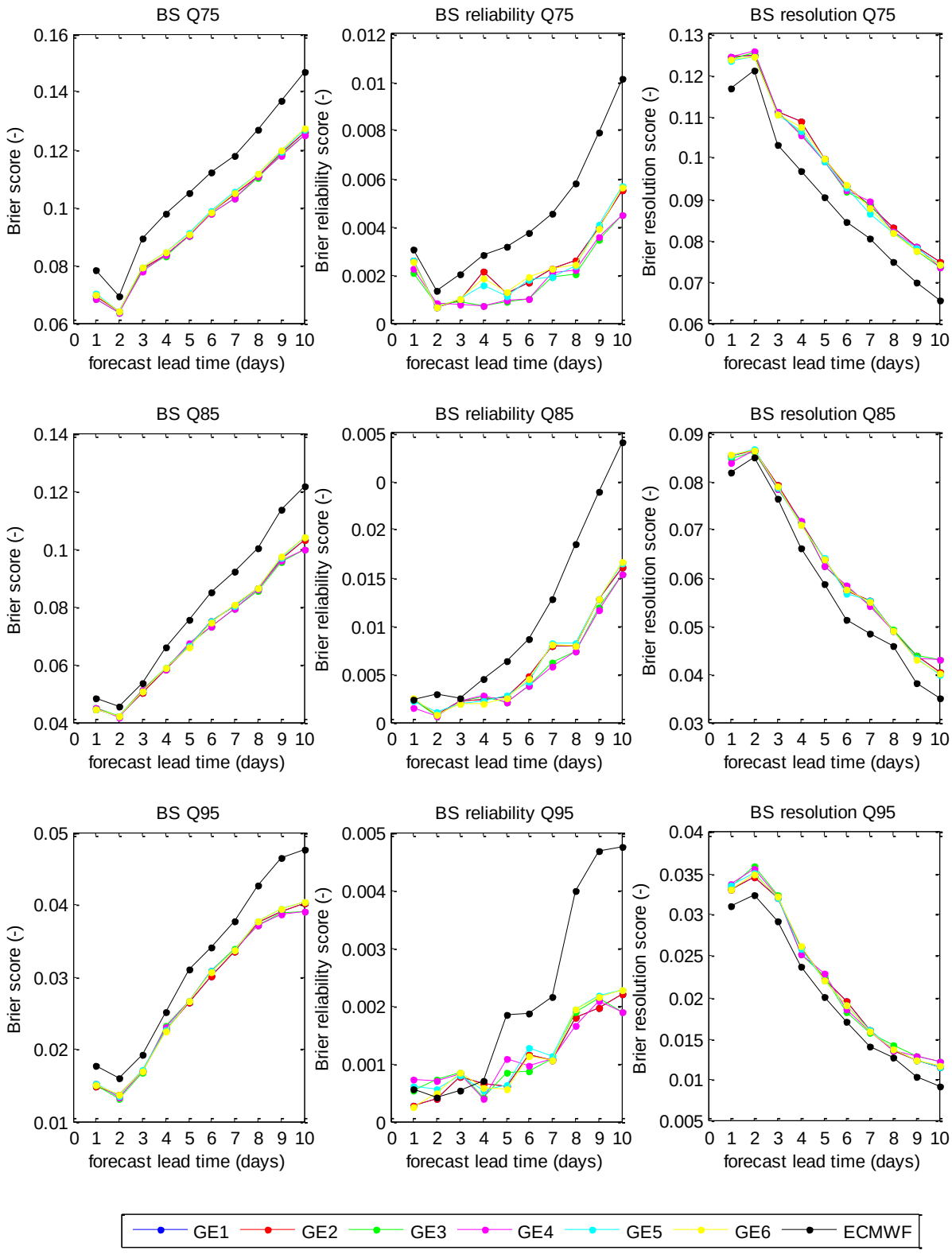

Figure 31 Brier score and decomposition into reliability and resolution of the grand ensemble hydrological forecasts and ECMWF for lead times 1-10 and thresholds Q75, Q85 and Q95. (GE1 = combination members, GE2 = combination models, GE3 = combination models (weights based on CRPS), GE4 = combination models (weights based on RMSE), GE5 = combination members (weights based on CRPS), GE6 = combination members (weights based on RMSE)).

4.5.2.4. Reliability and sharpness diagrams

Figure 32 and Figure 33 present the reliability diagrams and the sharpness diagrams of the grand ensemble hydrologic forecasts for different lead times (lead time 1, 3, 5 and 10 respectively). Lead times between 5 and 10 days are omitted, because there is the same trend visible as with the single model forecasts that the deviation with the diagonal grows with increasing lead time.

