TECHNICAL MEMORANDUM

557

TIGGE: Medium range multi model weather forecast ensembles in flood forecasting (a case study)

F. Pappenberger[1], J.Bartholmes[2], J. Thielen[2] and Elena Anghel[3]

Research Department

[1] ECMWF, Reading, UK
[2] Joint Research Centre of the European Commission, Ispra, Italy
[3] National Institute of Hydrology and Water Management of Romania

January 2008

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Abstract

The performance of a hydrological multi model flood forecast system with inputs provided by seven ensemble prediction systems of the THORPEX Interactive Grand Global Ensemble (TIGGE) archive is evaluated. A flood forecast of this multi-model ensemble has been computed for case study the October 2007 floods in Romania using the LISFLOOD model of the European Flood Forecasting System (EFAS) as hydrological component. None of the forecast centres (ECMWF, UKMO, JMA, NCEP, CMA, CMC, BOM and simple multi-model ensemble) predict the distribution of precipitation observations exactly, with most centres exhibiting an over prediction at day nine. The distribution of discharge predictions for observations and forecasts is similar at the lower flow. The percentage of precipitation and discharge forecasts above and below the 10^{th} and 90^{th} percentile of the predicted EPS distributions is very high. The UKMO, ECMWF and the multi-model ensemble forecast perform favourably in terms of root mean squared error of the ensemble mean discharge and precipitation predictions. All forecasts (apart from the one issued by BOM) would have lead to a correct flood warning about 8 days in advance. It can be demonstrates that the Multi-model ensemble has the best average properties, followed by the forecast of ECMWF and UKMO. The increased quality achieved by the multi-model ensemble is explained by the fact that it provides a better approximation at the tails of the distribution. All analysis is based on a set of criteria specific for this case study and might be different in more general cases.

1. Introduction

One major research challenge of the 21^{st} century is to mitigate the effects of natural hazards. Of all natural disaster, flooding is the most frequent, affecting the second largest number of people after droughts causing damage in excess of several billion Euros a year [4].

Flood forecasting based on observed precipitation or river levels, limits the lead time of the forecast to the natural response time of the catchment. However, longer lead times provide civil protection authorities with more time to prepare for the event and give an advanced warning to the public, and could reduce the socio-economic impact of the flooding. Although the incorporation of numerical weather forecasts into a flood warning system can significantly increase forecast lead time [for example 5,6-10], many hydrological services consider the use of forecasted rainfall to introduce an unacceptable degree of uncertainty into their forecasts making decision making problematic [11]. Currently, numerical weather predictions with a single deterministic forecast are mostly used for hydrological predictions in the short range, in which the quality of the forecasts is still high and the impact of uncertainties such as initial conditions on the weather predictions are low [12].

These limitations are addressed by ensemble prediction systems (EPS) which incorporate uncertainties in the initial conditions and factors of the modelling process in the numerical weather predictions and produce multiple weather forecasts [13]. It has been demonstrated that EPS have more value than a single deterministic forecast with the same resolution from the same modelling system [14,15]. EPS are used successfully at meteorological forecasting centres around the world [16-22]. Recently EPS are also increasingly applied in coupled meteorological-hydrological modelling systems. Examples of such integrated approaches of operational flood forecasting systems in are shown in Bangladesh [23] the flood forecasts of the Finnish Hydrological Service [24] and Swedish Hydro-Meteorological Service [25], the European Flood Alert System [2] and many more.

EPS forecasts from a single forecast centre only address some of the uncertainties inherent in numerical weather predictions and many other sources such as boundary conditions or numerical implementations exist. For example, model physics and numerics have substantial impact in generating the full spectrum of possible solutions [26]. A multi-model approach is an effective and pragmatic approach of incorporating some of these additional sources of uncertainty [27].

In this paper we aim for an evaluation of a hydrological multi model ensemble flood forecast. The meteorological input data are provided by the THORPEX (http://www.wmo.ch/pages/prog/arep/thorpex/) Interactive Grand Global Ensemble (TIGGE) data archive which collects global ensemble forecasts from more than seven meteorological centres around the world. This initial evaluation will be performed on a case study of a 2-5 year return period flood event that took place in Romania in October 2007 in tributaries to the Danube. We used the standard set-up of the European Flood Alert System (EFAS) [1], which provides early flood alerts for Europe pre-operationally [2,3] since 2005. The analysis will focus generally on the region, but also two individual locations with their flood warning status evaluated based on these results.

2. Experimental set-up

In this section the experimental set-up of the case study experiment will be presented. First a brief introduction to the TIGGE data set is given, then EFAS is introduced and finally the event and study region is described.

2.1 Thorpex Interactive Grand Global Ensemble (TIGGE)

TIGGE is a key component of THORPEX, a World Weather Research Programme to accelerate the improvements in the accuracy of 1-day to 2 week high-impact weather forecasts for the benefit of humanity. Part of the key objectives of TIGGE is to test concepts of a TIGGE Prediction Centre to produce ensemble-based predictions of high-impact weather, wherever it occurs, on all predictable time ranges. This paper contributes to this objective, by evaluating the case study with regard to real time flood forecasting. Table 1 shows the forecasts used in this study. All forecast centres provide forecasts of at least 10 days with the exception of Japan, which provides forecasts only up to day 9. Only the first 10 days of meteorological forecasts have been used to force the hydrological model in order to allow for a comparison with the standard set-up of the EFAS.

The usage of the multi-model TIGGE ensembles in flood forecasting (i) recognizes that multiple modelling structures may be equally valid representations of the system and application in question (ii) accepts that all models have their inherent weaknesses and strengths (iii) harvests the fact that each model makes use of different information and incorporates information in different ways (for example differences in data assimilation) [28]. In this way some of the deficits of using the output of a single model EPS are overcome. The caveat is that it will never be possible to represent the full range of all uncertainties present in the modelling process. This would not only require a very large number of different implementations of model structures, but also a full understanding and quantification of all sources of uncertainty, which is impossible [29]. Moreover, all models in any current multi-model prediction systems are based on the same paradigm of current scientific consensus of the physical processes in the atmosphere and land surface. Therefore, this approach inevitably neglects sources of uncertainty outside the current paradigm. Nevertheless, multi-model prediction systems have been proven successful in meteorology and hydrology [30,31]. Moreover,

methodologies have been suggested to incorporate the uncertainty of meteorological, hydrological and hydraulic models [5,7].

Table 1: Meteorological forecast centres and the data used in this study. For the hydrological forecasts only the first 10 days of lead time were used.

Centre	Abbreviation	Country / Domain	Ensemble Members	Horizontal Resolution	Vertical Levels	Forecast Length
Bureau of Meteorology	BOM	Australia	33	TL119○	19	10
China Meteorological Administration	CMA	China	15	T213	31	10
National Centre for Environmental Predictions	NCEP	USA	21	T126	28	16
UK MetOffice	UKMO	United Kingdom	24	1.25x0.83deg	38	15
Canadian Meteorological Centre	CMC	Canada	21	T254 (up to 3.5 days) then T170	64 (up to 3.5 days) then 42	16
Japan Meteorological Agency	JMA	Japan	51	TL159	40	9
European Centre for Medium-Range Weather Forecasts	ECMWF	Europe	51	TL399 (up to day 10)	62	15

The TIGGE archive contains all variables such as precipitation, evaporation, 2 metre temperature which are necessary to drive the hydrological model of EFAS.

2.2 European Flood Forecasting System (EFAS)

The repetitive occurrence of high impact trans-national floods in Europe prompted the European Commission to investigate new strategies for flood prevention and protection, with focus on coordinated actions among countries sharing the same river basin and funded the development of EFAS [1,5,10,11]. EFAS aims to increase preparedness in trans-national European river basins [2,3,32,33] by providing early flood warning information on a catchment scale. Its weather forecasting inputs include the full set of EPS forecasts from ECMWF as well as a poor-man ensemble consisting of ECMWF and DWD deterministic forecasts. These inputs allow EFAS to provide local water authorities with probabilistic medium-range flood forecasting information 3 to 10 days in advance.

The hydrological model of EFAS is the LISFLOOD model, a hybrid between a conceptual and a physical rainfall-runoff model with a channel routing [1,2]. It is set-up for the whole of Europe on a 5 km grid. Each weather forecast ensemble member is propagated through the hydrological model as a single deterministic forecast and the resulting distribution of discharges is explored. At each pixel all information is combined into early flood warning information [33] in the form of spatial overview maps or time series information at any user defined pixel.

EFAS runs without assimilation of observed real time river discharge data [as for example by 34] which allows the application of this warning system to ungauged catchments [35]. Four warning levels (*low, medium, high and severe*) are derived at each pixel from long-term model simulations with observed meteorological data. The highest simulated discharge of the entire simulated time series represents the highest critical threshold (*severe*) and the 99% quantile, the 2nd highest critical threshold (*high*), the *medium* warning level is defined by 98% quantile and the *low* warning level by the 97% quantile. At each pixel, EFAS determines the highest threshold exceeded during a 24 hour interval. The forecasted discharge distribution is compared against the thresholds at each pixel and threshold exceedance calculated.

Analysis of EFAS results through case studies [for example 32] and statistical skill score analysis [3] has shown that medium-range probabilistic flood forecasts can provide added value information that is complementary to local and national forecasting centres.

2.3 Description of study area and event

The study area is in the country of Romania (central/eastern Europe) and concentratres on rivers Siret, Jium, Olt and Arges which are tributaries of the River Danube. In October 2007 the hydrological regime of almost all the Romanian rivers was above the monthly mean multiannual values. In the last ten days of the month the precipitation regime exceeded the normal values of the whole month due to the high instability of the weather and torrential rainfall, especially during 20th -24th October and so the discharges in the rivers increased. Flooding was reported in parts of the country espesically the south west and east (see figure1).

