HAL
open science

Skill and relative economic value of medium-range hydrological ensemble predictions

E. Roulin

To cite this version:

E. Roulin. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrology and Earth System Sciences Discussions, 2007, 11 (2), pp.725-737. hal-00305048

HAL Id: hal-00305048

https://hal.science/hal-00305048v1

Submitted on 18 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hydrol. Earth Syst. Sci., 11, 725–737, 2007
www.hydrol-earth-syst-sci.net/11/725/2007/
© Author(s) 2007. This work is licensed
under a Creative Commons License.

Hydrology and Earth System Sciences

Skill and relative economic value of medium-range hydrological ensemble predictions

E. Roulin

Institut Royal Météorologique de Belgique, Brussels, Belgium

Received: 21 February 2006 – Published in Hydrol. Earth Syst. Sci. Discuss.: 10 July 2006
Revised: 23 October 2006 – Accepted: 22 December 2006 – Published: 17 January 2007

Abstract. A hydrological ensemble prediction system, integrating a water balance model with ensemble precipitation forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS), is evaluated for two Belgian catchments using verification methods borrowed from meteorology. The skill of the probability forecast that the streamflow exceeds a given level is measured with the Brier Skill Score. Then the value of the system is assessed using a cost-loss decision model. The verification results of the hydrological ensemble predictions are compared with the corresponding results obtained for simpler alternatives as the one obtained by using of the deterministic forecast of ECMWF which is characterized by a higher spatial resolution or by using of the EPS ensemble mean.

1 Introduction

The forecast of future precipitation amounts is one of the most challenging tasks of weather forecasting. In rainfall dominated hydrological regimes, it has long been the main limitation to flood forecasting at the medium-range and beyond. The improvement of numerical weather prediction has been made possible by the increase of spatial resolution, the assimilation of new remote-sensed information, better physics parameterizations, all being sustained by a continuous growth of computing power. An alternative use of these computing resources at several centres is the development of ensemble methods. By perturbing the initial conditions and the model physics, a set of different weather scenarios are computed, all assumed to be equally probable and providing additional and valuable information on the uncertainty about the future precipitation amounts. The Ensemble Prediction System (EPS) of ECMWF, whose archives will be used in this study, has been operational since 1992 (Molteni et al., 1996).

To tackle the problem of uncertainty in streamflow forecasting, different probabilistic systems have been developed varying by the underlying theoretical framework as well as by the methods of computation (e.g. Day, 1985; Krzysztofowicz, 1999). The advent of skillful probabilistic precipitation forecasts in the medium-range from ensemble prediction system (Buizza et al., 1999; Mullen and Buizza, 2001; Clark and Hay, 2004; Hamill et al., 2004) has led to the development of ensemble precipitation- based streamflow forecasts. The use of EPS in a large-scale hydrological model has for instance been tested for selected historical flood events (de Roo et al., 2003; Gouweleeuw, 2005; Pappenberger et al., 2005).

Using six years of archived EPS forecast, Roulin and Vannitsem (2005) demonstrated that hydrological ensemble prediction has a positive skill at forecasting the probability of high flow in two test catchments in Belgium. For winter, when large-scale precipitation prevails, this skill was found significant during the whole forecast range of 9 days whereas during summer when convective precipitation dominates, this skill vanishes after 6 to 8 days. The differences in skill between the two catchments were put in relation with the differences in their responses to rainfall events. The changes in spatial resolution of the Numerical Weather Prediction (NWP) in November 2000, from 120 km to 80 km had a positive impact on the forecasts for winter at lead-time of 3 to 6 days.

In this paper, further insight on the quality of the hydrological ensemble predictions will be given. The use of a simple cost-loss decision model will allow for the estimation of the value of flood forecast for decision-making. The skill and value of the ensemble will be compared with those of deterministic approaches. The hydrological prediction system is described in Sect. 2. The methods to analyze the probability

Correspondence to: E. Roulin
(emmanuel.roulin@oma.be)

Published by Copernicus GmbH on behalf of the European Geosciences Union.

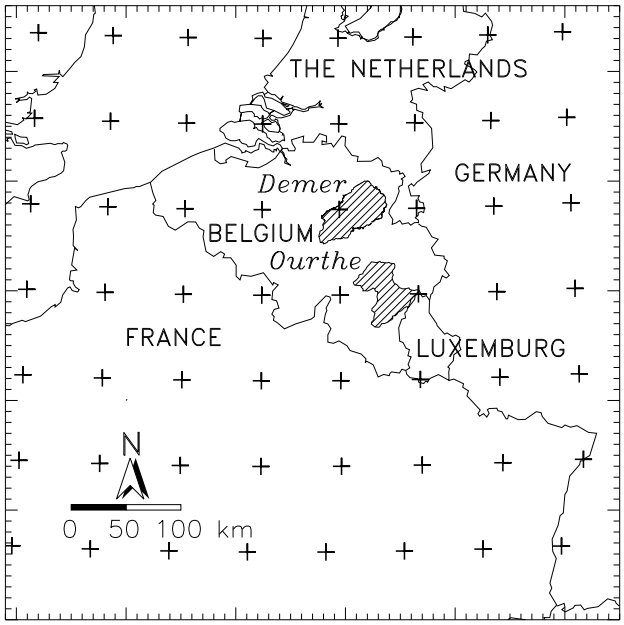

Fig. 1. Location map of the Demer and Ourthe catchments in Belgium; grid points (+) of EPS archives corresponding to the T255 NWP in use from November 2000 to January 2006 at ECMWF.

forecasts are outlined in Sect. 3. The results obtained on two Belgian test catchments using ECMWF archives are presented in Sect. 4. The conclusions are drawn in Sect. 5.

2 Hydrological ensemble prediction system

This study is based on hydrological hindcast using the archived precipitation forecasts from the ECMWF EPS since the date on which the T255 numerical weather prediction (NWP) model was operational (21 November 2000). This NWP model has a horizontal resolution of about 80 km. The EPS has been recently (1 February 2006) upgraded with a T399 NWP with ∼50 km resolution. Precipitation over a grid cell of the hydrological model is given by the value predicted at the nearest grid point of the NWP model. The differences between the scales of the NWP and the hydrologic model (80 km and 7 km, respectively) as well the possible biases in the precipitation forecasts deserve future developments. In the present study, skill and value will be estimated using directly the forecast total precipitation data, i.e. the sum of large scale and convective precipitation. These forecasts start at 12:00 UTC for a period of 240 h and are archived with a 6-h time step. For verification purposes, data from the dense ($\approx$100 km^{-1}) network of daily precipitation observations over Belgium are used; these data are available at 08:00 LT i.e. 07:00 UTC (or 06:00 UTC during the daylight saving period). Therefore, in this study, the forecast day D+1 covers the EPS forecast period between +18 h to +42 h. The last forecast day, D+9, corresponds to the forecast range from +210 h to +234 h. Observed precipitation is interpolated over a grid cell of the hydrological model by weighting measurements by its intersection with the Thiessen polygons of the neighbouring raingauges.

The hydrological model is adapted from the IRMB (Integrated Runoff Model – F. Bultot) water balance model (Bultot and Dupriez, 1976, 1985). The model is implemented with grid cells of 7 km×7 km, and is complemented with a routing procedure based on the width function (Naden, 1992). The conceptual scheme comprises a snow layer, nine land covers with an interception layer and two soil reservoirs for each vegetation cover, two underground reservoirs, and a unit hydrograph for simulating surface water runoff. The IRMB model parameters have been reduced to ten adjustable parameters, and optimized using observed streamflow on a set of medium-sized catchments. The first six parameters were obtained by flow separation and a deconvolution method (Bultot and Dupriez, 1976). The remaining four were optimized using the SCE-UA algorithm (Duan et al., 1992). Empirical non-linear relationships have been found between the values of the parameters and the physical catchment properties using the Nevprop artificial neural network algorithm (Goodman, 1996). These relationships have been used to estimate parameters for the entire hydrological model grid. Note that the high uncertainty introduced in the regionalization procedure has been estimated but not accounted for in the present study. The values of the parameters of the routing function have been obtained by trial and error using observed streamflow for larger basins. Daily precipitation data are the main input information. Other meteorological data (temperature, air humidity, wind speed, solar radiation) are used to account for snow accumulation and melt, and for evapotranspiration using a Penman potential evapotranspiration formulation (Bultot and Dupriez, 1985).