Overall there is a tendency visible that reliability and resolution decrease with increasing lead time. With increasing lead time, the graphs of the models deviate more from the diagonal (decreasing reliability) and get closer to the climatology line (decreasing resolution). The figures also show that the lines of the models lie more often under than above the diagonal, which means that the high flows of the forecasts are mostly underpredicted. The results of the grand ensembles are comparable to the other evaluation scores. The six grand ensembles have comparable skill. In general the reliability diagrams of the grand ensemble forecasts are closer to the diagonal and more deviated from the climatology when compared with the single model forecasts. Hence, the reliability and resolution have increased.

The sharpness diagrams below the reliability diagrams give insight in the sharpness of the forecasts. The sharpness diagrams of the different grand ensemble forecasts are generally the same depending on the lead time and threshold. The frequency of use here is also best distributed over the different probability categories for the Q75 threshold, thereafter Q85 and worst for Q95. For lower thresholds, forecast probabilities higher than zero become more frequent, because there are more forecasts with a value above the threshold value. So it is normal that the majority of the forecasts predict low probabilities for the threshold used here. These low probabilities are lower than the climatological probability of 25%, 15% and 5% respectively for Q75, Q85 and Q95 as can be seen by the starting point of the reliability graphs which is close to zero. The forecasts are also able to predict relatively high probabilities greater than 40% of the event in comparison with climatologies, although such forecasts are less frequent. The forecasts here thus have sharpness. In general the grand ensemble forecast of the simple combination with the models show the highest sharpness with higher bars in the higher probability categories. In general the sharpness and the distribution of the forecasts over the probability categories of the grand ensemble forecasts has improved in comparison with sharpness diagrams of the single-model forecasts. The bars of the middle and the higher probability categories are generally higher.