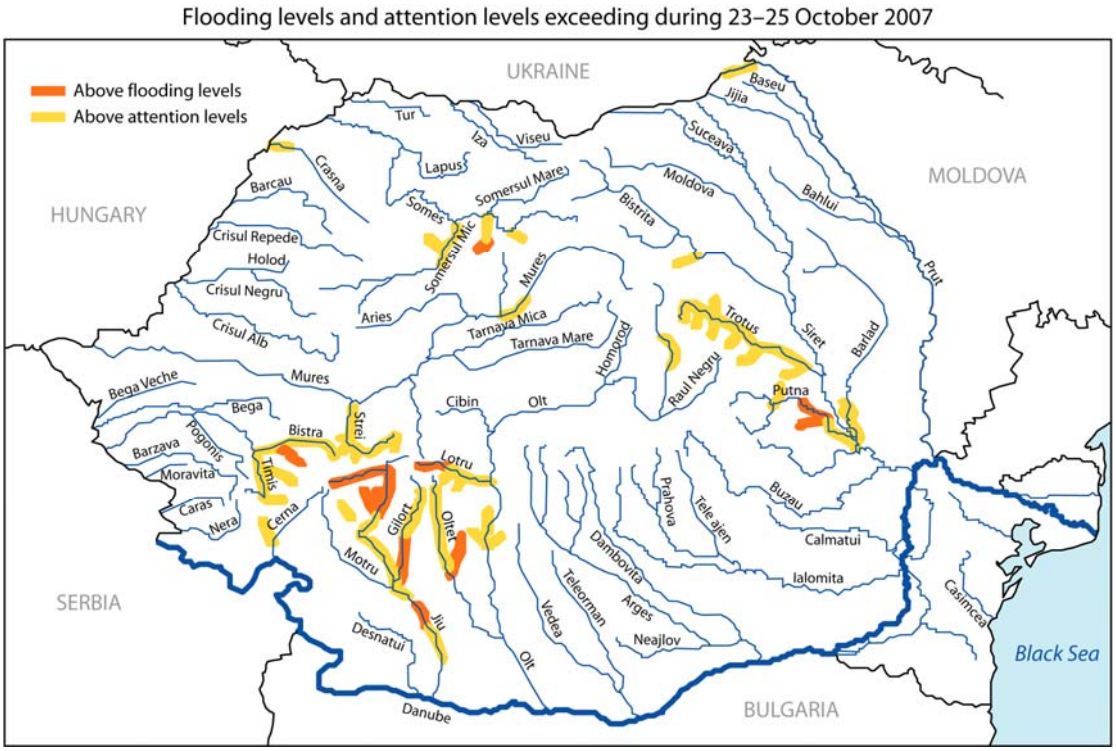

Figure 1: Flood Levels and Attention Levels in Romania exceeded during the 23rd to 25th of October 2007. The basemap for this figure has been kindly provided by E. Anghel from the National Institute of Hydrology and Water Management, Bucharest Romania.

3. Methodology

In this section, the methodology used to evaluate the forecasts will be explained in more detail, the method used to assemble a multi model explained and the difference in the evaluation of precipitation and discharge predictions are discussed, which is essential to understand the results.

3.1 Numerical Methods

The numerical methods are used in this study to evaluate the performance of the forecast systems include:

- **Quantile-quantile** plots in which the distribution of the observed values is plotted against the distribution of the forecasted values. In an EPS the forecasted values outnumber the observed values as each observation is predicted by multiple forecasts. Therefore, in this paper the 5th to 95th percentiles are plotted in steps of 5%. A good fit between both distributions would be indicated by this plot if the plotted values fall onto a straight line.

- **Outlier plots** in which the number of observations above or below the 10th and 90th percentile given by the EPS distribution is shown.

- The **Root Mean Squared Error** (RMSE) between the ensemble mean and the observations.

- The **Rank Probability Score** (RPS) which is equivalent to the Brier score, but measures accuracy of probability forecasts when there are more than two categories [36]:

$$RPS = \frac{1}{k-1}\sum_{k=1}^{k}\left(CDF_{fc,k} - CDF_{obs,k}\right) \qquad (1)$$

  where CDF is the cumulative distribution function of the forecasted or observed precipitation or discharge, k is the number of categories; $CDF_{fc,k}$ represents the probability of the forecasted precipitation below the threshold of k; $CDF_{fc,k}$ represents the probability of the observed precipitation below the threshold of k.

  The threshold for the precipitation score are derived from the 10th to 90th percentile of the observed distribution in 10th percentile steps. The threshold of the discharge score is given by the warning maps of the EFAS system (see section 2.2). The closer the RPS is to zero the better the score.

- **Warning profiles** for individual points are derived by computing the percentage of ensembles above a certain warning level for several consecutive days and lead times.

The methods used in this paper assume that all ensemble members are equally likely. They do not take account of the ensemble size and smaller distributions will have a less well defined cdf especially at the extremes, which are of most interest in flood forecasting. Therefore, results cannot be fully interpreted as probabilities.

3.2 Multi model ensemble (combination)

There have been many attempts to transform the output of EPS systems with error models to derive probability distributions [for a comparison see 37]. However, the low frequency of hydrological extremes and the short time period available in the TIGGE archive does not yet allow such an approach in this research. Nevertheless, it is possible to combine the results of all ensemble forecasts into a multi-model ensemble [38,39]. Evidence suggests that a multi-model concept is superior to the forecasts with individual models [38,39]. There are multiple ways to combine multiple forecasts in hydrology and meteorology and a comprehensive review can be found in Clemen et al. [40]. Abrahart and See [41] illustrated that the improvements achieved with different methodologies varies to a large degree with the dominant hydrological regime. The frequency of flood events as well as the short amount of available forecasts, does not justify any sophisticated approaches [42]. In its simplest form a multi-model ensemble forecast is produced by simply merging the individual forecasts with equal weights [43].

3.3 Comparing precipitation and discharge forecasts

In this paper, precipitation is analysed as the average precipitation upstream of certain locations, which are also used for the discharge predictions and results are compared. Despite this spatial similarity of the two variables, distinctive differences remain:

1. **Catchments are non-linear filters on precipitation inputs**. For example, in the presence of threshold processes, small errors in the precipitation forecast can be amplified or dampened [44,45]. Additionally, discharge predictions are spatial as well as temporal integrations of precipitation fields. Spatial distribution of precipitation fields can influence the shape of a hydrograph, although this will depend on the catchment characteristics as well as the antecedent conditions [46,47]. Temporal integration is caused by varying travel and residence time of water across most catchments. Thus different lead times of precipitation forecasts contribute to different lead times of discharge predictions.

2. **A flow hydrograph is strongly influenced by antecedent conditions**. Antecedent conditions such as soil moisture or ground water levels are more important in analysing discharge forecasts than in precipitation predictions. Especially at small lead times discharge predictions can be significantly influenced by antecedent soil moisture conditions and ground water could become an important control at later stages. No similar considerations are important for precipitation predictions in their application to hydrological models.

3. **Precipitation and discharge also vary in their statistical properties**. Discharge has a higher auto correlation than precipitation. For example, Kann and Haiden [48] have shown that higher autocorrelations positively influence skill scores.

4. **The autocorrelation of precipitation largely depends on the accumulation period** used in the evaluation stage. Pappenberger et al. [49] argued that the accumulation period should reflect the residence time of water in the catchment. However, this will vary from catchment to catchment and maybe difficult to establish in many cases. Therefore, a generic comparison between different catchments and will be only possible if a fixed accumulation time is chosen. Accumulation of discharge forecasts is less desirable when the flood is caused by an overtopping of embankments.,

However, it may be more important if dyke breaching scenarios are of interest or if water balances are analysed [for example 50].

5. **Hydrological models have to be calibrated and are an additional source of uncertainty**, which will have an impact on most of the issues mentioned previously. Moreover, acknowledging the uncertainty in the parameterisation and structure of the hydrological model can dampen or accentuate differences in the comparison of discharge and precipitation predictions. In this paper, we ignore this hydrological uncertainty as we assume that the uncertainty in precipitation forecasts is dominant for this flood event, long lead times and the catchment characteristics, as seen in an earlier study by Pappenberger et al. [5] for the Meuse catchment.

All arguments in this section have to be born in mind at the analysis of the results.

4. Results

In this section the results of the case study will be presented and precipitation and discharge forecasts will be compared.

4.1 Comparison of precipitation and discharge forecasts over the flooded area

In figure 2a and 2b the distribution of the observed and forecasted precipitation is compared with a quantile-quantile plot for each EPS prediction set. The distributions are the same when the dots fall on the straight dashed line.

Figure 2a shows that all ensemble forecasts are over-predicting precipitation at larger precipitation amounts at a lead time of one day. The ensemble forecasts by BOM and JMA are closest to the observations. A more detailed analysis (not shown here) reveals that the frequency of small precipitation amounts forecasted by all EPS is much higher than for observed precipitation (at a lead time of one and five days). At the nine day forecasts (figure 2b) a very mixed picture emerges with an under prediction of all ensembles at lower precipitation amounts. At high precipitation amounts most ensemble forecast systems still show an under prediction. NCEP, CMC and CMA over-predict at high precipitations. The ECMWF forecast follows the centre line most closely at high precipitation amounts. The maximum percentile is under-predicted by nearly all forecast systems but not CMC.

In table 2 the observed divided by forecasted precipitation is averaged for all percentiles above 70% over all lead times from day one to nine. This represents the upper end of the precipitation distribution representing the precipitation which, most likely, contributed most to the flood event.

The table illustrates that the forecasts by UKMO, ECMWF and the Multi-model ensemble are best for this case study. The ensemble forecast by BOM represents the worst precipitation forecast in this case study. In figure 3 the distribution of the discharge observations is compared to the distribution of the forecast models with a quantile-quantile plot.

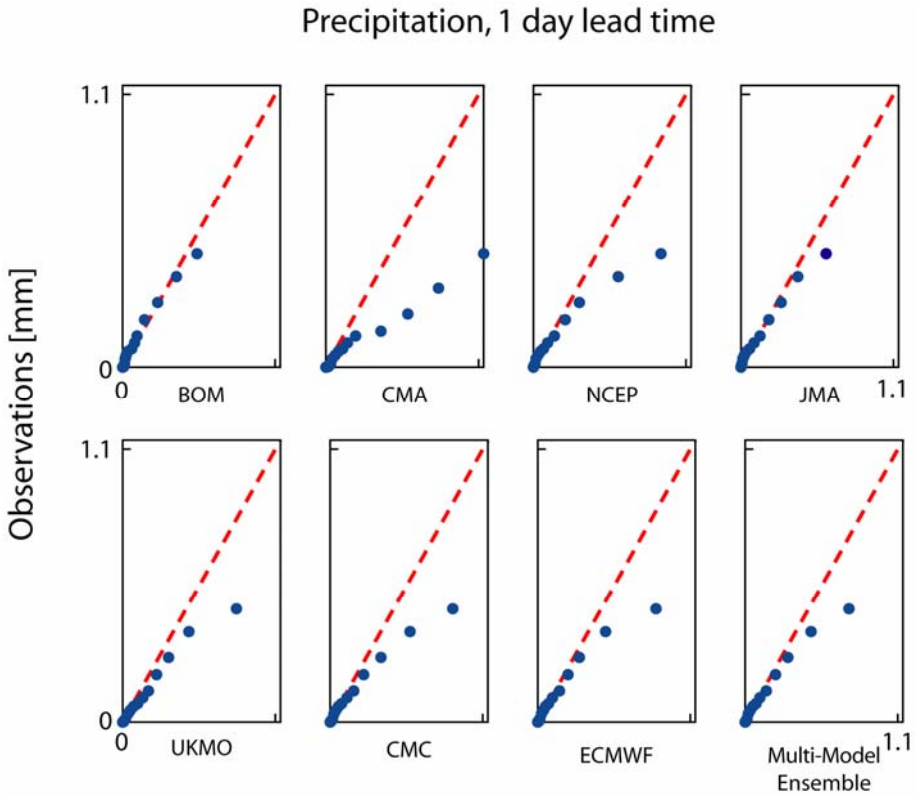

Figure 2a: Quantile-quantile plot for a lead time of 1 day of the forecasted and observed average upstream precipitation.

