

EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts

R. Buizza⁽¹⁾, H. Asensio⁽²⁾, G. Balint⁽³⁾, J. Bartholmes⁽⁵⁾, J. Bliefernicht⁽⁴⁾, K. Bogner⁽⁵⁾, F. Chavaux⁽⁷⁾, A. de Roo⁽⁵⁾, J. Donnadille⁽⁶⁾, V. Ducrocq⁽⁷⁾, C. Edlund⁽⁸⁾, V. Kotroni⁽¹⁶⁾, P. Krahe⁽⁹⁾, M. Kunz⁽¹²⁾, K. Lacire⁽¹⁰⁾, M. Lelay⁽¹³⁾, C. Marsigli⁽¹¹⁾, M. Milelli⁽¹⁵⁾, A. Montani⁽¹¹⁾, F. Pappenberger⁽¹⁾, D. Rabuffetti⁽¹⁵⁾, M.-H. Ramos⁽⁵⁾, B. Ritter⁽²⁾, J. W. Schipper⁽¹²⁾, P. Steiner⁽¹⁴⁾, J. Thielen-Del Pozzo⁽⁵⁾ & B. Vincendon⁽⁷⁾

(1) European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

(2) Deutscher Wetterdienst (DWD), Offenbach, Germany

(3) VITUKI, Budapest, Hungary

(4) University of Stuttgart, Germany

(5) Joint Research Center (JRC), Ispra, Italy

(6) NOVELTIS, Ramonville Saint-Agne, France

(7) Météo-France, Toulouse, France

(8) Swedish Meteorological and Hydrological Institute (SMHI), Sweden

(9) Federal Institute of Hydrology (BFG), Koblenz, Germany

(10) Infoterra, Toulouse, France

(11) Servizio Idro-Meteorologico Emilia Romagna (ARPA-SIM), Bologna, Italy

(12) Institute for Meteorology and Climate Research (IMK), University of Karlsruhe, Germany

(13) Laboratoire d'études des Transferts en Hydrology et Environment (THE), Grenoble, France

(14) Meteoswiss, Zurich, Switzerland

(15) ARPA-Piemonte, Torino, Italy

(16) National Observatory of Athens (NOA), Athens, Greece

Key words: verification, hydrological forecasts, meteorological forecasts, ensemble prediction

Corresponding author address: Dr R. Buizza, ECMWF, Shinfield Park, Reading, RG2-9AX, UK (email: Buizza@ecmwf.int).

PREVIEW is a research and development project co-funded by the European Commission (contract 516172 of the 6° Framework Programme). PREVIEW proposes to develop, at the European scale, new or enhanced information services for risk management in support of European Civil Protection Units and local or regional authorities, making the best use of the most advanced research and technology outcomes in Earth Observation. To reach that goal a European team encompassing major actors in risk management has been set up (the EURORISK Consortium), gathering the required technical skills from the scientific community, the operators and the industry with the expertise of main end-users bodies (Civil Protections). For further information, please visit www.preview-risk.com.

Abstract

One component of the PREVIEW project is the analyses of issues linked to the validation of the overall forecast value of probabilistic meteorological and hydrological forecasts. This report is a contribution to this discussion, with attention focused to the case of probabilistic meteorological and hydrological predictions.

Since a forecast is valuable only if it has a high technical and functional quality, the forecast value can be assessed only if both the technical and the functional qualities are assessed. Two frameworks are introduced to structure the problem of the assessment of the technical and functional quality:

- The ‘Technical Quality Framework’, based on the assessment of four (technical) attributes: ‘forecast characteristics’, ‘validation characteristics’, ‘metric’ and ‘significance test types’
- The ‘Functional Quality Framework’, based on the assessment of four (functional) attributes: ‘availability and means of distribution’, ‘content and format’, support, maintenance and training’ and ‘communication of product’s technical quality’

Although the quantification of a forecast functional quality is not always possible, when this is the case the forecast value can be quantified by computing a ‘generalized’ product of the functional quality and technical quality scores.

These frameworks are applied to six verification problems, to illustrate how they can be used to structure in a more appropriate way the problem of the validation of a forecasting system. Furthermore, some synthetic (idealized) examples where functional quality can be quantified are discussed, and the methodology that can be applied to assess the forecast value is presented.