The streamflow prediction experiment is performed in two steps. First, a continuous simulation is made using observed precipitation. This run constitutes the reference run. Every day of the verification period, the values of the variables of the hydrological model are stored and subsequently used as initial conditions for the forecast runs. An ensemble streamflow forecast is then made using the precipitation predicted by each of the 50 members of the ECMWF Ensemble Prediction System as input. For all other weather variables, the input data assume their observed values. Besides this ensemble streamflow, three separate predictions are made: the first using precipitation from the "operational" forecast from ECMWF obtained with a NWP at a higher horizontal resolution (∼50 km); a second, with the "control" run of EPS, i.e. with the same NWP as EPS but without any perturbation of initial conditions; finally, a third, with the mean of the ensemble.

Two catchments are studied (Fig. 1). They belong to the two main river basins in Belgium and possess contrasted hydrological regimes. The Demer at Diest (1775 km^2) in the River Scheldt Basin is located in the loamy region of

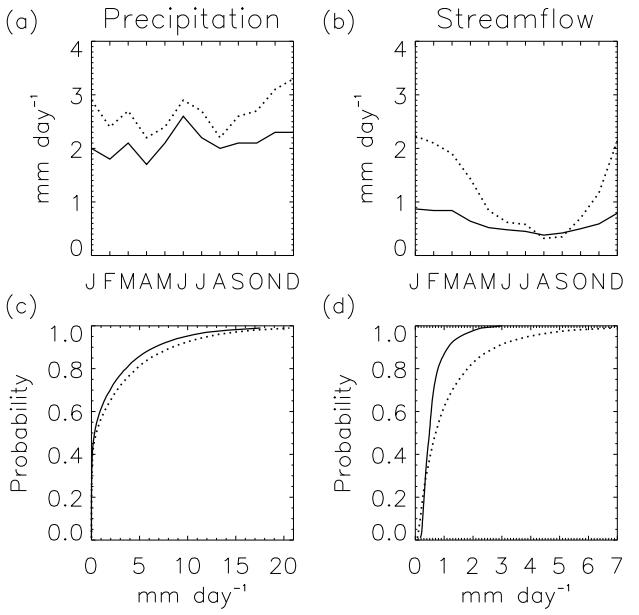

Fig. 2. Regime of observed daily precipitation **(a)** averaged over the Demer catchment (continuous line) and over the Ourthe catchment (dotted line) during 1971–2000, regime of observed streamflow **(b)** of the Demer during 1981–2000 and of the Ourthe during 1971–2000; corresponding probability distribution of precipitation **(c)** and streamflow **(d)**; the maxima are 66 and 57 mm day^{-1} for precipitation of the Demer and Ourthe, respectively and the corresponding maxima for the streamflow are 3 and 20 mm day^{-1}.

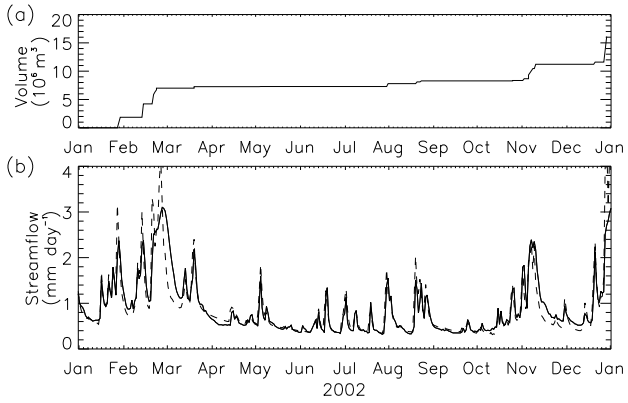

Fig. 3. Management of the Demer during 2002; **(a)** cumulated volume of water (in 10^6 m^3) diverted in the Schulensmeer reservoir (by courtesy of K. Cauwenberghs, AMINAL – afdeling Water); **(b)** daily streamflow of the Demer at Diest; the continuous line corresponds to the measured values (data from HIC) and the dashed line to the "reference" values i.e. simulated using observed precipitation.

Hesbaye and in the sandy region of Campine. This catchment is almost flat and covered with crops and pastures. The Ourthe at Tabreux (1616 km^2) in the River Meuse Basin is located in the Ardennes which consists of a plateau with deep valleys, where a bedrock of shale and sandstone is covered with thin stony loamy soils. The dominant covers are coniferous forests and pastures. Precipitation and streamflow averages based on measurements during 1971–2000 in the two catchments are displayed in Figs. 2a and b. Daily mean discharge values (m^3 s^{-1}) have been converted into daily runoff (mm day^{-1}) to make easier comparison with rainfall and streamflow of catchments of different area. Precipitation occurs evenly throughout the year. Due to the higher elevation of the Ardennes, precipitation over the Ourthe catchment is greater than over the Demer catchment, and the difference is on average slightly larger during winter. The smoother streamflow regime of the Demer is characteristic of a high baseflow contribution due to the high infiltration rate and the presence of aquifers. The larger seasonal differences of the streamflow regime of the Ourthe results from a larger contribution of surface flow. Precipitation and streamflow thresholds that will be used in the verification of the probabilistic forecast are computed from the data observed during 1971–2000 and summarized on Figs. 2c and d.

Note that the water balance model does not take water management actions into account. For the Demer basin, decisions to be taken by flood risk managers include diverting part of the streamflow into flood control reservoirs with a total capacity of about 11.10^6 m^3 (AMINAL – afdeling Water, 2004). On Fig. 3, the cumulated volume of water diverted into the main reservoir during 2002 is compared with the streamflow measured downstream at Diest and with the reference simulation. The effect of filling the reservoir and releasing with some delay the stored water may be noticed for the February and the November events. For this reason, in the following, the streamflow forecasts are compared to the values given by the reference simulation.

The autocorrelation function of simulated streamflow for the two catchments has already been compared to the ones for the observed values for subcatchments and for the Ourthe (Roulin and Vannitsem, 2005). During winter, the decorrelation time i.e. the lag for which the autocorrelation function drops below e^{-1} is around 10 days for sub-catchments during winter and the autocorrelation function is well reproduced by the water balance model. During summer, the decorrelation time varies from 3 to more than 10 days for subcatchments of the Demer. For the Ourthe, the decorrelation time is even longer but it is not properly reproduced beyond 2 days. As a consequence of the daily time step and parameter simplification, runoff generated by precipitation falling over the catchment does not reach the outlet until the next day. For this reason, precipitation observed on day D starts to contribute to the simulated streamflow at D+1 and precipitation forecast for D+1 is routed to the outlet starting at D+2 etc. Therefore, we will begin verification of the streamflow forecasts at D+2.

3 Validation measures

This study will restrict the verification to methods suited for probability forecasts. The methods are mainly from the field of meteorology even if the first validation measure, the Brier Skill Score (Brier, 1950), has been recently applied in hydrology (Bradley et al., 2003, 2004; Roulin and Vannitsem, 2005). The second method, a simple decision model (Richardson, 2000; Zhu et al., 2002) allows the hit and false alarm rates of the ensemble prediction system to be translated into an economic value of the forecast.

3.1 Brier skill score and derived measures

In this study, the probability forecast that the streamflow will exceed a threshold is estimated as the relative frequency of occurrence of this event among the ensemble members. The ensembles are taken at "face value" (e.g. Richardson 2001; Katz and Ehrendorfer 2006) and the uncertainty due to the finite size of the ensembles is ignored. This estimated forecast probability ($0 \leq f \leq 1$) is compared with the observation x that the streamflow has exceeded this threshold (x=1) or not (x=0). In the framework of the distribution oriented (DO) approach to forecast verification, the forecast quality consists of the totality of statistical characteristics embodied in the joint distribution of forecasts and observations $p(f, x)$ (Murphy and Winkler, 1987; Murphy, 1997). The first aspect of the forecast quality is the accuracy measured by the Brier score (Brier, 1950) or mean square error MSE(f, x). The Brier Skill Score is the fractional improvement in accuracy of the forecast over climatology:

$$\mathrm{BSS} = 1 - \frac{\mathrm{MSE}\,(f, x)}{\sigma_x^2} \tag{1}$$

where σ_x^2=MSE (μ_x, x) is the variance of the observations in the verification sample and is a measure of uncertainty; μ_x is the frequency of occurrence of the event during the verification period.