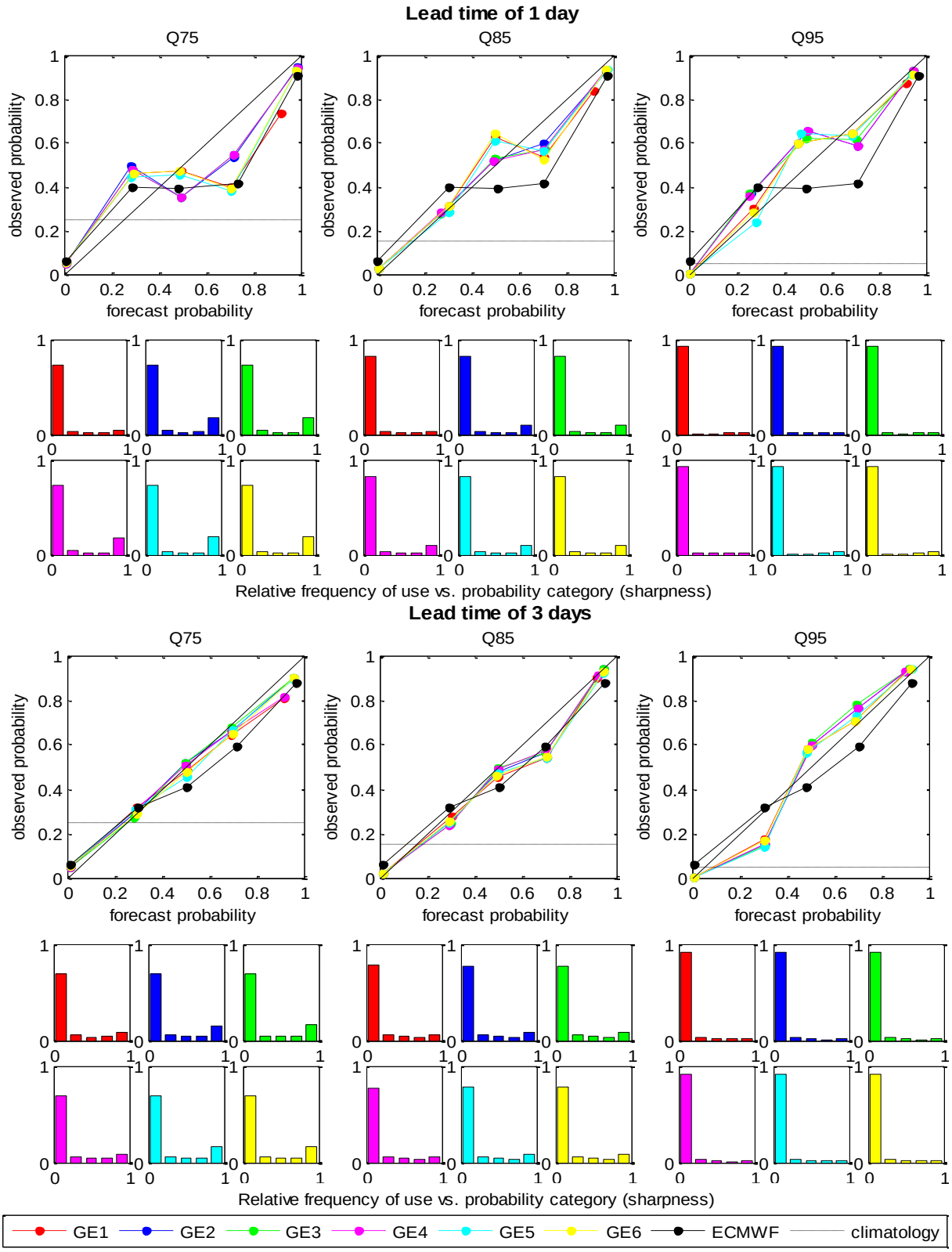

Figure 32 Reliability diagram and sharpness diagrams of the different grand ensemble hydrologic ensemble forecasts for a lead time of 1 and 3 days and thresholds Q75, Q85 and Q95. (GE1 = combination members, GE2 = combination models, GE3 = combination models (weights based on CRPS), GE4 = combination models (weights based on RMSE), GE5 = combination members (weights based on CRPS), GE6 = combination members (weights based on RMSE)).

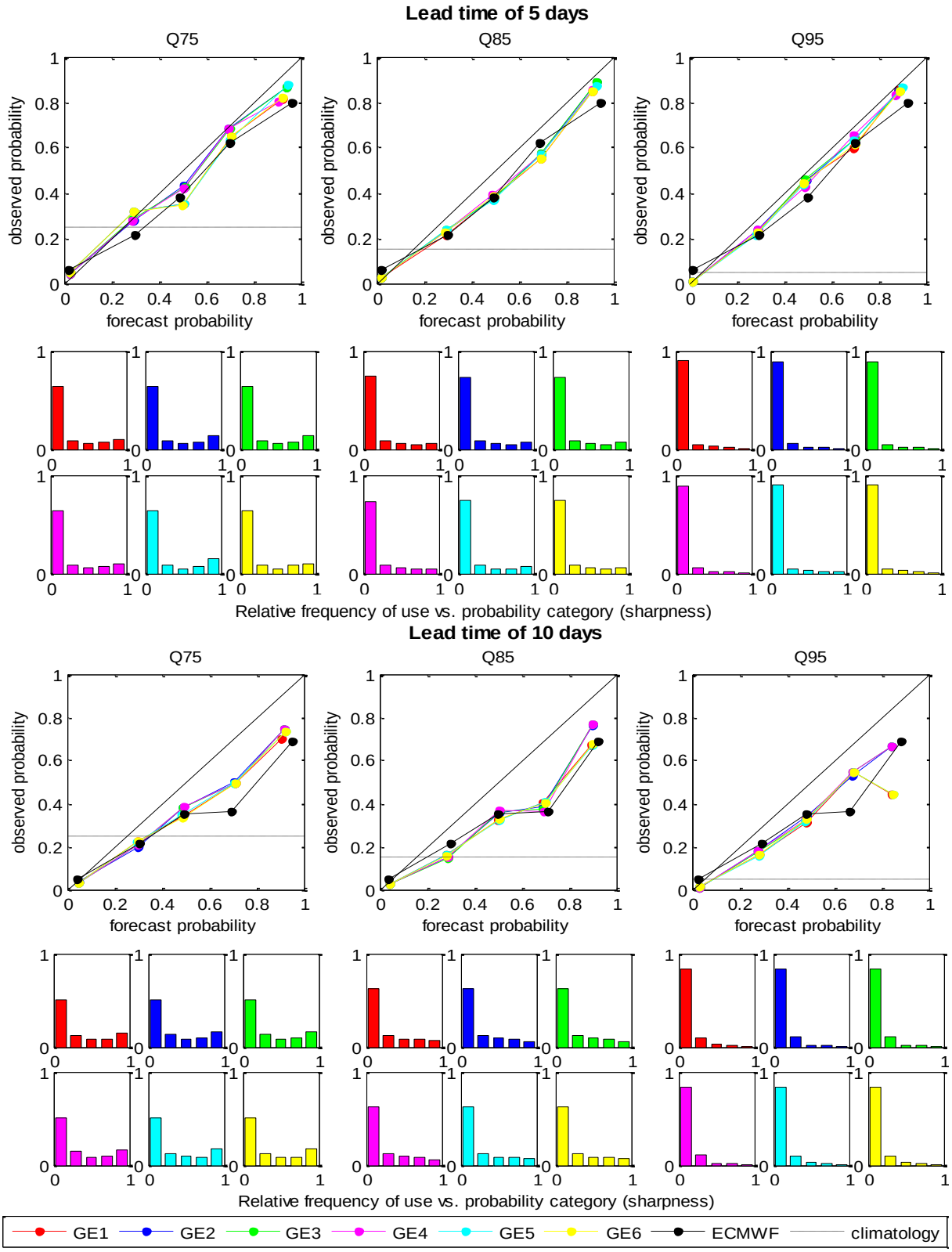

Figure 33 Reliability diagram and sharpness diagrams of the different grand ensemble hydrologic ensemble forecasts for a lead time of 5 and 10 days and thresholds Q75, Q85 and Q95. (GE1 = combination members, GE2 = combination models, GE3 = combination models (weights based on CRPS), GE4 = combination models (weights based on RMSE), GE5 = combination members (weights based on CRPS), GE6 = combination members (weights based on RMSE)).