1. ‘Technical quality’, ‘functional quality’ and ‘forecast value’

Meteorological predictions are often expressed in the form of deterministic or probabilistic forecasts of variables that can take one of a finite set of possible values (*Buizza 2001*). A typical example is the prediction of more than 10 mm of precipitation or of temperature below freezing. Considering an event X, a deterministic forecast is usually expressed as a statement like ‘the event X will (or will not) occur’, while a probabilistic forecast can be expressed as ‘there is a 30% probability of occurrence of event X’. Hydrological forecasts, such as the ones issued by the European Flood Alert System, have started using similar formats, with forecasts of probabilities of reaching a finite set of warning levels been issued by national or international agencies (e.g. flood warning, flood watch, all clear in the case of the Environment Agency in Britain).

The *technical quality* of a deterministic or probabilistic forecast is a measure of the accuracy of the forecast statement, whereby the accuracy is measured using one or a range of ‘metrics’ that quantify the similarity between the forecast and the observed value (examples of metrics are bias, mean error, root-mean-square error, ranked probability score and skill score, Brier score and skill score, area under a relative operating characteristic curve, rank histograms, Nash-Sutcliffe coefficient and explained variance).

Unfortunately, technical quality is not a guarantee of forecast value: for example, a perfect forecast that is communicated to the user too late (paradoxically after the event has already occurred) has zero forecast value although it has perfect technical. The *functional quality* of a forecast is a measure that depends on issues such as communication delays, or information content. For example, a less accurate forecast that is communicated to the user earlier enough to

allow the user to take protection actions to reduce the potential losses, is technically less accurate but functionally more valuable. The distinction between technical and functional quality is not academic, but reflects the real-time operational use of meteorological and hydrological forecasts: *forecast value* depends from both the technical and functional quality of a forecast.

Increasingly, both meteorological and hydrological forecasts have been used by decision-makers to judge whether or not to take an action to protect against a possible loss. Typically, a decision-maker would have the possibility to spend an amount C to protect against a loss L (with $L > C$), and would use a deterministic or a probabilistic forecast to decide whether to spend C to protect against the potential loss L . The potential economic value of a deterministic or a probabilistic forecast can be assessed using skill measures defined by coupling contingency tables and cost-loss decision models (*Katz et al 1982, Murphy 1985, Wilks & Hamill 1995, Richardson 2000*). This cost/loss decision models will be used in this work to illustrate the difference between technical and functional quality, and to quantify the functional quality of synthetic (idealized) forecasts.

In the first part of this report, two frameworks that can be used to assess the technical and functional qualities of a forecast are introduced. These frameworks are then applied to verify the quality of meteorological and hydrological forecasts (there could be cases when evaluation is possible although verification per se may not be, see e.g. *Beven 2006, Konikow & Bredehoeft 1992, Oreskes et al. 1994*), in particular of forecasts issued in the context of the PREVIEW project. In the second part of this report, the possibility to define a unique, objective measure of forecast value that depends on both the technical and functional qualities is discussed, and few 'synthetic' (idealized) examples are presented. Finally, the key issues discussed in this report are summarized, and a list of recommendations is presented.

2. Four-attribute frameworks to assess technical and functional quality

The problem of the assessment of the *technical quality* of a forecast can be addressed using the four-attribute framework illustrated in Fig. 1:

- **Forecast characteristics:** this attribute includes a description of the forecast field under assessment: the variable type, its forecast length, its resolution, covered area, origin;
- **Verification characteristics:** this attribute includes a description of the verification data used to assess the forecast: the variable type, its resolution, covered area, origin;
- **Metric:** this attribute describes the measure(s) used to compare each forecast with its verification, whether the verification refers to one case or an average of many cases, whether it refers to the skill of a single or of a probabilistic forecast
- **Significance test types:** this attribute includes a description of the statistical tests used to assess the robustness of the verification results.

Table 1 lists some of the components of the four attributes.

Functional quality depends on attributes that are not usually considered when measuring technical quality (e.g. availability and timeliness of forecast products, content and presentation of forecast information and credibility), and which may be difficult to assess in an objective way. For example, it may be difficult to give an objective measure to the content of a forecast bulletin, or to its presentation. Figure 2 illustrates a framework based on four attributes that can be used to assess the functional quality of a forecast:

- ***Availability and means of distribution:*** ‘availability’ depends on how frequently forecasts are available to the end user, e.g., 24 hours per day or 365 days per year. It is important to recall that the availability of a product is very important during a flood situation: if the product is a part of a production chain with different (technical) parts involved, its availability can be affected by the availability of previous parts in the chain. ‘Means of distribution’ depends on the delivery mode (e.g. internet or ftp), and should consider, for example, whether the product reaches the decision makers. For example, a forecast may be available at the production web site, but due to poor internet connection the product is not transmitted to the user.
- ***Content and format:*** these attributes depend on whether a product can be decrypted (i.e. clearly understood by the end user) and lead to a better decision-making process (i.e. it contains valuable information). Product’s format can vary widely from case to case (e.g. forecasts can be delivered as bulletins, maps, graphs or tables) and should be accompanied by all the information/documentation necessary to make it clearly understood by the end-users. ‘Content’ measures also whether the forecast product contains the necessary and sufficient information for the decision-making process (for example, in the framework of flood forecasting, precipitation forecasts should be supplemented with temperature forecasts to be able to consider snow-melt contributions).
- ***Support, maintenance and training:*** it’s important that a policy for support and maintenance of the products, as well as regularly training of the end-user are organized. This activity will ensure continuity, and improve the capacity of the end users to exploit the products, while its absence may lead to misunderstandings and communication break downs, possibly precisely during high-alert cases.
- ***Information about products’ technical quality:*** it is extremely important that a product’s technical quality is communicated to the users, so that users can take informed decisions

(there is considerable research into ways to communicate probabilistic and deterministic forecasts, see e.g. *Janssen et al. 2004; van der Sluijs et al. 1998*, and references in *Pappenberger & Beven 2006* and in *Pappenberger et al. 2006*).

The four functional-quality attributes could be assessed using a questionnaire (Table 2), where users are asked to give a numerical score (e.g. ranging from 0 for a useless forecast to 1 for a perfectly functional one), so that a final functional score could be obtained.

3. Technical and functional quality of meteorological and hydrological forecasts: six real-time examples

Hereafter, the two general frameworks introduced in sections 2 are applied to some verification problems of meteorological and hydrological forecasts.

3.1. ‘Technical quality’ of ECMWF probabilistic precipitation prediction

On 12 September 2006, ECMWF upgraded its operational Ensemble Prediction System to the new Variable Resolution EPS (VAREPS, *Buizza et al 2007*): following the implementation, the resolution of the first 10 forecast days was increased from T_L255L40 to T_L399L62, and the forecast length has been extended to 15 days, but with the last 5 forecasts days run with a T_L255L62 resolution. As part of its pre-operational testing, an earlier version of VAREPS with a truncation from T_L399 to T_L255 applied at forecast day 7 instead of 10, was used to generate precipitation forecasts for a set of cases covering the PREVIEW special period (20 July to 31 August 2002). As part of PREVIEW, the VAREPS forecasts were compared with forecasts issued by the T_L255 ensemble system, to assess whether the system upgrade lead to more accurate forecasts of meteorological variables such as precipitation, of interest for hydrological

probabilistic predictions. This section discusses few results obtained during the comparison, applying the verification framework of technical quality.

Table 4 summarizes some of the key attributes of the four components of the technical quality framework used in this assessment: in particular, note that the variable of interest is the 24-hour accumulated total precipitation, defined over a regular latitude/longitude grid with a 2.5 degree resolution. Since the number of cases and of grid points over Europe is rather limited, technical quality measured over Europe was compared to technical quality over the whole Northern Hemisphere. Note also that a proxy defined as the 0-24h forecast given by the ECMWF operational forecast run in summer 2002 (T_L511L60 resolution) was used as verification field instead of observed values. A set of metrics were used to assess different aspects of technical quality of both deterministic and probabilistic forecasts, without any significance test.

Figure 3 shows one of the results of this investigation, the average Brier skill score computed over Europe and Northern Hemisphere for four probabilistic prediction forecasts. The top panel of Fig. 3 indicates that for during the PREVIEW special period, on average EPS probabilistic precipitation forecasts of up to 10 mm/d are skilful up to about forecast day 8 when verified against a proxy for verification defined by a 24-hour T_L511L60 forecast.

3.2. ‘Technical quality’ of COSMO-LEPS probabilistic precipitation prediction

The COSMO-LEPS system (*Marsigli et al 2001, Molteni et al 2001*) has been used in the PREVIEW project to dynamically downscale the global VAREPS forecasts provided by ECMWF. One of the rationales for using COSMO-LEPS instead of simpler, purely statistical downscaling methods is that the COSMO model is better capable to simulate small-scale

phenomena, and can thus provide some extra, valuable information on the top of the global VAREPS system.