Two different factorizations of the joint distribution into conditional and marginal distributions conveniently relate accuracy to other aspects of forecast quality. The first is the calibration-refinement (CR) factorization:

$$p\,(f, x) = p\,(x|f)\,p\,(f) \tag{2}$$

In the context of the CR factorization, the mean square error can be written as:

$$\mathrm{MSE}\,(f, x) = \sigma_x^2 + E_f\left(\mu_{x|f} - f\right)^2 - E_f\left(\mu_{x|f} - \mu_x\right)^2 \tag{3}$$

where $\mu_{x|f}$ is the conditional mean of the observed event given the forecast. The second term of the right-hand side of (3) is a conditional bias and is a measure of reliability. A forecast is reliable (or well calibrated) if the occurrence frequency of an event, for a given forecast value, is close to its forecast probability of occurrence. The third term is a measure of resolution (or refinement) by which the relative frequency of an event given a forecast differs from the unconditional mean. Substituting (3) into (1), the Brier Skill Score may be expressed as:

$$\mathrm{BSS}=\mathrm{RRES}-\mathrm{RREL} \tag{4}$$

with RRES for relative resolution and RREL for relative reliability. A perfect forecast would have RRES=1 and RREL=0.

The second factorization is the likelihood-base rate (LBR) factorization:

$$p\,(f, x) = p\,(f|x)\,p\,(x) \tag{5}$$

With this factorization, the mean square error can be decomposed as:

$$\mathrm{MSE}\,(f, x) = \sigma_f^2 + E_x\left(\mu_{x|f} - x\right)^2 - E_x\left(\mu_{x|f} - \mu_f\right)^2 \tag{6}$$

where μ_f is the average forecast, $\mu_{f|x}$ its average given the observation and where σ_f^2 is the variance of the forecasts and is a measure of the sharpness or degree to which the probability forecasts are close to 0 or 1. The third term of the right-hand side of (6) is a measure of discrimination or degree to which the distribution of forecasts when the event occurs differs from the distribution when it doesn't. The second term is a measure of the bias conditional to the observation (or type 2 conditional bias). Substituting (6) into (1), the Brier Skill Score may be expressed as:

$$\mathrm{BSS}=1+\mathrm{RDIS}+\mathrm{RSH}-\mathrm{RTY2} \tag{7}$$

where RDIS is the relative discrimination, RSH the relative sharpness and RTY2 the relative conditional bias. Note that both the sharpness and the discrimination are desirables aspects of forecast quality but contribute differently to the MSE.

3.2 Relative economic value

The benefits gained by using streamflow forecasts depends not only on their skill but also on the impact of the streamflow on the system to manage as well as on the actions the user should take to mitigate the consequences. The value of the hydrological prediction system is studied in the framework of a simple static cost-loss ratio decision model whose formulation comes from Richardson (2000).

In this model, the decision has to be made either to take a protective action again an adverse event, let's say a flood event X, or to do nothing. If the event occurs and no action has been taken, then a loss L has to be incurred. Taking action has a cost C whatever the outcome but if the event occurs, a part of the loss, L_1 is prevented.

The strategy to be adopted should minimise the expense on the long term. If the only information available is μ, the frequency of occurrence of the event in the climatological record, the strategy reduces to: either always protect or

never. Always taking a protective action results in an average expense of $C+\mu(L-L_1)$ whilst the average expense of never protecting is μL. Therefore, the optimal strategy consists in always to act if $C+\mu(L-L_1)<\mu L$ that is if $C/L_1<\mu$ and never to act otherwise. The average expense is:

$$E_{\text{climate}} = \min\{C + \mu(L + L_1), \mu L\} \tag{8}$$

If the future were perfectly forecast, it could be possible to take actions only on those occasions when the event is going to occur; the average expense would be:

$$E_{\text{perfect}} = \mu(C + L - L_1) \tag{9}$$

The information given by actual forecasts may allow a decision maker to adapt the strategy in order to reduce the average expense. The reduction of expense is a measure of the value of the forecasts. The relative economic value compares the reduction of expense with the reduction which would be achieved by a perfect forecast as the ratio:

$$V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}} \tag{10}$$

The decision maker will benefit from the forecast if $V>0$.

A deterministic forecast of a dichotomous event like "the discharge of the Demer will exceed 68 $\text{m}^3\,\text{s}^{-1}$" simply consists in announcing that either it will or will not occur. The four possible combinations of deterministic forecasts and event outcomes are depicted on Fig. 4. The event is correctly rejected at a frequency f_1, it is missed at f_2, incorrectly forecast at f_3 and correctly at f_4. Noting that $\mu=f_2+f_4$, the hit rate H and the false-alarm rate F are defined as:

$$H = f_4/\mu \tag{11}$$

$$F = f_3/(1-\mu) \tag{12}$$

The average expense of the deterministic forecast system is given with:

$$E_{\text{forecast}} = f_2 L + f_3 C + f_4(C + L - L_1) \tag{13}$$

With the definition of the value in Eq. (10), of the expenses given in Eqs. (8), (9) and (13), and of the hit and false-alarm rate in Eqs. (11) and (12), the relative economic value of the deterministic forecast can be written as:

$$V = \frac{\min(\alpha,\mu) - F\alpha(1-\mu) + H\mu(1-\alpha) - \mu}{\min(\alpha,\mu) - \mu\alpha} \tag{14}$$

where $\alpha=C/L_1$, the cost-loss ratio, or cost of taking an action expressed as a fraction of that part of the potential loss which is protected by that action. The relative value depends on the parameter α which characterises the decision making situation, μ which describes the hydrologic (climatic) context and H and F which quantify the performance of the forecast system.

As far as probabilistic forecast is concerned, a decision maker has to choose the probability threshold p_t at which he will take an action. By this choice, the probabilistic forecast system is changed into a deterministic one. For any value of p_t, the relative economic value of the probabilistic system can be calculated with Eq. (14). For an optimal strategy, the decision maker will select the probability threshold which results in the largest value.

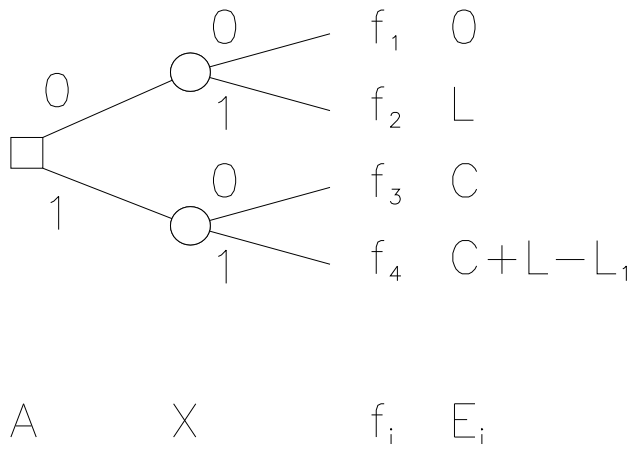

Fig. 4. Decision tree for a simple static decision problem: a decision A is taken on an action at the cost C on the base of forecast; this action is aimed at preventing an amount L_1 of the loss L that otherwise should be incurred if the event X happens; the expenses E_i are given for each combination i of forecasts and outcome occurring with frequency f_i.

3.3 Value of early warnings

How could a manager take advantage of a positive relative value at medium-range if hydrological ensemble forecast has a greater value at shorter range? It was suggested by Richardson (2000) that "users with more time to prepare for protective action may be able to reduce the cost of taking it and to reduce α towards the region where EPS has the greater advantage". Nevertheless, the situation may be such that the forecast changes or, if the forecast of high risk persists the following days, actions with gradually more implications may be undertaken. In other words, the situation is more likely to be dynamic rather than static.