4.5.3 Conclusions

In all evaluation methods it becomes clear that grand ensemble hydrological forecasts are beneficial. They show improved RMSE, CRPSS, reliability and resolution compared to the best single model EPSs. The sharpness is more or less similar to the single model forecasts. The results also show that the CRPSS and RMSE graphs are smoother. This is the result of the different biases of the single forecasts that cancel each other out in the grand ensemble forecasts. It was expected that a skill based grand ensemble is better than a simple grand ensemble combination of the models or combination of the members, however the simple grand ensemble show almost similar skill as the skill based grand ensemble. This shows that in this case EPSs with less skill than other EPSs still can add skill in a grand ensemble, because a model with less skill can add model structure errors that are missing in another EPS with good skill. In this study area the approach of combining the members for the grand ensemble is best performing, however the performance is almost similar to the other grand ensembles. This is in line with the results of the single model forecasts where the model with the most members (ECMWF) is performing best, while the model with the fewest members (CMA) is performing worst. In chapter 3 was described that in general increasing ensemble size leads to little improvement, however models with less members can be better than models with more members. Therefore it is better to use an approach where the models are weighted so that the influence is not dependent on ensemble size.

5. Discussion

5.1. Hydrological model and data

The GR4J model was already calibrated on observations of precipitation, temperature and discharge over a period from 1981 to 1990. It was found that the calibrated GR4J model had a high performance with a NS value of 0.93 for the Qu basin. The optimum simulation results were obtained through the parameter set with the highest Nash Suttcliffe (NS) efficiency coefficient. It is considered in this study that the calibrated GR4J model is good to use for forecasting of high flows, since NS is more sensitive to high flows than low flows. However, Tian et al. (2014) concluded that the GR4J simulations tend to underestimate the extreme high flows. Calibration focused on high flows probably results in a better parameter set to simulate extreme high flows. To improve the performance of GR4J and better simulate high flows an updating procedure is used in this study (section 5.2).

Observed data from 3 measurement stations have been used to calculate the mean daily areally averaged precipitation over the study area. There is no station in the southern part of the Quzhou basin. This might lead to biases in the mean observed daily areally averaged precipitation. This precipitation data is used in the simulation of the perfect flow forecasts, Quantile mapping method and the evaluation of the bias correction. It is also possible that other observation errors can be present. For the validation period newly obtained observed precipitation and discharge data are used. It is likely that there are errors present in these datasets, because this data is never used in previous studies. Some errors in the discharge dataset were already found in this study and are excluded in the validation period.

5.2. Hydrological updating and bias correction

In this study a simple effective updating procedure developed by Demirel et al. (2013) for the GR4J model is used to minimize errors in the initial model state. The updating procedure is a model storage updating procedure based on the observed discharge on the forecast issue day. This updating approach is chosen in this study, because popular updating procedures as Auto Regression and Kalman filtering are not suitable for short forecast periods and steep flood hydrograph characteristic which is typical for small, quickly reacting mountainous catchments such as the Quzhou catchment area (Wöhling et al., 2006). The updating procedure used here was effective (see Table 5). It improved both the NS and RVE over the validation period for all flow forecasts and for high flow forecast above the Q75 threshold.

The quantile mapping bias correction used in this study was also effective. It removed the drizzle and extreme precipitation forecast values were closer to the observed extreme precipitation values. However, it should be noted that the quantile mapping method is applied to the validation period with the use of observation data from this validation period. So the CDFs of the forecasts are corrected with the CDFs of the observation of the same period. This obviously results in good bias corrected values of the forecasts and is the best bias correction you can get with the quantile mapping method. However, in reality with forecasting the quantile mapping method have to be applied with the use of a training dataset of historical values of forecasts and observations, because the observations used in this study are

in reality observations in the future and therefore not available. For that reason sometimes hindcasts of forecast data are used to correct for bias. Hindcasts are reforecasts obtained with the forecast model for historical periods. In this study a hindcast dataset of areally averaged ECMWF ensemble precipitation forecast with lead time 1-10 days for the period 1981-1990 was used together with the observed areally averaged precipitation of 1981-1990 to use as training set for the quantile mapping method and correct the bias of the 'future' forecasts of ECMWF of the validation period in this study. However the ECMWF ensemble precipitation forecasts was not improved and even worsened. The difference between the hindcast data and the observations were too high and not comparable with the differences between the datasets in the validation period. The biases of the hindcast and the forecasts in the validation period were not very similar, with the result that the quantile mapping methods made wrong corrections. In this study the quantile mapping method is not used with a training period, because the TIGGE datasets are only available for a short period. In a few years it is expected that the TIGGE datasets can be bias corrected by a quantile mapping approach, because the datasets then will probably be long enough and/or hindcast datasets will be of better quality.