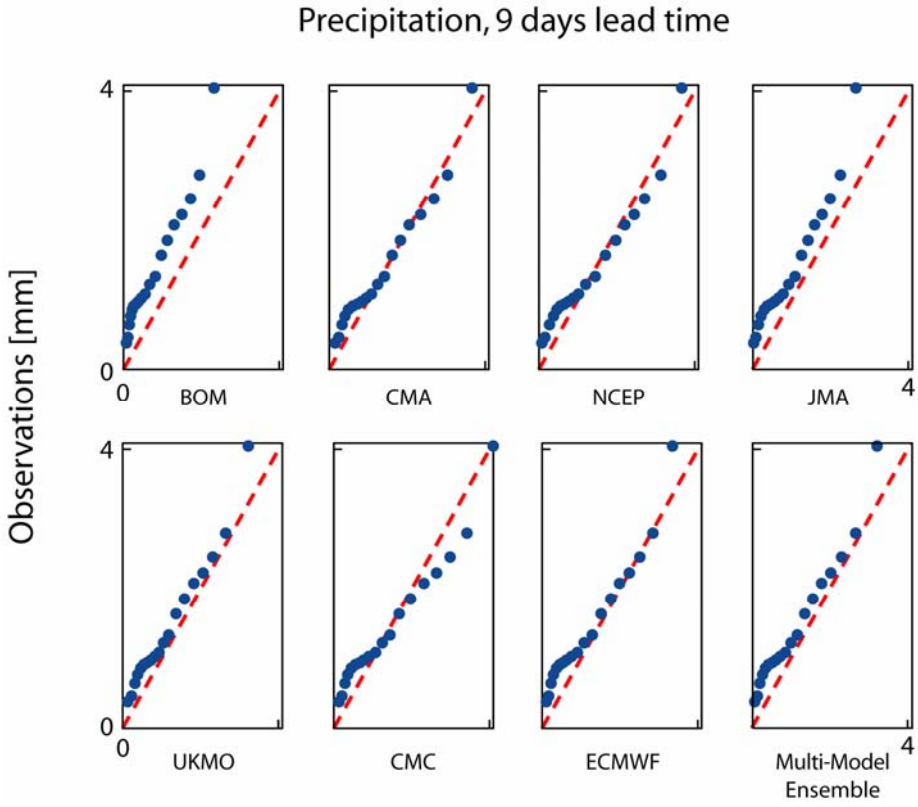

Figure 2b: Quantile-quantile plot for a lead time of 9 days of the forecasted and observed average upstream precipitation.

Table 2: Observed divided by forecasted precipitation averaged for all percentiles above 70% over all lead times from day one to nine

Rank	Centre	Error
1	ECMWF	0.02
2	UKMO	0.03
3	Multi-model	-0.03
4	NCEP	0.10
5	CMC	0.17
6	JMA	-0.17
7	CMC	0.18
8	BOM	-0.40

At a lead time of one day, discharge forecasts follow largely the observed distribution, which suggests a high reliability of the forecasts. Although, the quality of the precipitation forecasts influences the quality of the discharge forecasts, the errors are non-linearly transformed and thus results can expected to differ. All forecast centres seem to over predict at higher (flood related) discharges, which is similar to the distribution shown by precipitation predictions (figure 2a). At this lead time, precipitation, which has already fallen, mainly dominates the discharge predictions and thus the distributions between the different forecast centres look very similar. The systematic over predictions of discharges at the upper end of the distributions is more apparent at longer lead times. This magnitude of over prediction is controlled by the behaviour of the precipitation predictions (figure 2b). In contrast, lower discharges have the same pattern regardless of the quality of the precipitation predictions, which maybe explained by base flow component that is not mainly controlled by the uncertainty of the precipitation, but the hydrological model. The over prediction at large discharge amounts could have several reasons such as initial conditions, calibration of the hydrological model or discrepancy between observed and forecasted spatial rainfall distribution. Similar behaviour has been reported for other areas with different types of hydrological models (for example Pappenberger et al., 2005), which suggests that this error is due to the difference in spatial and temporal error structure between observed and forecasted field. No clear evidence for any of the hypothesis above could be found in this case study. In table 3 the observed divided by forecasted precipitation is averaged for all percentiles above 70% over all lead times from day one to nine. This represents the upper end of the precipitation distribution representing the discharge which, most likely, contributed most to the flood event.

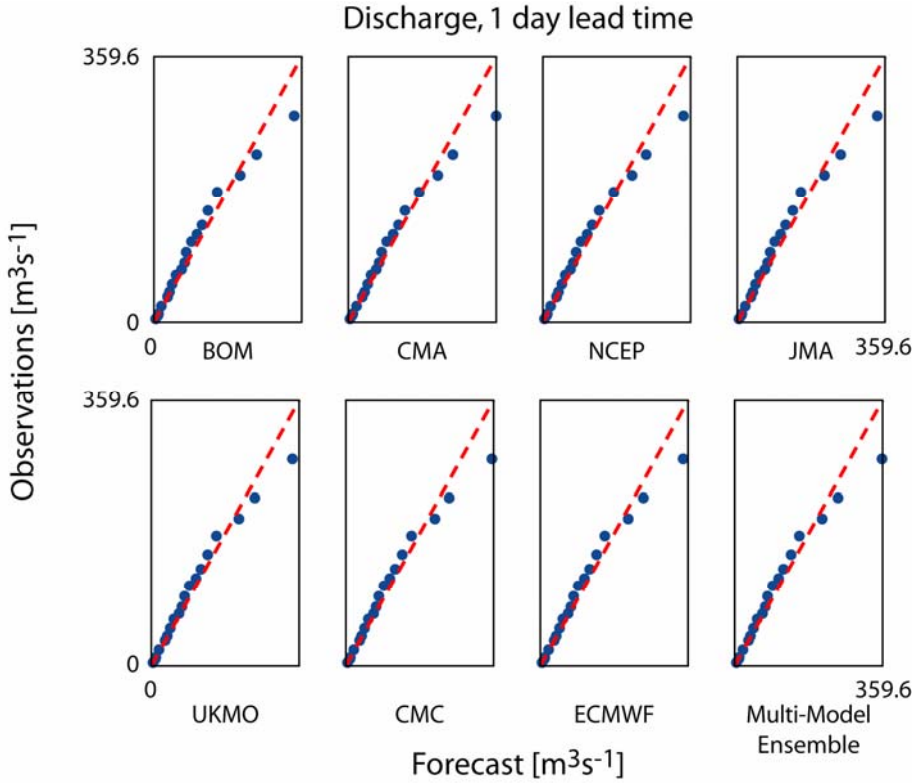

Figure 3a: Quantile-quantile plot for a lead time of 1 day of the forecasted and observed discharge.

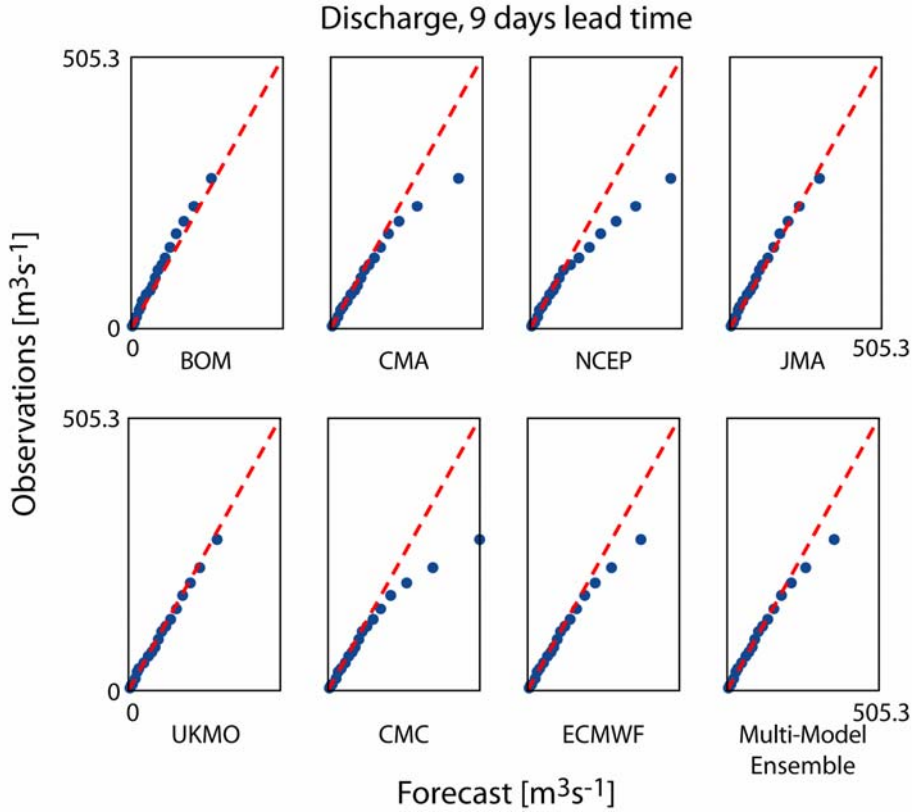

Figure 3b: Quantile-quantile plot for a lead time of 9 days of the forecasted and observed discharge.

Table 3: Observed divided by forecasted discharge averaged for all percentiles above 70% over all lead times from day one to nine

Rank	Centre	Error
1	JMA	0.01
2	UKMO	0.04
3	BOM	-0.05
4	Multi-model	0.06
5	ECMWF	0.08
6	CMA	0.13
7	NCEP	0.17
8	CMC	0.16

The error in discharge predictions is smaller in comparison to the error in precipitation predictions for nearly four of the eight ensemble prediction systems (BOM, CMA, JMA and CMC). NCEP, UKMO, ECMWF and the multi-model ensemble show a small increase. The best forecast is provided by JMA.

In conclusion, none of the forecast centres predict the distribution of precipitation observations exactly, with most centres exhibiting an over-prediction at day nine. The distribution of discharge predictions for observations and forecasts is similar for lower discharges. All centres over-predict discharge for higher discharges with the exception of BOM, which has the tendency to under predict at longer lead times. In a literature review by Pappenberger et al. [51] the maximum error for discharge measurements has been quoted as 8.5%. JMA, UKMO, ECMWF and the multi-model system all have an average error below 8.5% and thus could be seen as suitable EPS to predict this flood within the uncertainty of the measurements.