This section discusses results obtained by COSMO-LEPS system (*Montani et al. 2003*) for the period September-November 2004, when the population of the mesoscale system was set to 10 members, i.e. COSMO-LEPS was run in a configuration similar to that used for PREVIEW special period (Table 5). For this period, COSMO-LEPS forecasts are compared to ECMWF EPS forecasts to assess whether they can add valuable information to the one produced by the ECMWF EPS. Before comparing the two ensemble systems, it is worth reminding the reader that the two ensemble systems differ both in membership (10 for COSMO-LEPS and 51 for the EPS) and resolution (10 km for COSMO-LEPS and 80 km for the EPS). To assess the impact of ensemble size, COSMO-LEPS is compared to the full-size EPS, and to 10-member EPS consisting of the 10 Representative Members used to define the COSMO-LEPS initial and boundary conditions. To alleviate the fact that COSMO-LEPS has a higher resolution, both systems are verified on the same 1.5 degree grid: for each 1.5 degree box, grid point forecasts are aggregated and averaged. The aggregation of the forecast values is performed considering different features of the forecast probability distribution within the box. Similarly, observations within a box are treated, as the forecast values, as aggregated values (*Marsigli et al 2005*). The comparison is performed over a geographical region that includes Germany, Switzerland and Northern Italy, considering 24-h precipitation (accumulated from 06:00 to 06:00 UTC), verified against observations from a very dense network of rain-gauges (about 5000 observations per day).

Figure 4 shows the Brier skill score for different precipitation products given by the three forecasting systems, COSMO-LEPS (10 members), full-size EPS (51 members) and small-size EPS (10 members). Results confirm that forecast accuracy strongly depends on the type of

measure which used to assess it. Considering average precipitation, EPS performs better than the mesoscale system (top-left panel), indicating higher skill in predicting the total amount of precipitation deployed over a large area. On the other hand, if attention is focused on the prediction of maximum values, COSMO-LEPS BSSs are higher (bottom-right panel). This is probably due to the better capability of the mesoscale system to forecast precipitation peaks accounting for minor localisation errors. Results based on the prediction of the 50th percentile (Fig. 4, top-right panel) are similar to those obtained for average precipitation, while results based on the prediction of the 90th percentile (i.e. the of the precipitation distribution) shows that both COSMO-LEPS outperforms the other two systems, again possibly due to its higher horizontal resolution.

3.3. ‘Technical quality’ of COSMO-LME deterministic precipitation prediction

Classical skill scores, which are often based on a direct evaluation of the discrepancies between observations and the nearest model grid point results, are prone to misinterpretation when applied to high resolution NWP models. Effects like the so-called “double penalty” have a detrimental impact on the perceived skill of the forecast model but have little or no relevance for the quality of the forecast with regard to applications like medium range flood forecasting. Since flooding of larger rivers depends more on area averaged precipitation than on extreme events at isolated locations, a catchment based quantitative verification seems more appropriate as means to judge the technical quality of the NWP model.

As simple examples of catchment-based, continuous verification metrics, Figure 5 shows the daily mean precipitation of the COSMO-LME forecast and the associated mean error for the period 10 August 2002 to 16 August 2002 for the Elbe catchment area. The precipitation from a

12 UTC run was accumulated for 24 hours for the forecast range 18h to 42h and averaged over the Elbe catchment area. The corresponding synoptic observations at 6 UTC have been interpolated to the model grid and subsequently averaged over the same area. Another reference data set for verification was obtained from radar data precipitation estimates which were adjusted by the available rain gauges. Both observational datasets are illustrated in Fig. 5 together with the model results. Note that the associated scores like the mean error may vary depending on the choice of the sources of the observations (see bottom panel of Fig. 5). Even though the COSMO-LME model somewhat underestimated the observed extreme values of precipitation, the forecasts captured the overall evolution of the heavy rainfall event for this region quite well.

3.4. ‘Technical quality’ of high-resolution, deterministic forecasts of flash-floods

Within the framework of PREVIEW, the value of the next generation of high resolution NWP models is assessed for flash-flood forecast. Four non-hydrostatic kilometric-scale models (COSMO-ALMO2 for Meteoswiss, COSMO-LAMI for Arpa Piemonte, Meso-NH/AROME for Météo-France and MM5/WRF for NOAA) are compared against rain-gauge observations for flash-flood cases over the French Cevennes-Vivarais and the Italian Piemonte watersheds (i.e. two Mediterranean regions prone to heavy precipitation and flash-flooding), applying the verification framework of technical quality. For each flash-flood event, 18-h range forecast have been issued twice a day, i.e. at 00UTC and 12 UTC respectively.