5.3. Evaluation

Different evaluation methods are used in this study to evaluate the ensemble flow forecasts and the methodology used in this study. There are some restrictions in these evaluations. At first, observation errors can be present as described in 5.1. Another restriction is the validation period, which is relatively short (almost 5 years). This is because the TIGGE datasets are quite short. As a consequence threshold values for high flows are not extremely high, because enough numbers of events are essential to make confident evaluations. With a longer period of data availability also conclusions could be less uncertain. Additional, extra evaluation measures can lead to more certain conclusions, because skill has more components than the three (reliability, resolution and sharpness) evaluated here. For example discrimination as mentioned before in section 3.1 and uncertainty.

For the Brier score it is difficult to interpret the score, because forecasts exceeding higher thresholds have better scores. This is the consequence of the small probability of exceedance of the high flows and the forecast discharges, with the result that a large frequency of similarity between the forecast and the observation below the exceedance threshold will be present. However, the advantage of the Brier score is that it can be easily decomposed into reliability and resolution. For each threshold can be seen how the forecasts skill is determined by reliability and resolution and the different EPSs can be compared.

5.4. Evaluation results

In this section the performance of the ensemble flow forecasts are examined relative to other ensemble flow forecasts from other studies. The ensemble flow forecast performance in this study decreases with lead time. Other studies also show this trend. Not every evaluation measure can be directly compared to other studies. It is impossible to compare the Brier score, CRPS and RMSE to other studies. It is also difficult to compare the reliability and sharpness diagrams with other studies, because the thresholds have to be comparable. The Brier score is dependent on the number of forecasts exceeding the evaluated threshold and CRPS and RMSE is dependent on the magnitude of the evaluated variable. In this study also the CRPSS is used as evaluation measure. The CRPSS is dependent on the reference forecast. Bennett et al. (2014) used the same reference forecast generated by the use of historical

observations of precipitation and temperature to calculate the potential evapotranspiration, therefore the CRPSSs can be compared. The skills evaluated with the CRPSS have more or less the same magnitude as in the study of Bennett et al. (2014). Also the same tendency of decreasing skill for lead times 1-9 is visible (they evaluated the flow for lead times 0-9 days). The deterministic and probabilistic evaluation measures used in this study, the RMSE, CRPSS, and the reliability diagram, showed high performance of the forecasts for the different thresholds and lead times. The flow forecasting system of Bennett et al. (2014) show some similarities and also differences. They use an hourly lumped GR4 model instead of a daily lumped GR4 model; the meteorological forecasts are different, but the spatial resolution is comparable; they also use post-processing of the meteorological forecasts and an updating procedure for the flow forecasts, however these methods are different than in this study. Despite these differences, the results obtained with the CRPSS are comparable. In this study also the contribution of errors in the hydrological forecasts has been examined. In general, the meteorological error increases relative to the hydrological error and was more important than the hydrological errors except for the lead times 1-2 days where the hydrological error was more important. This is comparable with other studies. Demargne et al. (2010) conclude that hydrological model uncertainty is more important for short lead times, while Renner et al. (2009) and Olsson & Lindström (2008) both found that biases mainly originated from the meteorological forecasts.

In the introduction and in chapter 3 was described that several studies found grand ensemble forecasts a good strategy to improve the EPS forecasts. This is in agreement with the results in this study. It was expected that giving the different weather centres weight in the multi-model forecast based on skill scores would improve the performance of the grand ensemble forecasts. However, the six different weighting methods used in this study including weighting based on skill did not result in significant different performances. This is in contradiction with other studies as Johnson & Swinbank (2009), Raftery et al. (2005) and Stefanova & Krishnamurti (2002). However, they used different approaches to weight the models, used different forecast parameters and had another study area. The conclusion that the different biases of the single forecasts cancel out in the grand ensemble forecasts is on the other hand comparable with the other studies.

6. Conclusions and recommendations

This chapter presents the conclusions and recommendations obtained from this study.

6.1. Conclusions

Flood forecasting is becoming more important since more frequent floods have been experienced by regional communities in the catchment in recent decades. With Ensemble Prediction Systems it is possible to extend lead time and to better account for the uncertainty in forecasts. Chapter 1 describes that recently grand ensemble forecasts can be used to improve the forecasts as well.

In chapter 1 the objective of this study was defined as:

The purpose of this study is to develop an ensemble flood forecasting system for Quzhou (East-China) for lead times of 1 to 10 days and to evaluate different combined Grand Ensemble flood forecasts.