4.2 Outliers and Root Mean Squared Error (RMSE)

The quantile-quantile figures 2 and 3 ignore timing errors as all observations and forecast are lumped together. In figure 4, the timing error is explored as the percentage of precipitation observations above the 90^{th} percentile of the forecast distribution, the percentage of precipitation observations below the 10^{th} percentile of the forecast distribution and the RMSE is plotted against lead time.

The percentage of observed precipitation above the 90^{th} percentile of the forecasted distribution is very high; between 25% and 40% percent for predictions with a lead time of 1 day falling to below 20% -30% for predictions with a lead time of 9 to 10 days for most forecast centres. Exceptions are the BOM and JMA which both show a less significant drop. The percentage of observations which is below the 10^{th} percentile at a lead time of 1 day is between 30% and 60%. This reflects the findings of the figures above (figure 2 and 3), which indicated a under-prediction of observed precipitation by most forecast centres. The different forecast ensembles have varying number of members and therefore, the accuracy in the quantification of the extreme of the distribution varies. Figure 4 shows no clear dependency in the percentage of observations below or above the 10^{th} and 90^{th} percentile respectively. The percentage of observations below the forecast distributions of all forecast centres is very high and reflects the fact that an extreme event has been observed. The RMSE of the ensemble means increases from an average error of less than 0.1 mm to errors above 0.8

mm. All forecast ensembles show a similar distribution besides BOM, which performs significantly worse than the other centres. ECMWF, UKMO, CMC and the multi-model ensemble perform best with respect to the RMSE of average catchment precipitation. The multi-model ensemble also has a very good overall performance as it has a comparably low number of observations below and above the 10th and 90th percentile respectively.

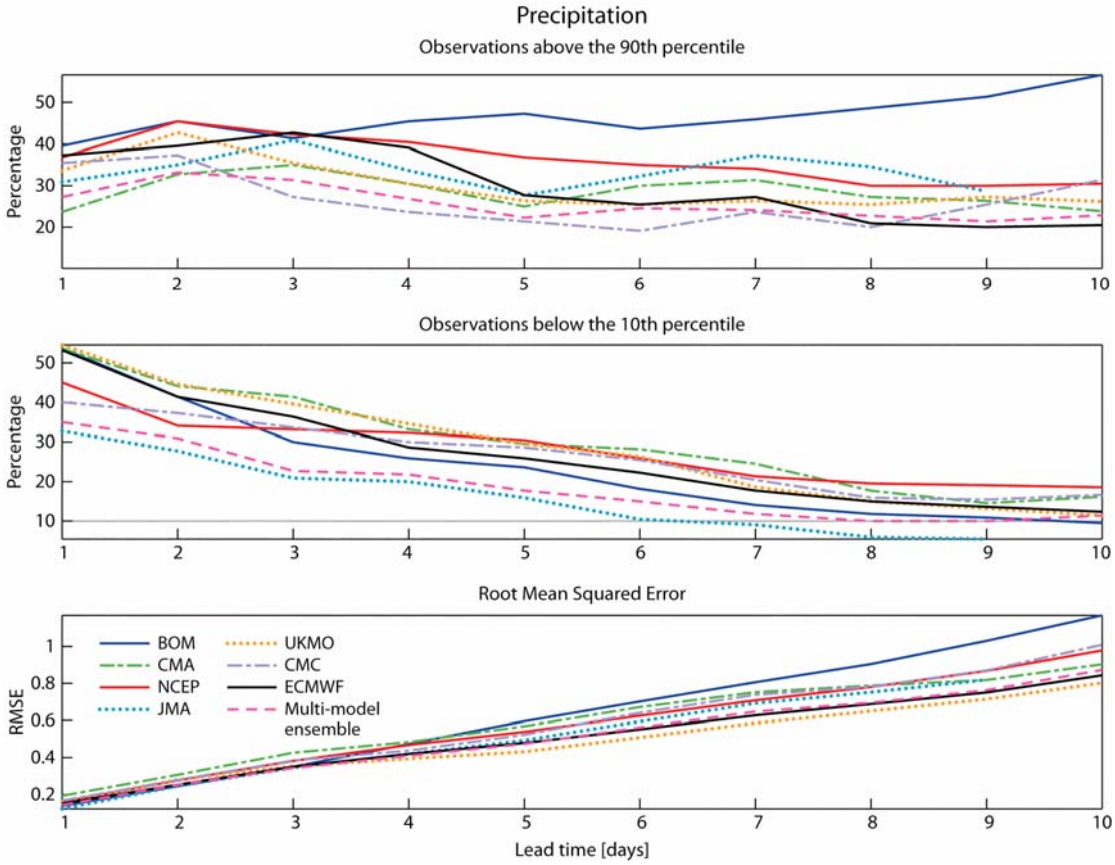

Figure 4: Percentage of observed precipitation above the 90th percentile of the forecast distribution (top figure), the percentage of observed precipitation below the 10th percentile of the forecast distribution (middle figure) and the RMSE of the ensemble mean against lead time (bottom figure).

The multi model ensemble also performs favourably when the analysis is repeated for discharge predictions. In figure 5, the percentage of observed discharge above the 90th percentile of the forecast distribution, the percentage of discharge observations below the 10th percentile of the forecast distribution and the RMSE of the ensemble mean discharge is plotted against lead time.

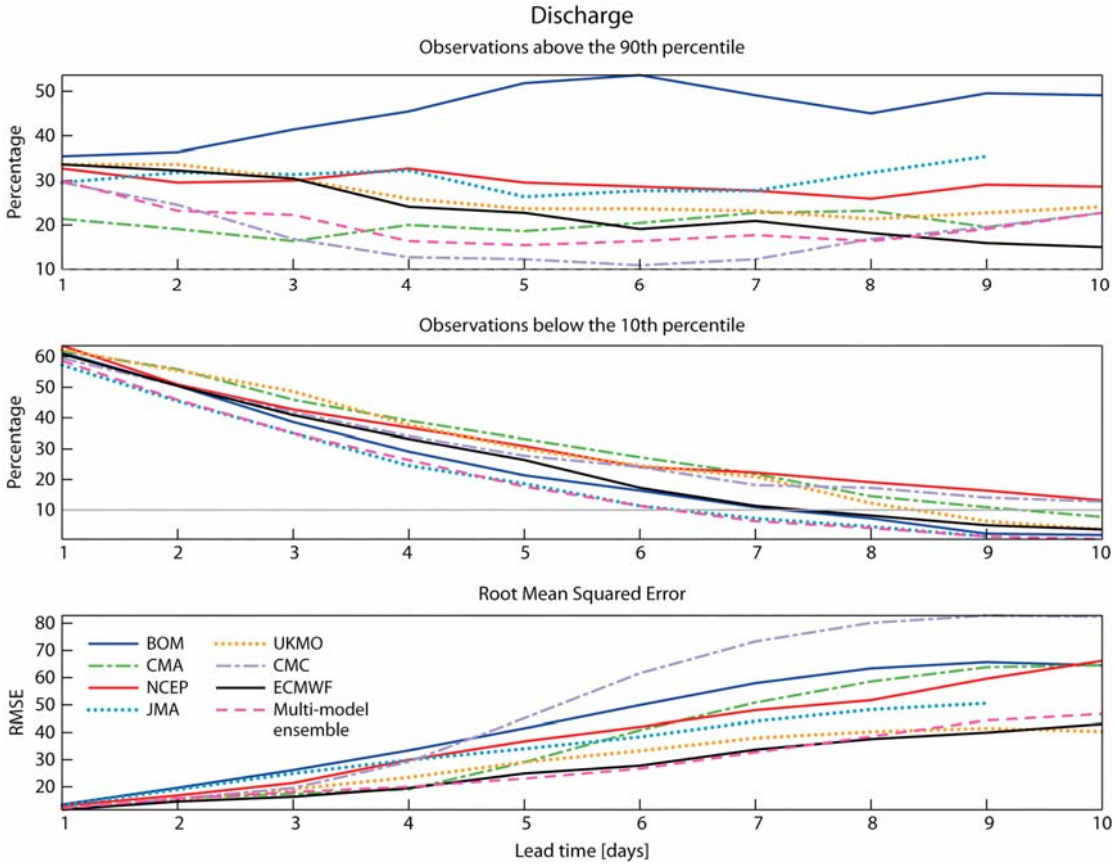

Figure 5: Percentage of observed discharge above the 90th percentile of the forecast distribution (top figure), the percentage of observed discharge below the 10th percentile of the forecast distribution (middle figure) and the RMSE of the ensemble mean against lead time (bottom figure).

Discharge shows in comparison to precipitation a lower number of observations above the 90th percentile, with between 20% and 40% at a lead time of one day and below 30% at a lead time of 10 days. Only, BOM has an increasing percentage over the lead time. The average percentage of observations below the 10th percentile is ~90% and drops below 20% at a lead time of 10 days. The percentage of outliers thus reflects the results above (quantile-quantile plots), with a significant over prediction of discharge by nearly all forecast centres. The RMSE is approximately 10 m^3/s at a lead time of 1 day and raises to values between 30 m^3/s and 60 m^3/s at a lead time of 10 days. The spread of the RMSE of different forecast ensemble mean discharge forecasts is significantly larger than the spread in the average precipitation forecasts. This is explained by the fact that discharge is a variable mainly integrates over time and is measured at one location, whereas upstream precipitation is dominated by spatial averaging. Thus errors can propagate to a much larger degree in discharge predictions. The RMSE of UKMO, ECMWF and the multi model ensemble are superior to the other forecast ensembles especially at long lead times.

The percentage of simulations above and below the 10th and 90th percentile is very high for discharge and precipitation predictions especially at lead times below 5 days. In fact, only a small proportion of the observations is bracketed by the forecasts. UKMO, ECMWF and the multi-model ensemble perform comparably well in terms of RMSE of the ensemble mean discharge and precipitation predictions. The multi-model ensemble also has a comparably low number of observations above and below the 10th and 90th

percentile respectively. This analysis suggests that the multi-model ensemble is the optimal choice to improve discharge predictions.

4.3 Rank Probability Score (RPS)

The previous analysis compared the ensemble predictions and supports the use of a multi model ensemble. However, only one deterministic score (the RMSE) has been computed so far and in what follows the Rank Probability Score is used to analyse the results in a probabilistic manner. In figure 6, the RPS of the upstream average precipitation predictions is computed over thresholds, which are based on the 10^{th}-90^{th} percentiles of the observed average precipitation amounts. The RPS increases from ~0.25 at lead time of 1 day to ~0.45 for lead times of day 10. The ECMWF model performs best at short range lead times and is overtaken by BOM and JMA at a lead time of around day 4-5. UKMO and NCEP perform worse than to the other ensemble systems. The multi model ensemble performs comparably well through all lead times.