The main attributes of the four components of the technical quality framework are listed in Table 7a for quantitative precipitation forecast of the September 2005 case in which two flash-floods occurred over the Cévennes-Vivarais region. In this example, the variable of interest is the 18-h accumulated precipitation (shorter accumulation period can also be considered), interpolated at the observation points shown in Fig. 6. The high-resolution of the forecast allow a direct

comparison of the model and raingauge values. Figure 6 shows that for that case, none of the models is better than the others for the whole period considering all verification metrics. For the “significant precipitation event” (defined as 20mm/18h), MM5 performs better according to the ETS on the first flash-flood, whereas COSMO-ALMO2 is superior for the second flash-flood of the September 2005 case. It is worth to note that some models seem to perform better according to one metric and not when using another metrics.

High-resolution models produce mesoscale structures, more intense cores of precipitation and gradients than larger scale models than those described in the previous sections. In that case, small space and timing errors lead to poorer scores than the smoother forecast of a low-resolution model. This is known as the “double penalty” problem (*see, e.g., Bougeault 2003*). Design of new approaches to assess the technical quality of high-resolution forecasts taken into account the double penalty problem is a current research topic (some of them are going to be assessed within PREVIEW , *see e.g. Theis et al 2005, Yates et al 2006*).

Another way to assess the value of high resolution precipitation forecasts is through the hydrological response to these rainfall forecasts. The hourly precipitation forecasts from the four models are supplied as input to two hydrological models within PREVIEW. The metrics used for the verification against observation and comparison between models are the Nash coefficient for the 1-3 December flash-flood event over the South-eastern France (Table 7b). For that event the MESO-NH forecast gives a better hydrological response for all the watersheds and sub-basins considered (Fig. 7). It is worth pointing out that the value of the forecast increases with the size of the river catchment.

3.5. ‘Technical quality’ of the European Flood Alert System (EFAS)

The evaluation of deterministic and probabilistic forecasted hydrological variables is usually done by adopting typical and well-established meteorological verification tools like the Brier skill score or the statistical scores derived from contingency tables of observed and forecasted occurrences (e.g, *Georgakakos et al* 2004). Whenever a continuous variable, like the river discharge or the water level, is converted to a discrete (or binary yes/no) one by applying some threshold filters, all these meteorological tools can be easily applied (*Thielen et al* 2006a). However a lot of information is lost by the application of thresholds and problems may occur when the hydrological forecasts have to be evaluated not only for the binary exceedances of thresholds (usually corresponding to flooding periods) but to the entire time series of forecasted values. The development of new validation methods for matching these additional hydrological demands is a current research topic in EFAS (for more information on EFAS¹ see *de Roo et al* 2000, 2002, and 2003) and some preliminary results of ongoing analyses are presented hereafter.

Basically, EFAS forecasts are used as a *pre-alert* to allow the receiving authorities to be aware of the possibility of a flood to take place before they are able to catch the event with their local forecasting system. In other words, with EFAS forecasts the authorities can already play through a number of different scenarii “what to do if” and, as the event approaches and location and magnitude become more certain, the authorities can act more quickly and accurately, increasing the economic value of the forecasts. Although the cost of the precautionary actions may be almost the same with or without the pre-alerts, the economic losses will decrease having a gain in time

¹ For details on EFAS, its set-up, research activities and dissemination products, the reader can refer to: *de Roo et al* (2000, 2002, and 2003), *Thielen* (2004), *Gouweleeuw et al* (2005), *Thielen et al.* (2006b), *Ramos et al.* (2006), and the EFAS website <http://efas.jrc.it/>.

for reacting. Therefore the evaluation of this gain in preparedness (see the definitions below) is one essential part of forecast validation and usefulness assessment in EFAS. It represents a preliminary attempt to bridge the gap between assessing the technical and functional qualities of hydrological forecasts. The main idea is to search for tools to apply in the verification of the technical quality of a forecast which take into account the most important aspects of usefulness related to the forecasts and/or the forecasting system.