An extension to the decision problem described in the previous section is proposed to investigate the relationships between the probability threshold chosen for the different forecast days, the costs of successive actions and the value of the forecasts. This problem is represented on Fig. 5 as a decision tree. Two stages are considered: a forecast is issued, let's say, six days before an eventual flood event; preliminary action can be decided accordingly at this first stage; three days later, a second forecast is issued, based on which complementary action can be undertaken; if no action was decided at the first stage, the same protection can be obtained with a late action at a higher cost. Depending upon the actions

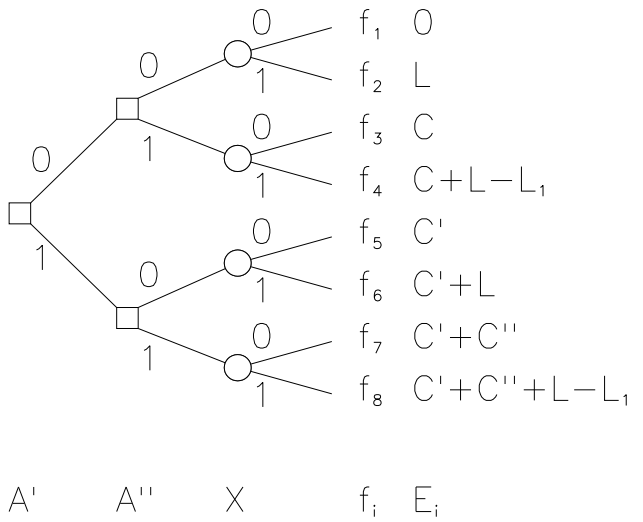

Fig. 5. Decision tree for a simple two-stage dynamic decision problem: a decision A' is taken on a preliminary action at cost C' on the base of a first forecast; later, depending on a second forecast, a decision A'' is taken on either a complementary action at cost C'' or a late action at cost C in case no action had been decided at the first stage; these actions at total cost C'+C'' or at cost C are aimed at preventing an amount L_1 of the loss L that otherwise should be incurred if the event X happens; the expenses E_i are given for each combination i of successive forecasts and outcome occurring with frequency f_i.

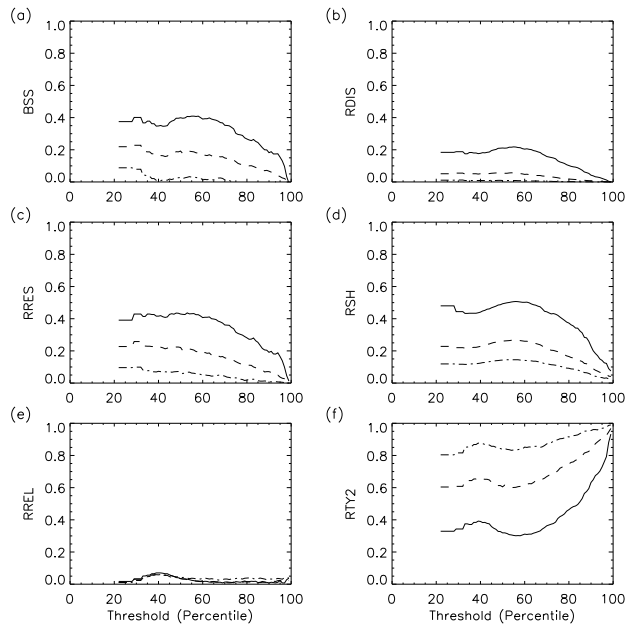

Fig. 6. Brier Skill Score **(a)** of the EPS precipitation for the Demer at D+3 (solid line), D+6 (dashed line) and D+9 (dash-dotted line); CR decomposition: relative measures of resolution **(c)** and reliability or conditional bias **(e)**; LBR decomposition: relative measures of discrimination **(b)**, sharpness **(d)** and type 2 conditional bias **(f)**.

taken at the two stages and upon the occurrence of the flood event, different total costs are incurred (Fig. 5). Note that this dynamic problem is different from other models (e.g. Wilks, 1997) consisting in a sequence of decision and event where losses can be sustained at each step and weather forecasts used for decision-making have a fixed range.

As in the static problem, if the flood occurs and no preventive action has been taken, the total loss is given by L and from that loss, an amount L_1 may be protected by the preventive actions. If only one action at the second stage is taken, its cost C is the same as in the static problem to which the dynamic problem will be compared. If taken sequentially, the actions have costs C' and C'' at the first and second stage, respectively, such that the advantage of timeliness is a lower total cost $C'+C''<C$. The average expense of this decision system is:

$$E_{\text{forecast}} = \sum_{i=1}^{8} f_i \times E_i \tag{15}$$

where f_i is the frequency of occurrence of the i-th combination of successive forecasts and outcome and E_i is its associated cost and loss given in Fig. 5. The frequencies of occurrence depend upon the choice of the probability thresholds at the two stages. The pair of thresholds resulting in the minimum average expense is to be selected.

4 Results

4.1 Skill of rainfall and runoff predictions

In the Figs. 6 to 8, measures of the accuracy and other aspects of probability forecasts quality are summarized for the precipitation over the Demer and for the streamflow of the Demer and of the Ourthe. Unlike in Roulin and Vannitsem (2005) these measures have been calculated over the whole year whatever the season and are presented over the entire range of thresholds defined by the percentiles (corresponding to Fig. 2).

For precipitation (Fig. 6), the skill is decreasing for increasing precipitation amounts and for increasing lead times. Nine days in advance, warnings based solely on the probability of heavy precipitation are of little use. The most interesting property is that forecasts of the probability of precipitation are very reliable. Note also that percentiles below a value of about 20 are not represented and correspond to no precipitation.

The probability that the streamflow exceeds a threshold is forecast with a significant skill over a wide range of thresholds corresponding to percentiles from 5–10 to about 95 (Figs. 7 and 8). This skill results from the high resolution of the forecasts (RRES), combined with their good reliability as seen with the small values of the relative bias conditional to the forecast (RREL). For high flows, the skill of the

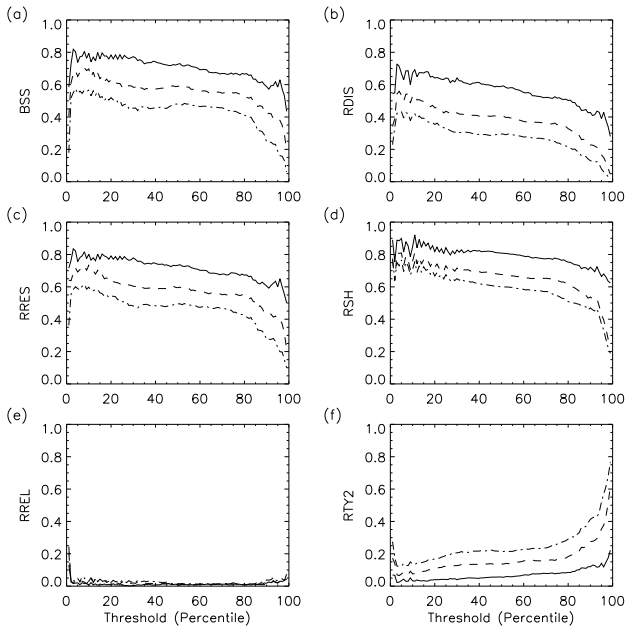

Fig. 7. Brier Skill Score **(a)** of the hydrological ensemble predictions for the Demer at D+3 (solid line), D+6 (dashed line) and D+9 (dash-dotted line); CR decomposition: relative measures of resolution **(c)** and reliability or conditional bias **(e)**; LBR decomposition: relative measures of discrimination **(b)**, sharpness **(d)** and type 2 conditional bias **(f)**.

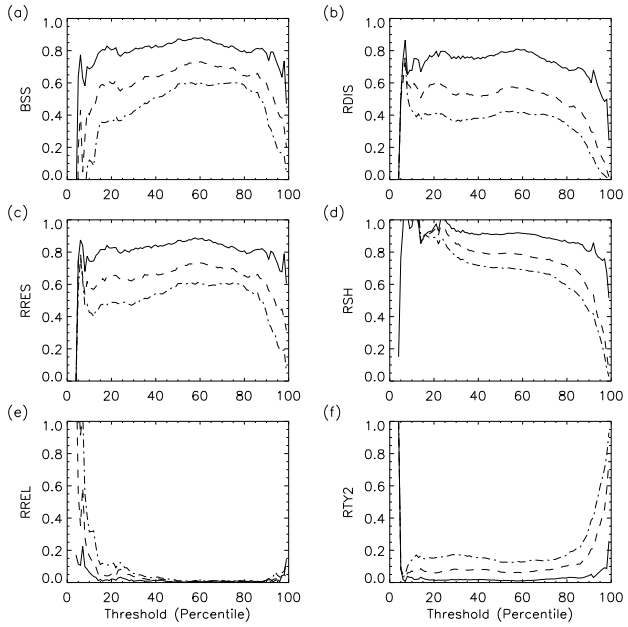

Fig. 8. Brier Skill Score **(a)** of the hydrological ensemble predictions for the Ourthe at D+3 (solid line), D+6 (dashed line) and D+9 (dash-dotted line); CR decomposition: relative measures of resolution **(c)** and reliability or conditional bias (e); LBR decomposition: relative measures of discrimination **(b)**, sharpness **(d)** and type 2 conditional bias **(f)**.