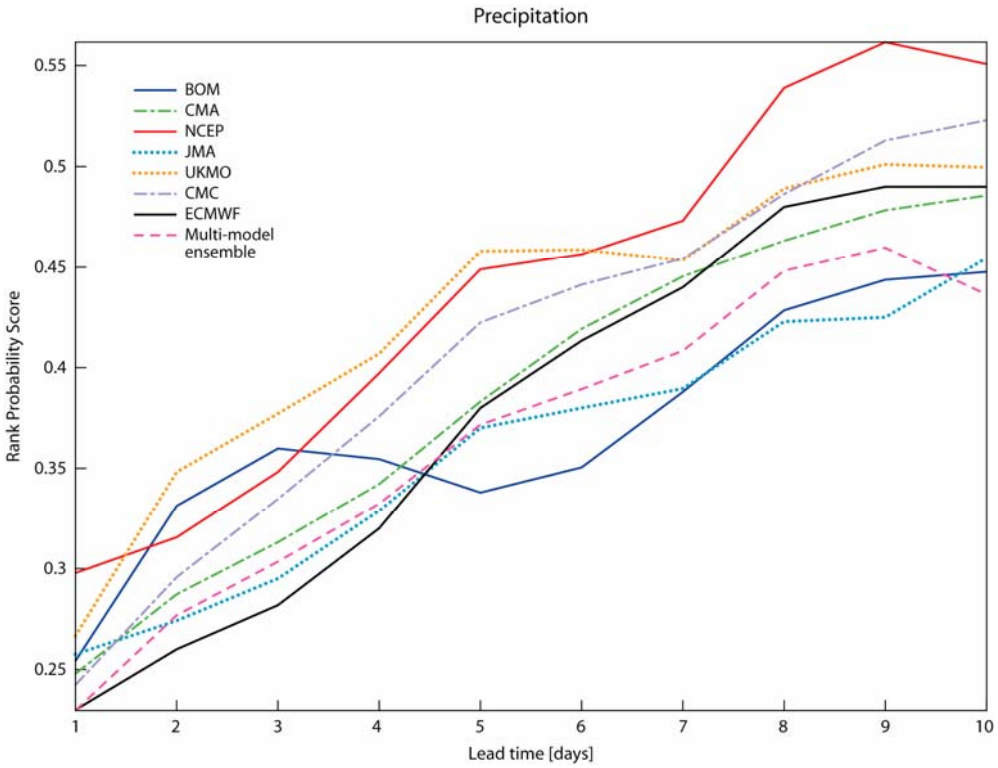

Figure 6: RPS of all ensemble prediction systems for precipitation. The thresholds are the 10th to 90th percentiles of the observed precipitation.

The threshold analysis of the precipitation analysis does not primarily focus on flooding as it is based on the 10^{th} to 90^{th} percentiles of the observed precipitation. The thresholds for the discharge forecasts are based on the four warning levels of low, medium, high and extreme (see section 2.2 for description) available for this area and are therefore, directly flooding related. In figure 7 the RPS for discharge is presented. The BOM model system performs exceptionally well, as does JMA and the multi-model ensemble. The fourth best model is presented by ECMWF. NCEP has the worst performing model. The high performance of BOM is explained by its inactivity. It tends to under predict observed precipitation (see figure 2 and 3) and thus scores well as most of the discharge measurements are below the extreme thresholds.

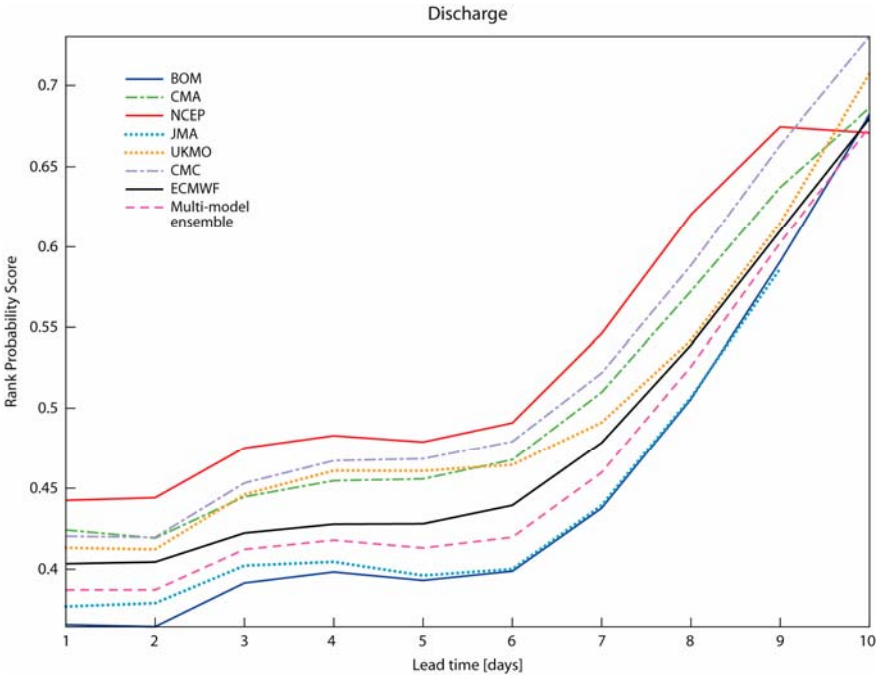

Figure 7: RPS of all ensemble prediction systems for discharge. The thresholds are based on the four warning levels, which are low medium, high and severe.

In summary, the analysis above seems to suggest that a multi-model ensemble should be favoured above the use of any single ensemble prediction system. The multi-model ensemble compares consistently favourable in terms of the percentage of observations bracketed by the 10th and 90th percentile, the RMSE and the RPS for discharge and precipitation.

4.4 Point Flood warnings

Flood warnings are often based on point predictions, representing the risk of an entire area of getting flooded. It is therefore of interest to analyse one location at which flooding has occurred and one in which it did not. Only discharge will be analysed in this section as the general trends in the relation of precipitation forecasts to discharge forecasts have been illustrated above. In figure 8, the 10^{th} and 90^{th} percentile of discharge predictions of the different forecasts with a 2 day lead time are shown. The dashed horizontal lines show the four warning levels. The observations (depicted as stars) clearly exceed these warning levels. In flood forecasting, the rising limb of the hydrograph is one of the most important feature to predict. It is also the most active period of a flood event. Of further importance is the maximum peak discharge and its timing as they mainly influence flood alleviation measures. The distribution of the two day forecast is very small. The uncertainty bounds of CMA, NCEP), CMC, ECMWF and the multi model ensemble bracket the rising limb reasonably well. All of those forecast ensembles also predict the correct timing and magnitude of the peak. BOM and JMA predict the flood one day to late. UKMO does not predict the onset of the flood correctly, but exhibits a good prediction at the peak. All ensemble forecasts predict a too high recession and CMA even seems to predict a second peak. This higher prediction in the recession can have an important impact on flood management as it can influence reservoir management, antecedent conditions for the next event and water release. It is significant that all forecasts exceed the high warning level and thus a correct flood alert could be issued.

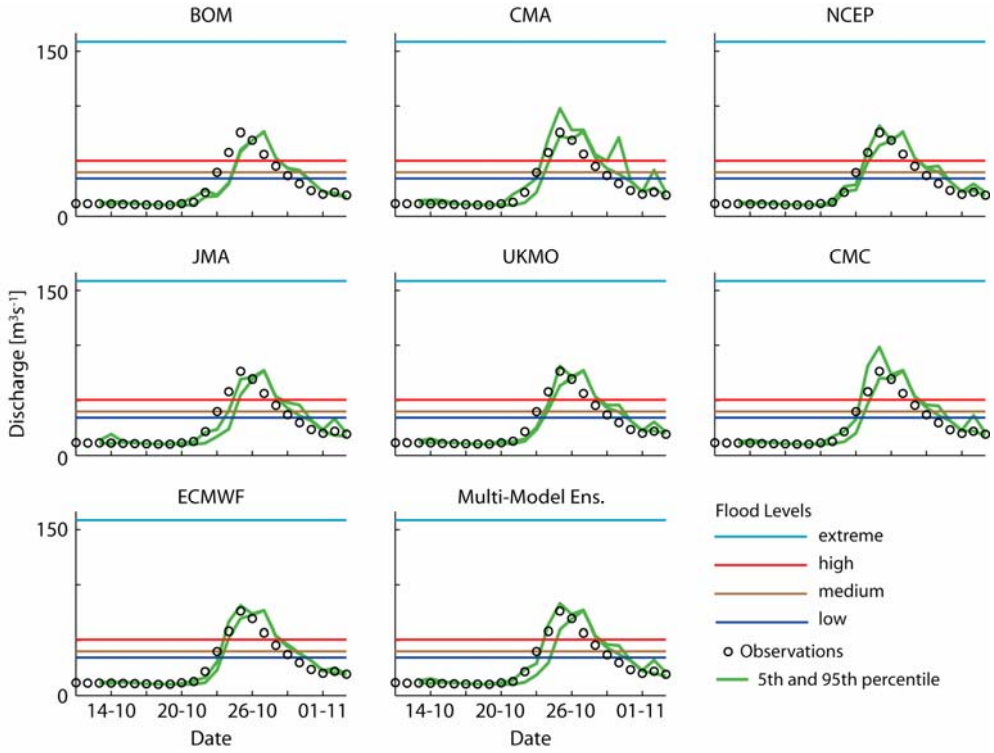

Figure 8: The 10th and 90th percentile of discharge predictions of the different forecasts with a 2 day lead time are shown. The dashed horizontal lines show the four warning levels.

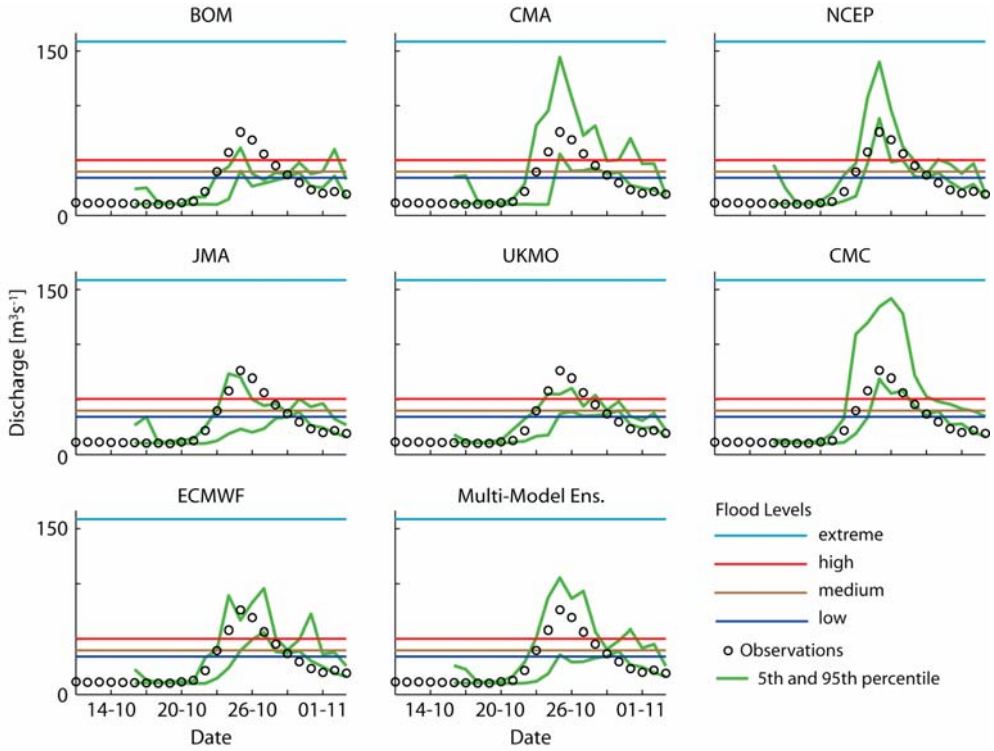

Figure 9: The 10th and 90th percentile of discharge predictions of the different forecasts with a 5 day lead time are shown. The dashed horizontal lines show the four warning levels.