From this objective three research questions have been formulated in chapter 1. The conclusions drawn from this study will be presented here with reference to the research questions.

1. *What is the performance of the meteorological forecasts and the hydrological model and how is this improving with the implementation of a bias correction method and a hydrological updating procedure?*

In this study the quantile mapping method is used to correct the biases of the raw ensemble precipitation data of the EPSs of CMA, ECMWF, UKMO and NCEP. Several evaluations of the corrected forecasts have shown that the bias correction results in improved ensemble precipitation forecasts. Cross correlation of the forecasts with observations results in relatively high correlation coefficients and shows that forecasts with lead times up to 10 days can still be meaningful. The frequency distribution has improved as shown in CDF plot comparisons between raw data and corrected data. After bias correction there is a good match between the CDFs of the forecasts and the observations of precipitation. Differences in dry day frequencies are corrected and the most extreme events are closer to the observations. Also the high amounts of low precipitation values close to zero in the forecasts (drizzle) are corrected. From the evaluation of the ensemble models with the Brier score it can be concluded that the reliability and resolution improve by bias correction for most lead times and threshold values. ECMWF proves to be the most skilful model and CMA the least skilful model for the Quzhou catchment area for precipitation forecasts for the raw as well as the bias corrected forecasts. However, for short lead times NCEP is least skilful. In general, the forecast skills decrease with lead time, as meteorological errors increase with lead time.

The hydrological updating procedure leads to an increase of the model performance. The NS improves from a validated value of 0.91 to a value of 0.94 for a lead time of 1 day. For high flows above the Q75 threshold this increase is from 0.89 to 0.92. The RVE also improves. The performance determined by NS and the RVE decreases with lead time to approximately the values they had without the implementation

of the updating procedure as a consequence of the smaller influence of the hydrological updating procedure with longer lead times.

2. *What are the performances of the ensemble flood forecasting system for the different TIGGE ensemble prediction models in the study area?*

The performance of the hydrological forecasts show similar performance trends as for the EPSs of the precipitation forecasts. The skill decreases with lead time and ECMWF generally has the most skilful hydrological ensemble forecasts, showing both good reliability and resolution. CMA is the least skilful hydrological ensemble forecast for the evaluation period and the Qu river basin. NCEP is performing worst for the shortest lead times (1-2 days). The sharpness diagrams show that the forecasts have sharpness and that ECMWF and CMA show the highest sharpness. When comparing the Brier scores of the precipitation forecasts with those for the hydrological forecasts it becomes clear that the hydrological model leads to increased errors because of the model errors. The Brier score graph for precipitation forecasts flattens for lead times over 5 days, while the Brier score graphs of the hydrologic forecasts do not show this flattening. This leads to less skilful forecasts especially for the longer lead times where the hydrological updating procedure has less effect. So the meteorological error as well as the hydrological error increases with lead time. This study shows that the meteorological error increases relative to the hydrological model error with lead time, even when the hydrological error itself is increasing. The meteorological error is higher than the hydrological error for high flows.

3. *What are the performances of grand ensemble flood forecasts with different weighting methods?*

All evaluation methods make clear that grand ensemble hydrological forecasts are beneficial. They show improved RMSE, CRPSS, reliability and resolution compared to the single model EPSs. The sharpness of the grand ensemble hydrological forecasts is more or less similar to that of single model forecasts. The results also show that the CRPSS and RMSE graphs are smoother. This is the result of the different biases of the single forecasts that cancel out in the grand ensemble forecasts. It was expected that a skill based grand ensemble yields better forecasts than a simple grand ensemble combination of the models or combination of the members, but this was not confirmed. This shows that EPSs with less skill than other EPSs still can add skill in a grand ensemble. In this study area the approach of combining the members for the grand ensemble is best performing, however the performance is almost similar to the other grand ensembles. This is in line with the results of the single model forecasts where the model with the most members (ECMWF) is performing best, while the model with the fewest members (CMA) is performing worst. In the grand ensemble where all members are combined, models with a high number of members get more weight. Hence, ECMWF is weighted more than CMA when all members are combined because of the difference in ensemble size. Generally it can be concluded that there is no significant difference between the different combination methods. In chapter 3 was described that in general increasing ensemble size leads to little improvement, however models with less members can be better than models with more members. Therefore it is better to use an approach where the models are weighted with the method of equal probability of selection so that the influence is not dependent on ensemble size.

Summarized, the results show that a simple combination of ensembles from different EPSs give a significant improvement in the forecast skill in comparison with single model ensemble forecasts and that this is a promising approach to use in flood forecasting in the Quzhou river basin.