Ourthe is greater than the skill of the Demer. This difference was linked to the faster response of the Demer to the precipitation events (Roulin and Vannitsem, 2005), as illustrated in Figs. 7a and 8a for percentiles above 40. Below 30, the reverse is true. The Demer is characterized by a large contribution of the aquifers to the baseflow; therefore, the streamflow of the Demer is more predictable in low flow situations if proper values are given to the initial conditions of the hydrological model.

With the LBR decomposition, it can be seen that the skill of the two catchments behave differently. For the Demer, with increasing value of the threshold, both the sharpness (RSH) and the discrimination (RDIS) are continuously decreasing whereas the bias conditional to the observations increases. For the Ourthe, the bias conditional to the observations is almost constant up to around a threshold corresponding to the percentile 80 depending on the forecast day. The sharpness is high in the same range and even very high for low threshold values indicating that the probability forecasts are not far from either 0 or 1. This good sharpness does not correspond to a good discrimination. The discrimination is nearly constant in this range and the skill is therefore increasing slightly up to around the percentile 50.

The different aspects of the quality of the hydrological ensemble prediction system are compared with the corresponding aspects of the three deterministic alternatives. On Fig. 9, the Brier Skill Score and measures derived from the CR and LBR factorization are plotted with the forecast days for a threshold corresponding to the 95th percentile. The hydrological ensemble prediction system outperforms any of the three deterministic forecast systems. Its skill remains significantly positive up to the forecast day D+9 whereas the skill of the deterministic forecasts vanishes at D+7 for forecasts using ECMWF "operational" and the EPS control and D+8 for forecasts using the EPS ensemble average.

The CR decomposition shows that this difference in skills between deterministic and ensemble forecasts comes mainly from the better resolution of the latter. The forecasts of the hydrological ensemble prediction system range between 0 and 1 and are binned in 0.1 intervals in order to compute the resolution. The deterministic forecasts are set to either 0 or 1. According to Eq. (3), a better resolution means that the frequency of high flows given the probability forecasts of the ensemble differs from the unconditional mean to a higher degree than the frequency given the deterministic forecasts. The bias conditional to the forecasts using the EPS control or the ECMWF operational precipitation is higher than the corresponding bias for the EPS ensemble average and for the hydrological ensemble prediction. It can be noted that, for a fairer comparison, the results for a deterministic forecast could be improved by adding small random noise (Dr. Kees Kok, personal communication; see also end of Sect. 4.2).

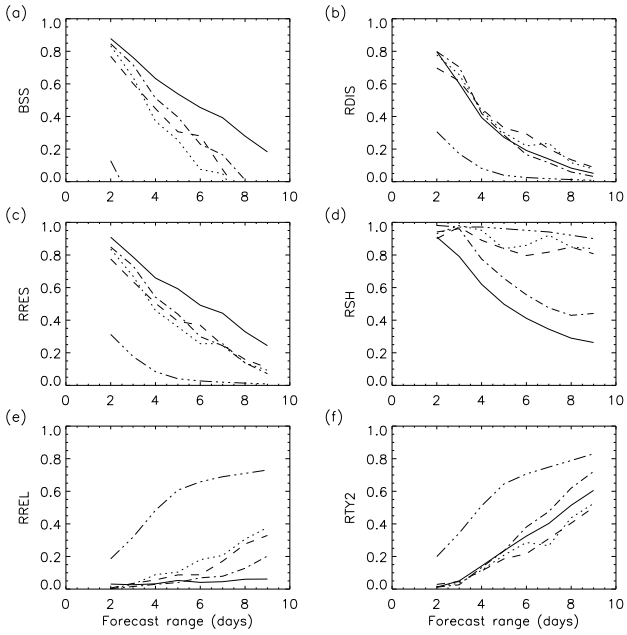

Fig. 9. Brier Skill Score **(a)** of the probability forecast that the streamflow of the Ourthe will exceed a threshold corresponding to the 95th Percentile; calibration-refinement decomposition: relative resolution **(c)** and relative reliability **(e)**; likelihood-base rate decomposition: relative discrimination **(b)**, relative sharpness **(d)** and relative type 2 conditional bias **(f)**; the solid lines correspond to the hydrological ensemble, the dashed lines correspond to the prediction using ECMWF deterministic forecast, the dotted lines, using the EPS control, the dash-dotted lines, using the EPS ensemble average, and the dash-triple dotted lines correspond to the streamflow persistence (verification period 2000–2004).

The LBR decomposition shows that the sharpness of the forecasts using ECMWF operational and EPS control are close to 1 as can be expected for deterministic forecasts. For probability forecasts at this high streamflow threshold, the sharpness measure is not exactly 1 because the unconditional means of the forecasts and of the observations are not equal ($\mu_f \neq \mu_x$). Nevertheless, this good sharpness is not associated to a good discrimination. The discrimination of the four systems is almost the same and decreases with forecast lead time. Therefore, despite a slightly lower bias conditional to the observations compared to the hydrological ensemble prediction system, the deterministic system using ECMWF operational and EPS control have less skill. The superiority of the hydrological ensemble prediction system over the deterministic system using the EPS ensemble average has another origin: the sharpness of this deterministic system is only moderately larger and the bias conditional to the observation is slightly larger than for the hydrological ensemble prediction system; combined with a discrimination of the same order, the skill of the hydrological ensemble prediction remains significantly better.

The skill has been conveniently measured using the sample uncertainty as a reference even if known a posteriori and assumed for all times of the year. There is a range of alternative forecasts that would be skillful in this situation: persistence of the streamflow, persistence of the weather used as forcing to the hydrological model, ensemble streamflow predictions using weather series drawn from past climate at the same time of the year. The latter has skill but less than ensemble streamflow predictions based on ECMWF-EPS (Roulin and Vannitsem, 2005): the skill of ensemble streamflow predictions using EPS compared to this alternative probabilistic forecast was above 0.5 after three days during winter and remained significantly positive after nine days. In the present study, the Brier Skill Score and its CR and LBR decompositions for the persistence forecasts are included in Fig. 9: persistence has a relative resolution decreasing steadily and, as soon as after two days, the skill of persistence is worst than climatology due also to increasing biases.

In ensemble forecasting, the choice of perturbed initial conditions is a crucial step to produce an ensemble with a spread representing properly the uncertainty about the future state of the atmosphere. In a perfect ensemble, the observations should be taken as an additional independent realization of the same underlying probability distribution (Talagrand et al., 1997; Hamill and Colucci, 1997). The histogram can be calculated with the positions of observed values among the ensemble. This histogram allows one to check the statistical consistency of the ensemble and may be used to calibrate the ensemble (Hamill and Colucci, 1997). In this study, the priority has been given to preserve the consistency in time of each precipitation scenario of the ensemble. The calibration method has been tentatively applied to the output ensemble streamflow but not reported here in detail. Only a marginal improvement has been obtained by this post-processing. Indeed, the calibration changes the reliability of the probability forecasts and not their resolution; since the reliability is already very good – at least for both cases studied – there is little room for improvement. Future improvements are rather to be expected through an increased resolution of the probability forecasts and such an increase might only follow future improvements of the NWP model.