In figure 9, forecasts for a 5 day lead time for the same location are shown. The distributions are significantly larger and can bracket flows below the warning levels as well as far above. CMA, NCEP, JMA, CMC, ECMWF and the multi model ensemble predict the rising limb correctly. CMA, CMC and the multi model ensemble bracket the peak. NCEP and CMC perform very well for the recession limb. The ensemble spread is much larger in comparison to figure 8 and also more observations are bracketed. The widening distribution also means that a lower percentage of models are above the warning levels and thus any issuing of warning has to be based on decreasing number of ensemble members with lead time.

In table 4, the skills of all forecast systems are summarized with a tick at each property which performs adequately. The analysis is subjective based on the experience of the authors as such a short time period prohibits the usage of more complex methodologies. The table shows quite clearly that none of the models has been able to predict the recession for a lead time of 2 days adequately. Table 2 indicates that one would not choose the forecasts by JMA and BOM for this location. The table also indicates that a multi model combination strategy for hydrological applications may have to compute weights which optimize the combination to get these desired properties. For example, some forecasts are better in predicting recession limbs than others and thus should have more weight in a combined analysis.

Table 4: Skill of all forecast systems for the lead time of 2 and 5 days in respect to representations of the rising limb, peak discharge, timing of the peak discharge and recession. A tick indicates that the process has been represented adequately.

	Rising Limb		Peak Discharge		Timing		Recession	
	2	5	2	5	2	5	2	5
BOM			√					
CMA	√	√	√	√	√	√		
ECMWF	√	√	√		√	√		
NCEP	√	√	√	√	√	√		√
JMA			√					√
UKMO		√	√		√			
CMC	√	√	√	√	√	√		√
Multi-model	√	√	√	√	√	√		√

Of further interest are false alarm rates, which will be analysed with respect to another location shown in figure 10 for a forecast with a two days lead time.

Figure 10 shows a point at which all observations are clearly below the warning levels. In fact only a gently increasing hydrograph can be seen. The most important feature of this forecast is whether the alarm levels are exceeded. Only, BOM and JMA are staying well below all alert levels and only CMA exceeds the high alert level. CMA and UKMO have a high proportion of their forecast above the low alert level, whereas other ensemble systems seem to have a high percentage of forecasts below the lowest alert level.

A larger number of ensembles exceed the lowest alert level at forecasts with a lead time of 5 days (see figure 11). CMA, CMC and ECMWF all exceed the high alert level and only the BOM model stays significantly below. At the previous (‘flooding’) location BOM also under predicted discharge, however, this cannot be verified to be consistent property as we only investigate two locations. Not only hydrographs are used for flood warning, but also probability of exceedance maps.

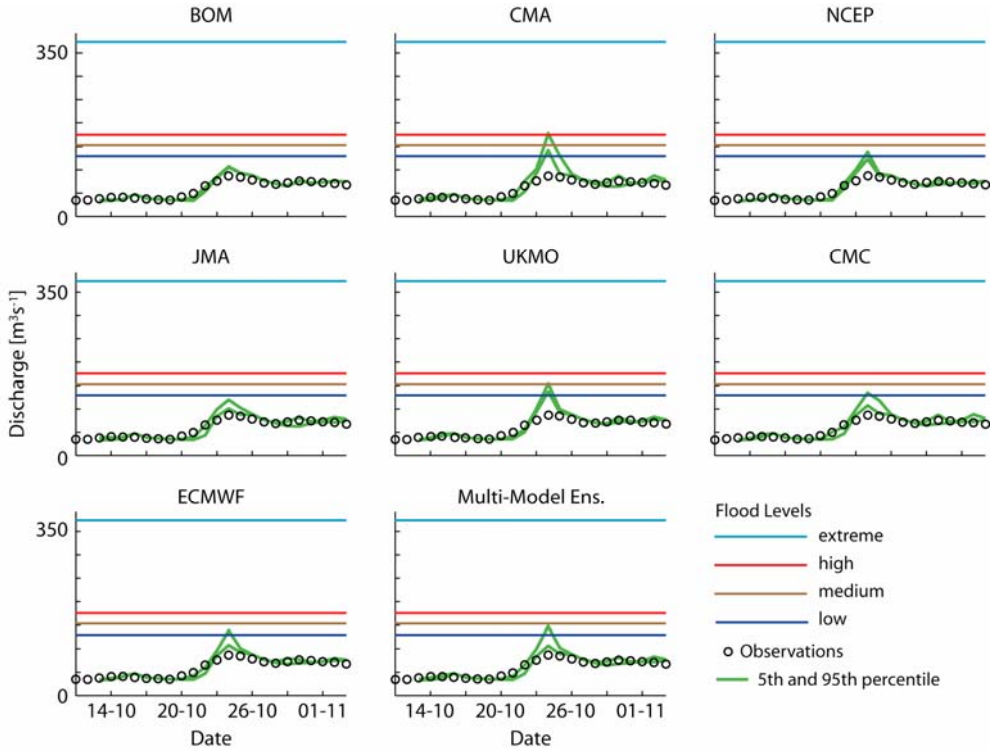

Figure 10: The 10th and 90th percentile of discharge predictions of the different forecasts with a 2 day lead time are shown. The dashed horizontal lines show the four warning levels.

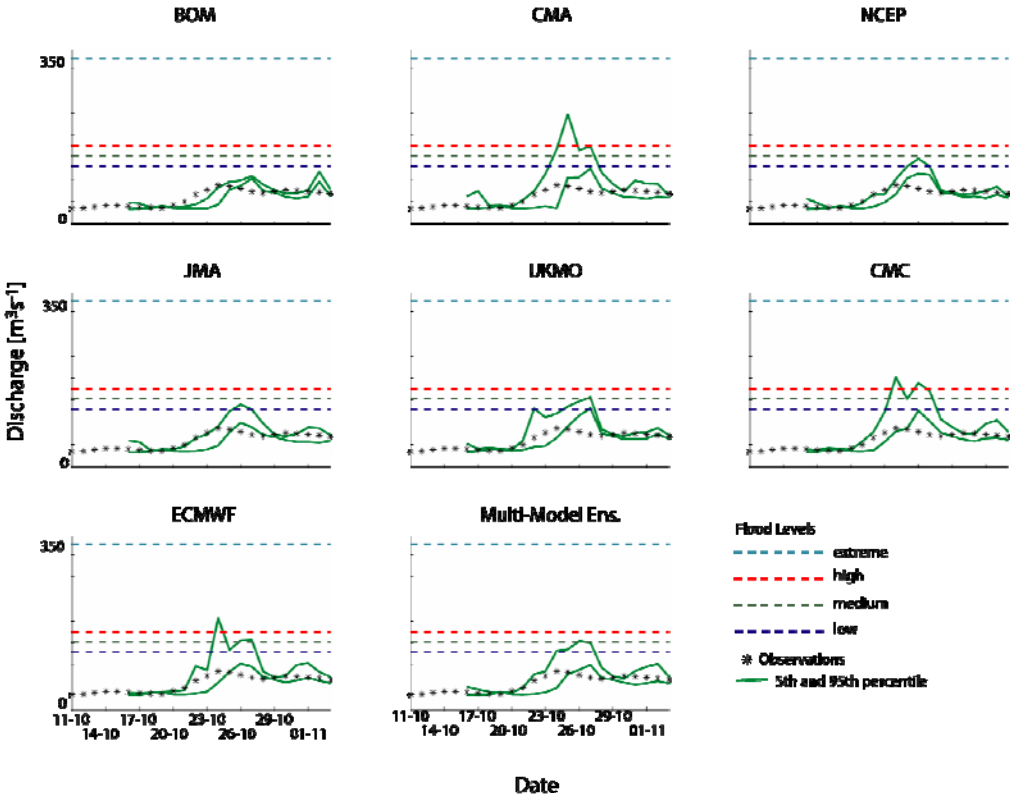

Figure 11: The 10th and 90th percentile of discharge predictions of the different forecasts with a 5 day lead time are shown. The dashed horizontal lines show the four warning levels.

In Figure 12, the probability of exceedance of the highest warning level for each forecast centre for 13 consecutive forecast dates is shown. This figure concentrates on the onset of the flood (24th of October) and therefore only shows the forecasts for the 11th of October to the 23rd of October. The exceedance levels indicate that most forecast ensembles predict flooding from the 14th of October to the 17th of October. The signal re-occurs from forecast to forecast, which provides the necessary reassurance. This type of persistency is one method used by EFAS [2] to decide whether warnings will be issued. From the 19th October onwards, the signal is very strong, although initially the flooding is predicted one day too early. This means that there is an efficient flood warning five days in advance and a possible warning 8 days in advance. The only ensemble prediction system used in the current EFAS set up is produced by ECMWF. No clear advantage over a warning based on ECMWF forecasts in comparison to a multi-model approach can be seen. A forecast based only on the ensemble of BOM would probably have led to no warning as the signal was inconsistent. Therefore, the use of multiple systems allows for more reassurance and could have led to a better forecast for this location and event.