6.2. Recommendations

The developed ensemble flood forecasting system can be used in the Quzhou catchment to produce skilful predictions of high flows. However, it is recommended to do further research on higher flows (e.g. Q99) when more data are available. With a longer period of available data conclusions can be drawn with better confidence, especially for high flows. Because the period of available data was only 5 years high flows above the Q99 thresholds were not included in this research. Also it is recommended to carry out further research on bias correction of the meteorological ensemble forecasts, since here bias correction has been done with the observations from the same period as the forecasts due to the small dataset. In a few years it is expected that a quantile based mapping bias correction approach with the use of a training dataset can be applied, because the data availability will be higher.

Based on the results from this study it is recommended to use grand ensemble forecasts over single model ensemble forecasts in flood forecasting in the Quzhou river basin. The results of this study show that the predictive skill has significantly improved when using a grand ensemble. For future implementation of grand ensemble forecasts it is recommended to construct simple grand ensemble forecasts without the use of weights based on skill in the Quzhou catchment. The improvement with the use of skill was not significant, even with weights based on skill scores determined over the same evaluation period. The difference between the two simple combination approaches (combination of the models and combination of the members) was also not significant. For future implementation the order of best performing models are unknown because the forecast period is different. In addition, the EPSs are continuously improving their models as described in the introduction with the result that other models might be better in the future. Therefore, the best combination approach to use for future implementation in flood forecasting systems in the Quzhou basin is the combination method where the models are combined with an equal probability in the grand ensemble where all models have the same weights in the grand ensemble. This is the most common used approach to combine models and is a promising approach.

It is expected that the recommendation to use grand ensemble flood forecasts for the Quzhou basin also applies to other basins (small basins as well as large basins), because other studies came to the same conclusion that grand ensemble forecasts are beneficial in comparison with single model ensemble forecasts. It is recommended to perform additional research on other catchments to validate or refine general conclusions.

It is recommended to calibrate the hydrological model on high flows specifically when using the hydrological model for flood forecasting. The hydrological model used here is calibrated on all flows using NS efficiency which is more sensitive to high flow.

The hydrological updating procedure used in this study resulted in an improvement of the hydrological model and the forecasts for all lead times in this study (1-10 days). It is therefore recommended to use an updating approach in the ensemble flood forecasting approach.

References

Addor, N., Jaun, S., Fundel, F., & Zappa, M. (2011). An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.*, *15*(7), 2327–2347. http://doi.org/10.5194/hess-15-2327-2011

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. http://doi.org/10.1016/j.jhydrol.2014.06.035

Bao, H.-J., Zhao, L.-N., He, Y., Li, Z.-J., Wetterhall, F., Cloke, H. L., … Manful, D. (2011). Coupling ensemble weather predictions based on TIGGE database with Grid-Xinanjiang model for flood forecast. *Adv. Geosci.*, *29*, 61–67. http://doi.org/10.5194/adgeo-29-61-2011

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 *20th International Congress on Modelling and Simulation* (pp. 2611–2617).

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 9days. *Journal of Hydrology*, *519*, 2832–2846. http://doi.org/10.1016/j.jhydrol.2014.08.010

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. http://doi.org/10.1175/2010BAMS2853.1

Bröcker, J., & Smith, L. A. (2007, June 1). Increasing the reliability of reliability diagrams. *Weather and Forecasting*. American Meteorological Society. Retrieved from http://eprints.lse.ac.uk/22217/#.Vj_uSwdWoAM.mendeley

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

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

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. http://doi.org/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. http://doi.org/10.1002/hyp.9419

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2013). 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. http://doi.org/10.1002/wrcr.20294

Du, N., Ottens, H., & Sliuzas, R. (2010). Spatial impact of urban expansion on surface water bodies—A case study of Wuhan, China. *Landscape and Urban Planning*, *94*(3-4), 175–185. http://doi.org/10.1016/j.landurbplan.2009.10.002

ECMWF. (2013, January 03). *Precipitation*. Retrieved from Software ECMWF: https://software.ecmwf.int/wiki/display/EMOS/Precipitation

ECMWF. (2015). ECMWF - TIGGE Data Retrieval. Retrieved from http://apps.ecmwf.int/datasets/data/tigge/

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. http://doi.org/10.5194/hess-11-939-2007

He, C., Tian, J., Shi, P., & Hu, D. (2011). Simulation of the spatial stress due to urban expansion on the wetlands in Beijing, China using a GIS-based assessment model. *Landscape and Urban Planning*, *101*(3), 269–277. http://doi.org/10.1016/j.landurbplan.2011.02.032

He, Y., Wetterhall, F., Bao, H., Cloke, H., Li, Z., Pappenberger, F., … Huang, Y. (2010). Ensemble forecasting using TIGGE for the July–September 2008 floods in the Upper Huai catchment: a case study. *Atmospheric Science Letters*, *11*(2), 132–138. http://doi.org/10.1002/asl.270