4.2 Value of streamflow predictions

The relative economic value of the hydrological ensemble prediction system is now examined for the Ourthe for streamflow exceeding a threshold corresponding to the 95th percentile. This value is shown on Fig. 10 for the forecast day D+6 as a function of the cost-loss ratio. The value is shown for different probability thresholds p_t amongst which the threshold resulting in the largest economic value has to be selected. It can be seen on this graph that the hydrological ensemble prediction system has a positive value for most of the cost-loss situations. The optimal probability threshold depends on the cost-loss ratio (Fig. 14). The decision maker

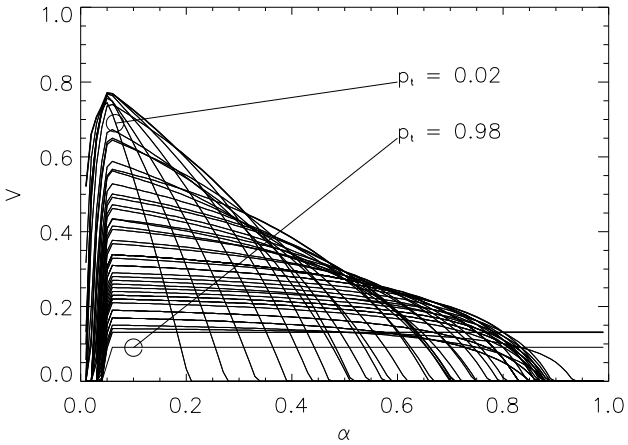

Fig. 10. Relative value V of the hydrological ensemble prediction for different cost-loss ratios α: probability that the streamflow of the Ourthe will exceed a threshold corresponding to the 95th Percentile on D+6; the different curves correspond to different probability thresholds p_t; the overall value is shown with the envelope (thick line).

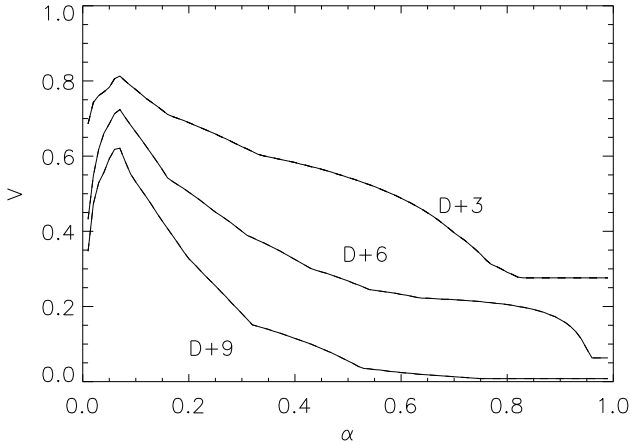

Fig. 11. Relative value V of the hydrological ensemble prediction for different cost-loss ratios α: probability that the streamflow of the Demer will exceed a threshold corresponding to the 95th percentile for the forecast days D+3, D+6 and D+9; the probability forecasts are compared with 1 or 0 if the "reference" streamflow exceeds the threshold or not.

faced with large potential loss compared to the cost of protection (small α) will benefit by taking actions even when the forecast probability is low. On the contrary, in a situation where costly action is to be decided (high α), the user would benefit from the forecast if she/he decides to act only when the probability forecast is high. The value of the forecast can substantially drop with an inappropriate choice of the probability threshold. The overall value of the hydrological ensemble prediction system is shown by the envelope of the different curves on Fig. 10. To benefit from this value, the decision-maker must have a good perception or knowledge of his situation and select the appropriate probability threshold.

Note that the highest value of the probability threshold (p_t=0.98) is rarely reached or exceeded and all these occasions correspond to streamflow higher than the 95th percentile. The false-alarm rate is therefore zero and it follows from Eq. (14) that the relative economic value is equal to the hit rate. This is due to sampling limitation. Note also that the resolution of the probability thresholds is linked to the size of the ensemble (Richardson, 2000) and, for the 51 members EPS, the best resolution is achieved with p_t intervals of about 0.02 as plotted on Fig. 10.

The overall relative economic value is compared for the forecast day D+3 to D+9 on Fig. 11 for the Demer. This value is positive in most of the cost-loss situations. As the forecast horizon increases, the value for situations where taking an action is costly decreases more rapidly than for situations characterized by a cost-loss ratio close to the frequency of occurrence of the damaging event. For $\alpha \approx \mu_x$, where the relative economic value is optimum, this value remains very high even at D+9. The results presented in this study may be considered as an upper limit to the skill and relative economic value of the hydrological ensemble predictions because the streamflow forecast using EPS precipitation is compared to the streamflow simulated using the observed precipitation. For the Demer, due to the flood control actions, the streamflow forecast has not been compared to the observed streamflow.

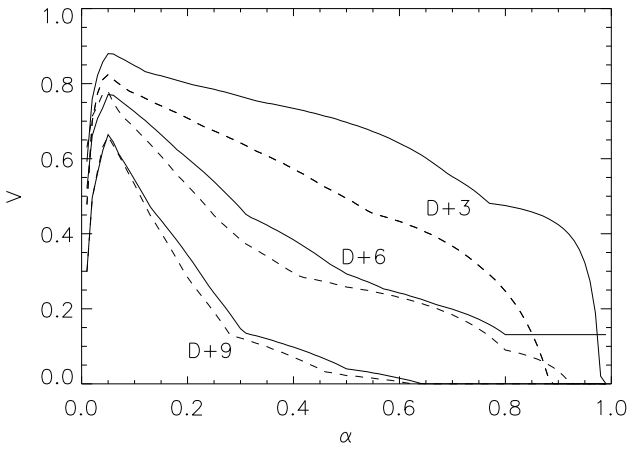

Fig. 12. Relative value V of the hydrological ensemble prediction for different cost-loss ratios α: probability that the streamflow of the Ourthe will exceed a threshold corresponding to the 95th Percentile for forecast days D+3, D+6 and D+9; continuous line: the probability forecasts are compared with 1 or 0 if the "reference" streamflow exceeds the threshold or not; dashed line: the comparison is made with the observed streamflow.

The overall relative economic value for the Ourthe is shown on Fig. 12. At D+3, this value is larger for the Ourthe

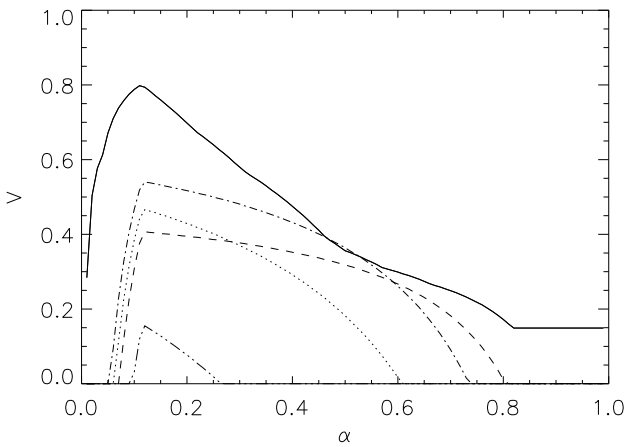

Fig. 13. Relative value V of the hydrological ensemble prediction (solid line) for different cost-loss ratios α: probability that the streamflow of the Ourthe will exceed a threshold corresponding to the 95th Percentile on D+6; the dashed line is the relative value of a deterministic prediction obtained by using the ECMWF deterministic forecasts; the dotted line corresponds to the prediction with the control run of EPS; the dash-dotted line to the prediction with the EPS ensemble average, and the dash-triple dotted line to the streamflow persistence (verification period 2000–2004).

than for the Demer as was shown for the skill. The relative economic value has also been estimated on the basis of the observed values of the streamflow. Due to errors in both input data and hydrological modelling, the simulated streamflow at the start of the forecast differs from the observed streamflow. At D+3, the comparison of the probability forecast with the observed streamflow results in a relative economic value which is much lower than the values based on the reference streamflow. Nevertheless, after several days, uncertainties about the precipitation forecasts themselves dominate and the differences between the verification measures made with the reference and the observed streamflow are reduced. A further development could complement this hydrological ensemble forecast system with an updating module (e.g. Goswami et al. 2005).

The overall relative economic value of the hydrological ensemble prediction is compared with the value of the deterministic alternatives on Fig. 13, for the forecast day D+6. This analysis did not account for forecasts made when a river is exceeding a flood threshold and this makes the economic value inflated. This problem is linked with the persistence of the streamflow already mentioned. The value of persistence is also included in Fig. 13 where it is shown marginal at D+6. At a threshold corresponding to the 95th Percentile, flood events last on average three days. The superiority of the ensemble forecasts over the deterministic forecasts lies on the flexibility allowed by a proper choice of the probability threshold. The advantage of using hydrological ensemble predictions for decision-making is the highest for the situation with low cost-loss ratio, where the optimum probability threshold is also low and whose use is made possible if the ensemble size is sufficient. For values of α lower or equal to μ_x, the advantage extends for ensemble predictions made at the medium-range over deterministic forecasts issued within a shorter delay (e.g. comparing D+9 on Fig. 12 with D+6 of any deterministic forecast on Fig. 13). Note that selecting a probability threshold which results in the highest value is equivalent to some calibration. Zhu et al. (2002) use independent verification data to calibrate the probability forecasts and then evaluate the economic value with the use of a probability threshold equal to the cost-loss ratio. Regarding the deterministic forecasts, their skill and value can be improved if error statistics are used to estimate a probability distribution (Atger, 1999; Smith et al., 2001).