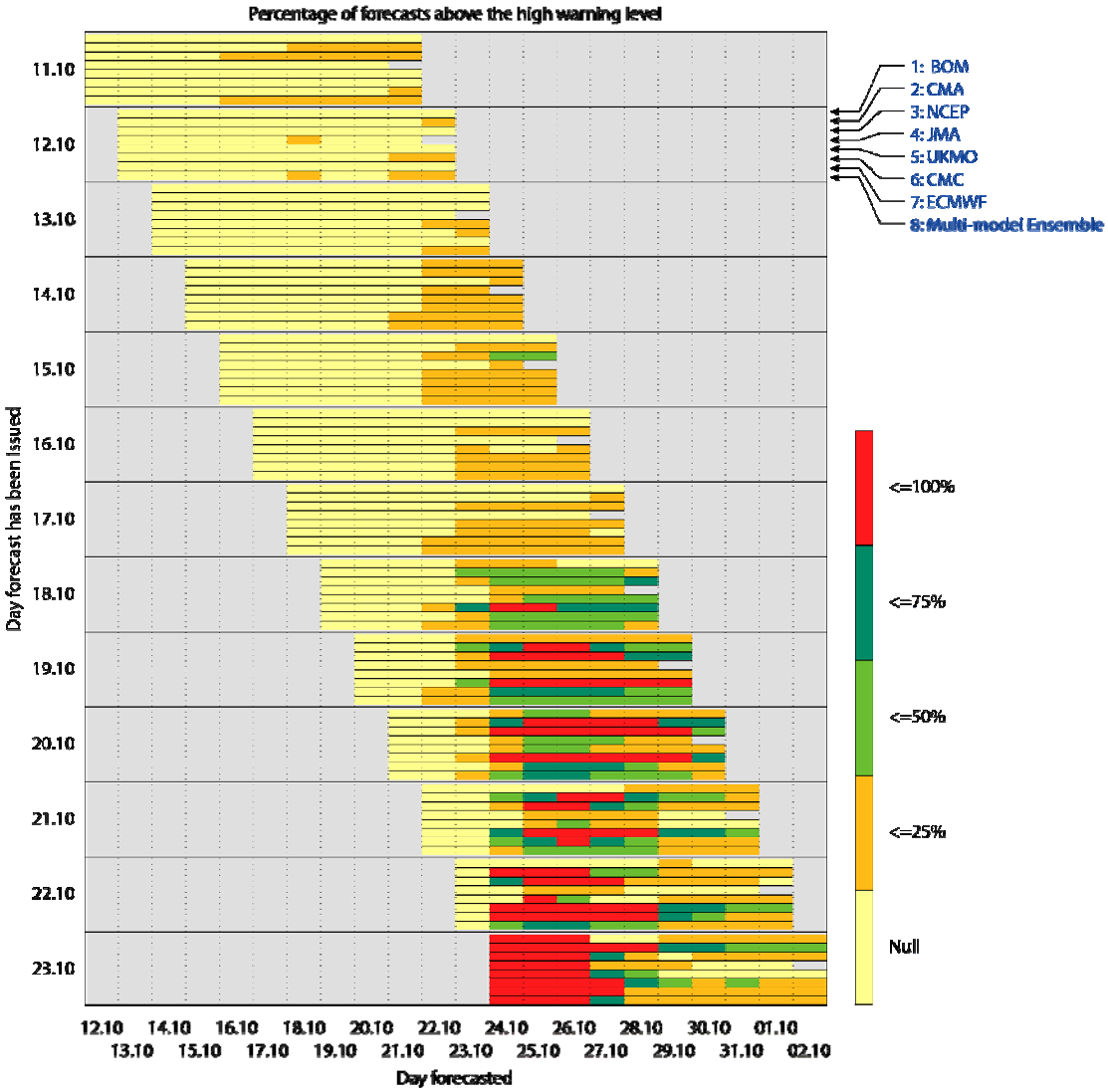

Figure 12: Percentage of forecasts exceeding the high thresholds from the 11th October to the 23rd October for all forecast systems.

In figure 13 the threshold exceedance for the location with no flood is shown. Some ensemble systems exceed the threshold as early as on the 14th, however, the percentage remains low and only a limited number of ensemble systems (3 out of 8) predict flooding. A forecast system based solely on the ECMWF model would have probably prompted any forecaster to increase the attention at this location. However, a comparison with the other forecasts would have prevented any issuing of an alert.

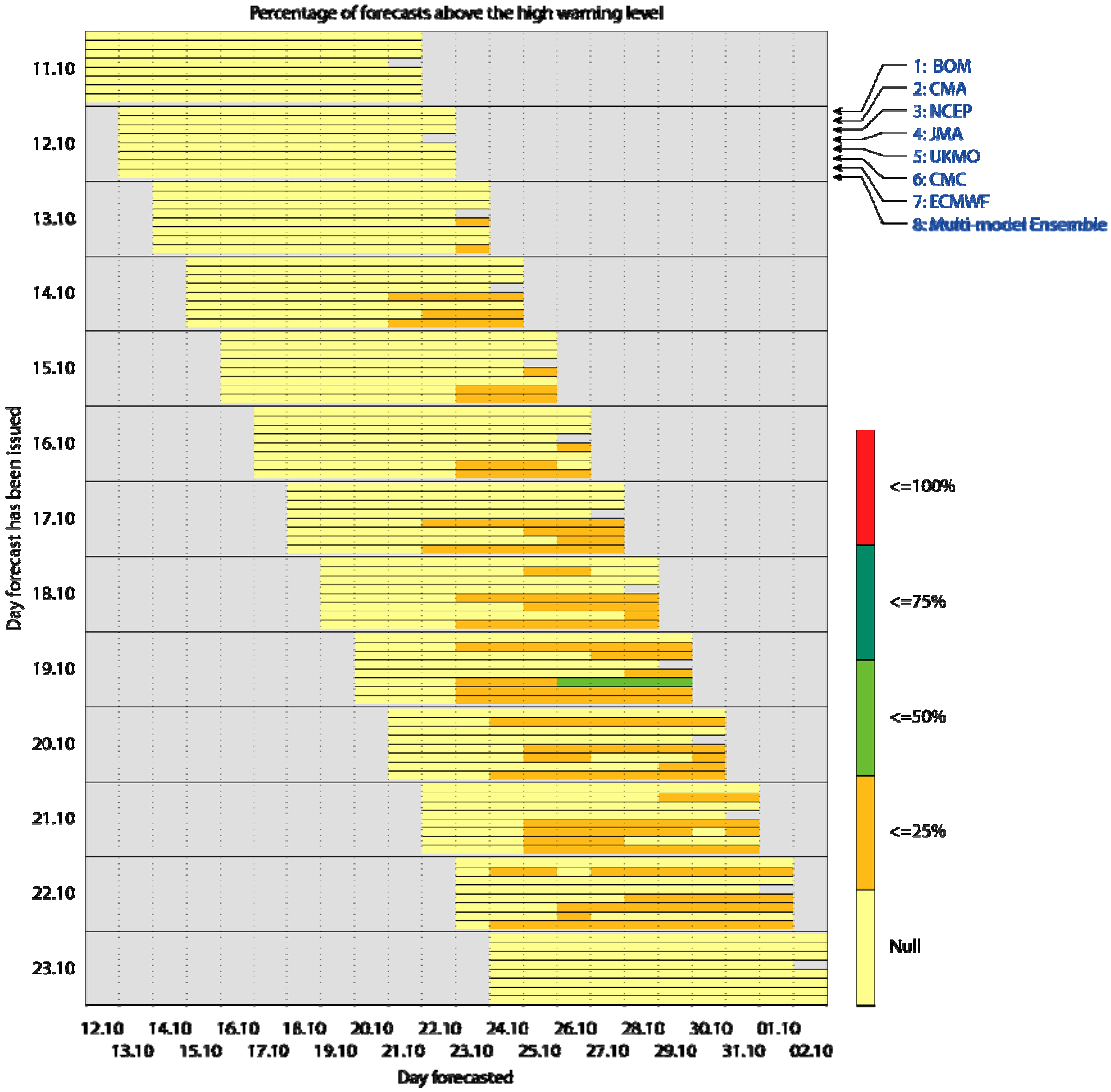

Figure 13: Percentage of forecasts exceeding the high thresholds from the 11th October to the 23rd October for all forecast systems.

5. Discussion

This paper evaluates ensemble prediction systems of weather forecasts as input into a hydrological model to predict flooding of a specific case study. Different aspects of the forecasts are presented including the analysis of two specific locations in the area of interest. In table 5 the analysis is summarized by computing the rank of each forecast system for each aspect of the analysis. For example the ranks of each system according to the precipitation error shown in table 2 is computed and summarized in table 5. The lower a rank the better is the forecast system. Some ranks are not integers as they have been computed over a time series (the rank for each forecast has been computed at each lead time and then averaged).

Table 5: Summary of performance expressed in ranks according to each criteria

Description	BOM	CMA	NCEP	JMA	UKMO	CMC	ECMWF	Multi-model	Detailed Description
Quantile Error (precipitation)	8	7	4	6	**2**	5	**1**	3	Rank according to absolute error in table 2
RMSE (precipitation)	6	6.8	5.5	3.7	**2.4**	6.1	2.9	**2.6**	Average Rank over all lead times of RMSE in figure 4
RPS (precipitation)	3.3	4.3	7.3	**2.3**	7.2	5.8	3.2	**2.6**	Average Rank over all lead times of RPS in figure 6
Quantile Error (discharge)	3	6	8	**1**	**2**	7	5	4	Rank according to absolute error in table 3
RMSE (discharge)	7.1	4.1	5.8	5.6	3	6.4	**1.5**	**2.5**	Average Rank over all lead times of RMSE in figure 5
RPS (discharge)	**1.1**	5.9	8	**1.9**	4.3	6.8	4	3	Average Rank over all lead times of RPS in figure 7
Location 1	8	4	**1**	7	6	1	5	**1**	Ranked after the number of OKs in table 4
Average	**5.2**	**5.4**	**5.7**	**4**	**3.8**	**5.4**	**3.2**	**2.7**	**Average of above**

The average over all analysis is computed in the last line. It demonstrates that the Multi-model ensemble has the best average properties, followed by the forecast of ECMWF and UKMO. NCEP provides the worst average forecast. The increased quality achieved by the multi-model ensemble may well be explained by the fact that a flood forecast is investigated and thus the multi-model ensemble may provide a better approximation at the tails of the distribution. This is based on a set of criteria specific for this case study and might be different in more general cases.

The case study does demonstrate the potential benefit of a multi-model forecasting system and has implications of future work. This includes the development of comprehensive framework which all uncertainties in this model cascade (numerical weather predictions, calibration of error models, factors of hydrological models and the merging of multi model ensembles) are treated in the same uncertainty framework. Each component of this cascade also needs additional attention. The error model of the numerical weather prediction models has to be tuned to hydrological applications. The uncertainty of all hydrological factors introduced. Various approaches to merge ensemble predictions tested.

6. Conclusions

None of the forecast centres (ECMWF, UKMO, JMA, NCEP, CMA, CMC, BOM and a multi-model ensemble) predict the distribution of precipitation observations exactly, with most centres exhibiting an over prediction at day nine. The distribution of discharge predictions for observations and forecasts is similar at the lower flown. All centres over predict discharge at high flow with the exception of BOM, which has the tendency to under predict at longer lead times. JMA, UKMO, ECMWF and the multi-model system all have an average error below 10% and thus can be seen as suitable EPS systems to predict this flood.