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

Johnson, C., & Swinbank, R. (2009). Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, *135*(640), 777–794. http://doi.org/10.1002/qj.383

Kahl, B., & Nachtnebel, H. P. (2008). Online updating procedures for a real-time hydrological forecasting system. *IOP Conference Series: Earth and Environmental Science*, *4*(1), 12001. http://doi.org/10.1088/1755-1307/4/1/012001

Khan, M. M., Shamseldin, A. Y., & Melville, B. W. (2014). Impact of Ensemble Size on Forecasting Occurrence of Rainfall Using TIGGE Precipitation Forecasts. *Journal of Hydrologic Engineering*, *19*(4), 732–738. http://doi.org/10.1061/(ASCE)HE.1943-5584.0000864

Liu, J., & Xie, Z. (2014). BMA Probabilistic Quantitative Precipitation Forecasting over the Huaihe Basin Using TIGGE Multimodel Ensemble Forecasts. *Monthly Weather Review*, *142*(4), 1542–1555. http://doi.org/10.1175/MWR-D-13-00031.1

Nester, T., Komma, J., Viglione, A., & Blöschl, G. (2012). Flood forecast errors and ensemble spread-A case study. *Water Resources Research*, *48*(10), n/a–n/a. http://doi.org/10.1029/2011WR011649

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

Orientplus. (2015). Meteorology: High-speed networking: helping to win the race against severe weather. Retrieved March 4, 2015, from http://www.orientplus.eu

Oudin, L., Michel, C., & Anctil, F. (2005). Which potential evapotranspiration input for a lumped rainfall-runoff model? *Journal of Hydrology*, *303*(1-4), 275–289. http://doi.org/10.1016/j.jhydrol.2004.08.025

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 prediction. *Journal of Hydrology*, *522*, 697–713. http://doi.org/10.1016/j.jhydrol.2015.01.024

Park, Y.-Y., Buizza, R., & Leutbecher, M. (2008). TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, *134*(637), 2029–2050. http://doi.org/10.1002/qj.334

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1-4), 275–289. http://doi.org/10.1016/S0022-1694(03)00225-7

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. http://doi.org/10.1175/MWR2906.1

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. http://doi.org/10.1016/j.jhydrol.2009.07.059

Stefanova, L., & Krishnamurti, T. N. (2002). Interpretation of Seasonal Climate Forecast Using Brier Skill Score, The Florida State University Superensemble, and the AMIP-I Dataset. *Journal of Climate*, *15*(5), 537–544. http://doi.org/10.1175/1520-0442(2002)015<0537:IOSCFU>2.0.CO;2

Su, X., Yuan, H., Zhu, Y., Luo, Y., & Wang, Y. (2014). Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008-2012. *Journal of Geophysical Research: Atmospheres*, *119*(12), 7292–7310. http://doi.org/10.1002/2014JD021733

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*, 2890–2905. http://doi.org/10.1016/j.jhydrol.2014.04.040

Tian, Y., Booij, M. J., & Xu, Y.-P. (2014). Uncertainty in high and low flows due to model structure and parameter errors. *Stochastic Environmental Research and Risk Assessment*, *28*(2), 319–332. http://doi.org/10.1007/s00477-013-0751-9

Tödter, J., & Ahrens, B. (2012). Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition. *Monthly Weather Review*, *140*(6), 2005–2017. http://doi.org/10.1175/MWR-D-11-00266.1

Voisin, N., Schaake, J. C., & Lettenmaier, D. P. (2010). Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts. *Weather and Forecasting*, *25*(6), 1603–1627. http://doi.org/10.1175/2010WAF2222367.1

Wilks, D. (2006). *Statistical Methods in the Atmospheric Sciences* (Vol. 91). New York: Elsevier.

WMO. (2015, February 3). *Forecast Verification: Issues, Methods and FAQ*. Retrieved from WWRP/WGNE Joint Working Group on Forecast Verification Research: 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. *Hydrol. Earth Syst. Sci.*, *10*(6), 783–788. http://doi.org/10.5194/hess-10-783-2006

Wu, L., Seo, D.-J., Demargne, J., Brown, J. D., Cong, S., & Schaake, J. (2011). Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *Journal of Hydrology*, *399*(3-4), 281–298. http://doi.org/10.1016/j.jhydrol.2011.01.013

Xu, Y.-P., Zhang, X., Ran, Q., & Tian, Y. (2013). Impact of climate change on hydrology of upper reaches of Qiantang River Basin, East China. *Journal of Hydrology*, *483*, 51–60. http://doi.org/10.1016/j.jhydrol.2013.01.004

Yang, L., Scheffran, J., Qin, H., & You, Q. (2015). Climate-related flood risks and urban responses in the Pearl River Delta, China. *Regional Environmental Change*, *15*(2), 379–391. http://doi.org/10.1007/s10113-014-0651-7

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. http://doi.org/10.1002/qj.2243