4.3 Value of early warnings

Numerical examples of the dynamic decision model are given with ficticious costs and losses but based on the probability forecast six days and three days in advance that the streamflow of the Demer exceeds the 95th percentile. In the following, the day when a terminal consequence is analyzed will be referred as T and the days when the successive forecasts are issued will be referred as T-6 and T-3. In these examples, the total cost of the two actions taken successively is set to $C'+C''=C-0.05\,L_1$. The model is applied for a range of cost-loss ratio ($\alpha=C/L_1$) from 0.10 to 0.95. Two cases are tested: in the first case, the cost of the earlier action is taken fixed (C'=0.01 L_1) for the whole range of α, the cost-loss ratio of the corresponding static problem, whereas in the second case, the costs of both actions are proportional ($4\,C'=C''$).

The values of the pair of thresholds selected for the different cost-loss ratios are compared with the values of the single threshold of the static problem on Fig. 14. The thresholds for the probability forecast at T-3 for the dynamic and static model follow almost the same relationship with α. The more expensive the action compared to the protected loss, the higher the probability threshold to optimise the value of the forecast system. For a perfectly reliable forecast in the static problem, the optimal threshold would be equal to α (Richardson, 2000). The threshold for the probability forecast at T-6 follows a different relationship than the threshold at T-3 and depends upon the rate of the cost of the action at the first stage, C', with α (excepted at α=0.15 where both cases have the same C'). For a cost constant with α, the threshold increases slowly with α and is always much lower than the threshold at T-3. For a cost C' proportional to the total cost, the threshold increases steeply with α, and, in the range of α from 0.2 to 0.8, it is greater than the threshold at T-3, even if the cost C' at T-6 in this case is still much lower than the cost C'' at T-3.

The relative economic values of the hydrological ensemble predictions for the Demer in the framework of the static prob-

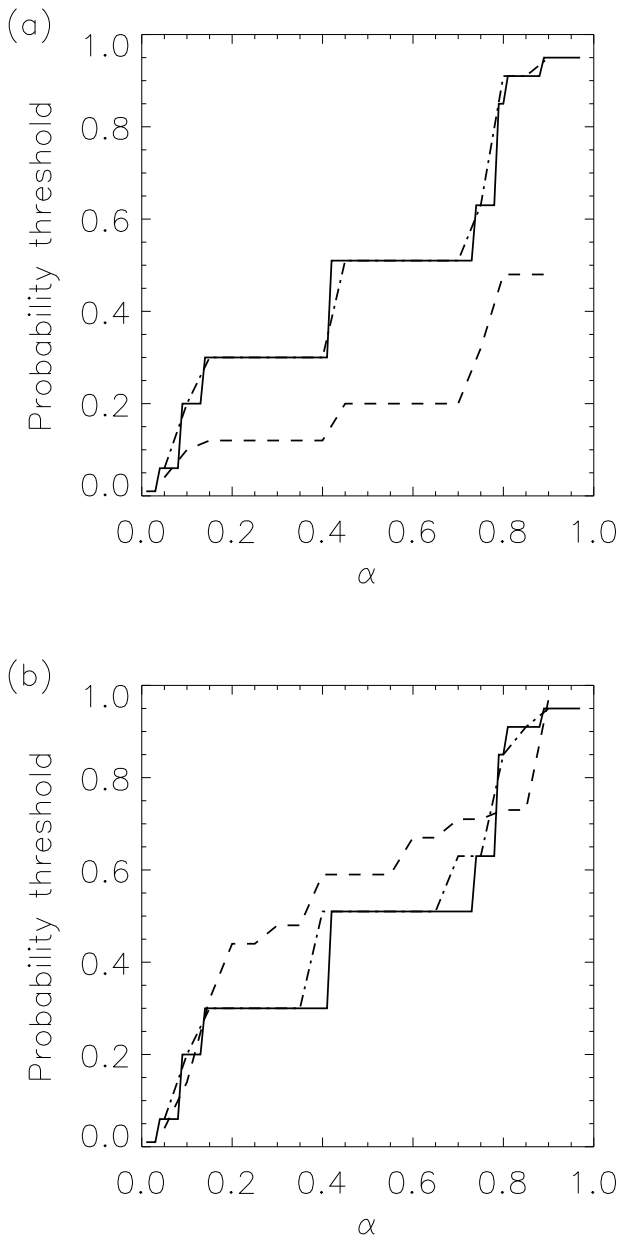

Fig. 14. Probability thresholds resulting in optimal relative economic values for different cost-loss ratio α; the continuous line corresponds to the static decision problem (α=C/L_1); the other lines correspond to the two-stage dynamic decision problem (α=$(C'+C'')/L_1$): the dashed and the dash-dotted lines are for the thresholds used at the first and second stage, respectively; two cost sharing figures are considered: **(a)** C'/L_1=0.01 and **(b)** C'=C''/4; both dynamic problems have a total cost reduced compared to the static problem (C'+C''=$C-0.05\,L_1$); example of the Demer with the 95th percentile as streamflow threshold.

lem and of the dynamic problem are presented on Fig. 15. In this figure the relative economic value is obtained with

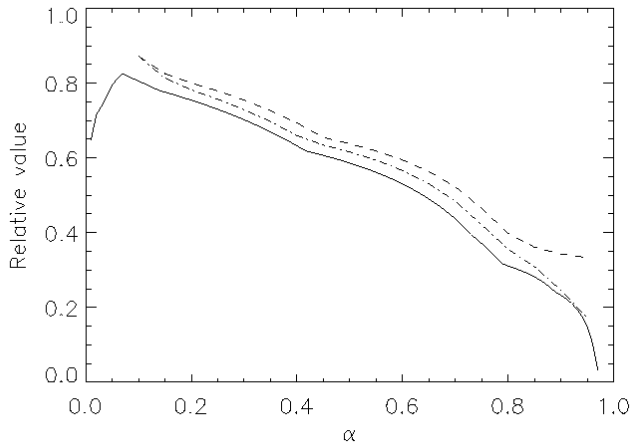

Fig. 15. Relative value of the hydrological ensemble prediction for different cost-loss ratio α; the continuous line corresponds to the static problem (T-3); the dashed line corresponds to the dynamic problem with a cost at the first stage (T-6) set to C'/L_1=0.01; the mixed line corresponds to the dynamic problem with costs at the two stages proportional: C'=C''/4; same example as Fig. 14.

Eq. (10) in which the forecast system is compared to two other decision systems – the first with only climatology as information and the second with perfect forecasts. For ease of comparison, in these two decision systems, the cost of action is set to C, for both the static and the dynamic problems. For a same reduction of the total cost, a cost of an earlier action fixed to a low value gives the highest advantage.

The preliminary results of a dynamic problem in the context of medium-range hydrological predictions presents numerical results based on hindcasts for the Demer catchment. The event chosen is a streamflow greater than a threshold defined by the 95th percentile. In the hindcast period, the frequency of occurrence of this event is 0.07. The results are presented for a range of cost-loss ratio from 0.10 to 0.95. Further study will focus on low cost/loss ratios, where the relative economic values are likely to be the highest and will have to take uncertainty into account. A formal analysis of this dynamic problem will also broaden the scope of possible conclusions towards a better use of medium-range forecasts. Finally, it is worth investigating the plausible values to be assigned to the parameters of the economic model in order to meet the operational context.

In this dynamic model, the second stage has been defined at T-3. The value of the early warning could also be compared with the situation at T for which decision are taken, in most operational river management centers, on the base of the values of precipitation and streamflow measured in real time. Likewise, the streamflow threshold has been selected so that a sufficient number of events are included in the verification period and that statistically meaningful conclusions may be drawn. Clearly, the choice of the streamflow threshold should be addressed and included in the study of

the interrelationships between probability threshold, forecast lead-time and value.