The percentage of simulations above and below the 10^{th} and 90^{th} percentile of the predicted EPS distributions is very high for discharge and precipitation predictions especially at lead times below 5 days. In fact, only a small proportion of the observations is bracketed by the forecasts. UKMO, ECMWF and the multi-model ensemble perform comparably well in terms of root mean squared error of the ensemble mean discharge and

precipitation predictions. The multi-model ensemble also has a comparably low number of observations above and below the 10th and 90th percentile respectively. This analysis suggests that the multi-model ensemble is the optimal choice to improve discharge predictions.

In this paper the forecasts for two individual locations at the rivers are shown. At the first location flooding has been observed. CMA, NCEP, JMA, CMC, ECMWF and the multi model ensemble predict the rising limb correctly. CMA, CMC and the multi model ensemble bracket the peak. NCEP and CMC perform very well at the recession limb. None of the models has been able to predict the recession for a lead time of 2 days adequately. It can be shown that one would not choose the forecasts by JMA and BOM for this location. Signal re-occurrence from one forecast to another is generally used by hydrologists to decide whether flood warnings will be issued (or retracted). All forecasts (apart from the one issued by BOM) would have lad to a flood warning about 8 days in advance.

At the second location no flooding has been observed. Only, the forecasts by BOM and JMA are staying well below all alert levels and only CMA exceeds the high alert level. CMA and UKMO have a high proportion of their forecast above the low alert level, whereas other ensemble systems seem to have a high percentage of simulations below the lowest alert level. An analysis of persistency suggest that no warning would have been issued for this location if all forecasts would be considered (which is correct).

In a final analysis, the different properties of the different forecasts have been ranked and an average rank computed. It demonstrates that the Multi-model ensemble has the best average properties, followed by the forecast of ECMWF and UKMO. NCEP provides the worst average forecast. The increased quality achieved by the multi-model ensemble may well be explained by the fact that a flood forecast is investigated and thus the multi-model ensemble may provide a better approximation at the tails of the distribution. This is based on a set of criteria specific for this case study and might be different in more general cases.

Acknowledgments

Florian Pappenberger was funded by the PREVIEW project. We thank the seven forecast centres to provide data for a data base, which is also useful for hydrological applications. We further would like to thank Dr. Hannah Cloke (Kings College London) for her insightful review, which helped to improve the manuscript. We thank Elena Anghel from the National Institute of Hydrology and Water Management, Bucharest Romania for her valuable help and provision of figure 1.

References

1. de Roo A, Gouweleeuw B, Thielen J, et al. Development of a European Flood Forecasting System. *Int J of River Basin Management* 2003;**1**:49-59

2. Thielen J, Bartholmes J, Ramos M-H, de Roo A. The European Flood Alert System. I. Concept and development. *Hydrological and Earth System Sciences Discussions* 2007

3. Bartholmes J, Thielen J, Ramos H, Gentilini S. The European Flood Alert System EFAS - Part II. Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrological and Earth System Sciences Discussions* 2007

4. EM-DAT. The OFDA/CRED International Disaster Database. Université Catholique de Louvain, Brussels, Belgium. 2007. www.em-dat.net

5. Pappenberger F, Beven KJ, Hunter N, et al. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Science* 2005;**9**:381-393

6. Gourley JJ, Vieux BE. A method for evaluating the accuracy of quantitative precipitation estimates from a hydrologic modeling perspective. *J Hydrometeorol* 2005;**6**:115-133

7. Krzysztofowicz R. Bayesian system for probabilistic river stage forecasting. *J Hydrol* 2002;**268**:16-40

8. Ahrens B, Jaun S. On evaluation of ensemble precipitation forecasts with observation-based ensembles. *Advances in Geosciences* 2007;**10**:139-144

9. Verbunt M, Zappa M, Gurtz J, Kaufmann P. Verification of a coupled hydrometeorological modelling approach for alpine tributaries in the Rhine basin. *J Hydrol* 2006;**324**:224-238

10. Gouweleeuw B, Thielen J, de Roo A, Buizza R. Flood forecasting using probabilistic weather predictions. *Hydrology and Earth System Science* 2005;9, 365-380

11. Demeritt D, Cloke H, Pappenberger F, et al. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards*;In Press, Corrected Proof

12. Martini F, de Roo A. Good Practice for delivering flood-related information to the general public. In, *European Exchange Circle on flood forecasting, early warning: Flood forecasting in Europe: Current practices, needs and proposed actions to go forward*, http://exciffjrcit/downloads/exciff-related-documents/EXCIFF_Current_practices_Doc1_2005doc European Commission EUR22760 EN; 2007

13. Palmer TN, Doblas-Reyes FJ, Hagedorn R, Weisheimer A. Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society B-Biological Sciences* 2005;**360**:1991-1998

14. Zhu YJ, Toth Z, Wobus R, Richardson D, Mylne K. The economic value of ensemble-based weather forecasts. *Bull Am Meteor Soc* 2002;**83**:73-84.

15. Roulin E. Skill and relative economic value of medium-range hydrological ensemble prediction. *Hydrol Earth Syst Sci* 2007;**11:** 725-737

16. Buizza R, Richardson DS, Palmer TN. The new 80-km high-resolution ECMWF EPS. *ECMWF Newsletter* 2001;**90**:2-9

17. Buizza R, Miller M, Palmer TN. Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Q J Roy Meteor Soc* 1999;**125**:2887-2908

18. Houtekamer PL, Mitchell HL. Data assimilation using an ensemble Kalman filter technique. *Mon Wea Rev* 1998;126:796-811

19. Toth Z, Kalnay E. Ensemble forecasting at NCEP and the breeding method. *Mon Wea Rev* 1997;**125**:3297-3319

20. Houtekamer PL, Lefaivre L, Derome J, Ritchie H, Mitchell HL. A system simulation approach to ensemble prediction. *Mon Wea Rev* 1996;**124**:1225-1242

21. Molteni F, Buizza R, Palmer TN, Petroliagis T. The ECMWF Ensemble Prediction System: methodology and validation. *Q J Roy Meteor Soc* 1996;**122**:73-119

22. Toth Z, Kalnay E. Ensemble Forecasting at Nmc - the Generation of Perturbations. *Bull Am Meteor Soc* 1993;**74**:2317-2330

23. Hopson T, Webster P. ThreeTier flood and precipitation forecasting scheme for south-east Asia http://cfab2.eas.gatech.edu/ . 2008

24. Vehvilainen B, Huttunen M. Hydrological forecasting and real time monitoring in Finland: The watershed simulation and forecasting system (WSFS). Helsinki; 2002

25. Olsson J, Lindström G. Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *J Hydrol* 2007;in press

26. Fritsch JM, Hilliker J, Ross J, Vislocky RL. Model consensus. *Wea & Forec* 2000;**15**:571-582

27. Doblas-Reyes FJ, Hagedorn R, Palmer TN. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. *Tellus Series A*- 2005;**57**:234-252

28. Goswami M, O'Connor KM, Bhattarai KP. Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment. *J Hydrol* 2007;**333**:517-531

29. Beven KJ. A manifesto for the equifinality thesis. *J Hydrol* 2006;**3**20:18-36

30. Clemen RT, Murphy AH. Objective and Subjective Precipitation Probability Forecasts: Some Methods for Improving Forecast Quality. *Weather Forecsting* 1986;**1**:213-218

31. Goswami M, O'Connor KM. Real-time flow forecasting in the absence of quantitative precipitation forecasts: A multi-model approach. *J Hydrol* 2007;**334**:125-140

32. Kalas M, Ramos M-H, Thielen J, Babiakova G. Evaluation of the medium-range European flood forecasts for the March - April 2006 flood in the Morava River. *J Hydrol & Hydromech* 2008;in press

33. Ramos MH, Bartholmes J, Thielen-del Pozo J. Development of decision support products based on ensemble weather forecasts in the European Flood Alert System. *Atmospheric Science Letters* 2007;**8**:113-119

34. Komma J, Reszler C, Bloschl G, Haiden T. Ensemble prediction of floods - catchment non-linearity and forecast probabilities. *Natural Hazards and Earth System Sciences* 2007;**7**:431-444

35. Sivapalan M, Takeuchi K, Franks SW, et al. IAHS decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. *Hydrol Sci J* 2003;**48**:857-880

36. Jolliffe IT, Stephenson DB. *Forecast verification: a practitioner's guide in atmospheric science*. Chichester: J. Wiley; 2003:xiii, 240 p.

37. Wilks DS, Hamill TM. Comparison of ensemble-MOS methods using GFS reforecasts. *Mon Wea Rev* 2007;135:2379-2390

38. Doblas-Reyes FJ, Deque M, Piedelievre JP. Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q J Roy Meteor Soc* 2000;**126**:2069-2087

39. Palmer TN. Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics* 2000;**63**:71-116

40. Clemen RT. Combining forecasts: A review and annotated bibliography. *Internat J of Forecasting* 1989;**5**:559-583

41. Abrahart RJ, See L. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrol Process* 2000;**14**:2157-2172

42. Young PC, Parkinson S, Lees M. Simplicity out of complexity in environmental modelling: Occam's razor revisited. *J Appl Stat* 1996;**23**:165-210

43. Hagedorn R, Doblas-Reyes FJ, Palmer TN. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus Series A*- 2005;**57**:219-233

44. Bloschl G, Zehe E. Invited commentary - On hydrological predictability. *Hydrol Process* 2005;**19**:3923-3929

45. Zehe E, Elsenbeer H, Lindenmaier F, Schulz K, Bloschl G. Patterns of predictability in hydrological threshold systems. *Water Resour Res* 2007;**43** W07434, doi:10.1029/2006WR005589

46. Segond M-L. Stochastic modelling of space-time rainfall and the significance of spatial data for flood runoff generation. Thesis, Department of Civil and Environmental Engineering. London: Imperial College London; 2006:222

47. Beven KJ, Hornberger GM. Assessing the Effect of Spatial Pattern of Precipitation in Modelling Stream-Flow Hydrographs. *Water Resources Bull* 1982;**18**:823-829

48. Kann A, Haiden T. The August 2002 flood in Austria: sensitivity of precipitation forecast skill to areal and temporal averaging. *Meteorologische Zeitschrift* 2005;**14**:369-377

49. Pappenberger F, Scipal K, Buizza R. Hydrological aspects of meteorological verification. *Atmospheric Science Letters* 2007

50. Pappenberger F, Buizza R. The skill of ECMWF predictions for hydrological modelling. ECMWF Technical Memorandum, No.558. 2007

51. Pappenberger F, Matgen P, Beven KJ, et al. Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources* 2006; **29**:1430-1449