The dynamic aspects associated with the medium-range forecasts are worth being addressed with regard, for instance, to the belief that persistent probabilistic forecasts have more value. A simpler two-stage dynamic problem to test this hypothesis would consist in taking a single action provided not only that a probability threshold is exceeded but also that a probability threshold, possibly different, has been exceeded in a previous forecast. In this situation and with the variables defined in Fig. 5, the expenses could be summarized with $E=\{0, L, 0, L, 0, L, C, C+L-L_1\}$. Tests not reported here using this model for successive probabilistic forecasts at T-6 and T-3 but also for T-4 and T-3, haven't led to additional value. An even simpler static problem related to this view consists in pooling successive forecasts into lagged super-ensembles. Pooling ensembles forecasts issued at ECMWF at 00:00 UTC and 12:00 UTC did not improve the skill of hydrological ensemble predictions.

5 Conclusions

A hydrological ensemble prediction system has been developed based on the use of ECMWF EPS in a hydrological model. This system has been tested on two Belgian catchments with contrasted hydrological cycle, by performing extensive hindcasts with archives from ECMWF for the period from November 2000 to January 2006, which corresponds to the operational NWP model with ∼80 km spatial resolution. This study focuses on the skill and the relative economic value of the probability forecasts.

Ensemble forecasting requires the adaptation of the management rules towards the use of probability forecasts instead of deterministic discharge forecasts. This implies also an increased number of simulations with hydrological/hydraulic modelling systems. For comparison, alternative deterministic hindcasts have been also performed on the basis of either the archives of the ECMWF operational deterministic runs at higher spatial resolution or on the basis of the mean of EPS ensembles or on the control (not perturbed) run of EPS.

The value of early warning and the issue raised by the choice of probability thresholds has been tackled with a simple dynamic two-stage decision model. Numerical examples have been given with ficticious cost and loss amount, but with forecast probability and event occurrence taken from the hindcasts for the Demer. It has been shown that the optimum value is obtained with a probability threshold for the decision at the first stage depending upon both the cost of the preliminary action and the total cost to prevent damage; the probability threshold for the decision at the second stage has been found similar to the corresponding threshold in the static problem.

The hydrological ensemble predictions have greater skills than deterministic ones. Their greater relative economic value for many users lies on the choice of a probability threshold appropriate to the decision-making situation. Optimal thresholds can only be selected on the base of hindcasts. As the forecasting systems are regularly upgraded, there is a need to reiterate studies based on hindcasts.

The adaptation of the timestep from daily timestep to 6 h at which EPS precipitation forecasts are available, the addition of a streamflow updating model, and the inclusion of flood control rules are part of future developments. The benefit would be expected in the first forecast days beyond which uncertainties on precipitation predictions dominate. Hindcasts will be used in more realistic management situations in order to define the probability thresholds corresponding to the optimal use of forecasts for decision-making.

Acknowledgements. I thank G. Demarée for his continuous encouragements and S. Vannitsem for discussions and comments. I am grateful to E. Todini, K. Kok and an anonymous reviewer whose useful remarks and suggestions improved the manuscript. This research contributes to the European Union COST 731 concerted action on "Propagation of Uncertainty in Advanced Meteo-Hydrological Forecast Systems".

Edited by: E. Todini

References

AMINAL – afdeling Water, De Overstromingsvoorspeller van de Demer (in Dutch), edited by: Thomas, P., Brussels, 44 pp., 2004.

Atger, F.: The skill of ensemble prediction systems, Mon. Wea. Rev., 127, 1941–1953, 1999.

Bradley, A. A., Hashino, T., and Schwartz, S. S.: Distributions-oriented verification of probability forecasts for small data samples, Wea. Forecast., 18, 903–917, 2003.

Bradley, A. A., Schwartz, S. S., and Hashino, T.: Distribution-oriented verification of ensemble streamflow predictions, J. Hydrometeorol., 5, 532–545, 2004.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Wea. Rev., 78, 1–3, 1950.

Buizza, R., Hollingsworth, A., Lalaurette, F., and Ghelli, A.: Probabilistic predictions of precipitation using ECMWF Ensemble Prediction System, Wea. Forecast., 14, 168–189, 1999.

Bultot, F. and Dupriez, G. L.: Conceptual hydrological model for an average-sized catchment area, J. Hydrol., 29, 251–292, 1976.

Bultot, F. and Dupriez, G. L.: Daily effective evapotranspiration from a river basin, in: Casebook on Operational Assessment of Areal Evapotranspiration, Oper. Hydrol. Rep., 22, (WMO-No 635), 80–105, 1985.

Clark, M. P. and Hay, L. E.: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, J. Hydrometeorol., 5, 15–32, 2004.

Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Resour. Plan. Manage., 111, 157–170, 1985.

de Roo, A., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongiannini-Cerlini, P., Todini, E., Bates, P. D., Horritt, M., Hunter, N., Beven, K., Pappenberger, F., Heise, E., Rivin, G., Hils, M., Hollingsworth, A., Holst, B., Kwadijk, J., Reggiani, P., Van Dijk, M., Sattler, K., and Sprokkereef, E.: Development of a

European flood forecasting system, Int. J. River Basin Manage., 1, 49–59, 2003.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015–1031, 1992.

Goodman, Ph.: NevProp software, version 3. Reno, NV, University of Nevada, http://brain.cs.unr.edu/publications/NevProp.zip, 1996.

Goswami, M., O'Connor, K. M., Bhattarai, K. P., and Shamseldin, A. Y.: Assessing the performance of eight real-time updating models and procedures for the Brosna River, Hydrol. Earth Syst. Sci., 9, 394–411, 2005, http://www.hydrol-earth-syst-sci.net/9/394/2005/.

Gouweleeuw, B. T., Thielen J., Franchello, G., De Roo, A. P. J., and Buizza, R.: Flood forecasting using medium-range probabilistic weather prediction, Hydrol. Earth Syst. Sci., 9, 365–380, 2005, http://www.hydrol-earth-syst-sci.net/9/365/2005/.

Hamill, T. M. and Colucci, S. J.: Verification of the Eta-RSM short-range ensemble forecasts, Mon. Wea. Rev., 125, 1312–1327, 1997.

Hamill, T., Whitaker, J. S., and Wei, X.: Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts, Mon. Wea. Rev., 132, 1434–1447, 2004.

Katz, R. W. and Ehrendorfer, M.: Bayesian approach to decision making using ensemble weather forecasts, Wea. Forecast., 21, 220–231, 2006.

Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, Water Resour. Res., 35, 2739–2750, 1999.

Molteni, F., Buizza, R., Palmer, T. N., and Petrogliatis, T.: The ECMWF Ensemble Prediction System: methodology and validation, Q. J. R. Meteorol. Soc., 122, 73–119, 1996.

Mullen, S. L. and Buizza, R.: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System, Mon. Wea. Rev. 129, 638–663, 2001.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Mon. Wea. Rev., 115, 1330–1338, 1987.

Murphy, A. H.: Forecast verification, in: Economic value of weather and climate forecasts, edited by: Katz, R. W. and Murphy, A. H., Cambridge University Press, London, 19–74, 1997.

Naden, P. S.: Spatial variability in flood estimation for large catchments: the exploitation of channel network structure, Hydrol. Sci. J., 37, 53–71, 1992.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J., and de Roo, A. P. J.: 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), Hydrol. Earth Syst. Sci., 9, 381–393, 2005, http://www.hydrol-earth-syst-sci.net/9/381/2005/.

Richardson, D. S.: Skill and relative economic value of ECMWF ensemble prediction system, Q. J. R. Meteorol. Soc., 126, 649–667, 2000.

Richardson, D. S.: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, Q. J. R. Meteorol. Soc., 127, 2473–2489, 2001.

Roulin, E. and Vannitsem, S.: Skill of medium-range hydrological ensemble predictions, J. Hydrometeorol., 6, 729–744, 2005.

Smith, L. A., Roulston, M. S., and von Hardenberg, J.: End to end ensemble forecasting: Towards evaluating the economic value of ensemble prediction system, ECMWF Tech. Memo. No. 336, 29 pp., 2001.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, In Proceedings, Workshop on predictability, 20–22 October 1997, ECMWF, Reading, UK, 1–25, 1997.

Wilks, D. S.: Forecast value: prescriptive decision studies, in: Economic value of weather and climate forecasts, edited by: Katz, R. W. and Murphy, A. H., Cambridge University Press, London, 109–145, 1997.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The economic value of ensemble-based weather forecasts, Bull. Amer. Meteor. Soc., 83, 73–83, 2002.