A first detailed investigation of EFAS forecasts was conducted based on the classical analysis of contingency tables (see Table 3), in terms of misses and false alarms, where:

- An **observed event** is defined as YES (*or NO*) when discharges simulated with observed meteorological data as input exceed (*or do not exceed*) EFAS high flood alert levels.
- A **forecasted event** is defined as YES (*or NO*) for:
 - Deterministic ECMWF (named EUD) and/or DWD: when forecasted discharges exceed (*or do not exceed*) EFAS high alert levels in simulations based on weather forecasts issued at 12:00
 - Probabilistic ECMWF-EPS (named EUE): when forecasted discharges exceed (*or do not exceed*) EFAS high alert levels in simulations based on weather forecasts issued at 12:00 and with at least N_{th} EPS simulations above EFAS high levels (N_{th} ranges from 1 to 50). A contingency table is thus computed for each N_{th} considered in the analysis.

Since EFAS is a medium-range forecast system, it is possible to add a criterion of persistence in consecutive forecast to define a forecasted event. In this case, a forecasted event is defined as above, BUT with high alert levels forecasted by two consecutive EFAS forecasts.

Table 8 summarizes the key entries of the general framework used to assess the technical quality of EFAS, and Fig. 8 illustrates, for a single case, the methodology described above. [Note that, in this example, hits, misses, false alarms and correct rejection for $N_{th} = 5$ (i.e., at least 5 EPS-based simulations show discharges above EFAS high alert level) when a criterion of persistence is considered in the analysis; note also that the first and last 24 hours for which forecasts apply are not considered in order to avoid initial conditions effects and to allow considering persistence.] Figure 9 shows the hit and false alarm rates for EFAS deterministic ECMWF forecasts for 2 different upstream areas. The occurrences are calculated considering persistence and as a function of upstream area and lead time. The whole of Europe is considered in a 5x5-km pixel grid and computations were done for the period from June 2005 to May 2006. By computing hits, misses and false alarms, a number of scores can be derived. Figure 10 shows the result of applying some classical meteorological metrics to the EFAS. The scores are calculated without considering persistence and taking all lead times together in the analysis. Computations are done for the month of August 2005 and for 70 locations spread over the Danube river basin.

3.6. ‘Functional quality’ of EFAS medium-range ensemble flood forecasts

In order to assess EFAS forecasts in terms of its “functional quality”, the four attributes of the functional quality framework introduced in section 2 have been assessed in the following way:

- *Availability and means of distribution* - This has been assessed via the establishment of a network and a platform for open exchanges, with a formal agreement (Memorandum of Understanding, MoU) between the JRC, the data providers and the national authorities. Since the beginning of EFAS project, the creation of a network has been designed to establish a close contact with end users and an official platform for exchanges on the needs, use and communication of ensemble forecasts. The formal agreement is not only a

legal mean of exchanging data, but it also helps to establish shared responsibilities and to clarify the role of each part in the process of setting up a forecasting system.

- *Content and format* - On a yearly basis, a one-day technical meeting is programmed and supported by the JRC to gather EFAS partners and discuss on future improvements of the forecasting system. During these meetings, national authorities are encouraged to report on EFAS success/failure in forecasting events in their countries and on how the information is locally being used. Since EFAS is a research project, any constructive remarks are included, when possible, in the development of the final operational system. The archive and analysis of “feedback questionnaires” requested to partners whenever a flood event is forecasted by EFAS and external alert reports are sent out.
- *Communication of products’ technical quality* - Generally, if EFAS simulates potential flooding more than 48 hours in advance and for an upstream area larger than 30,000 km² in an area covered by a MoU, an EFAS information report is sent to the partner organization. At the end of each flood event for which EFAS reports were issued, a feedback questionnaire is sent out to the authorities to inquire about different aspects of the provided information, e.g. questions if the stated information was correct, the information was presented clearly, the reports were received on time or the information was used in some way. Free-format emails and other feedbacks, such as personal communications, are also collected and individually analyzed. Furthermore, workshops and seminars are held to inform users’ of the past performance of EFAS’ forecasts, and to understand the user’s needs.
- *Support and training* - The production of training documents and regular informative bulletins to clarify and stress on the aims, the possible benefits and adverse effects of EFAS forecasts, as well as the way the system works and presents its results. It is considered that the functional value of a forecast can be improved if the users have a

good knowledge and understanding of the aims and set-up of the forecasting system, as well as of the visualisation tools used for presenting the forecasts and their derived products. EFAS bulletins are issued typically bi-monthly, summarizing the most important EFAS events during a two- to three-month period and reporting on system development issues. Training documents are prepared whenever the need for specific topics is expressed (e.g., how to correctly read EFAS reports, how to interpret EFAS maps and diagrams of forecasted alert levels, understanding the use of EPS in flood forecasts, combining spatial and temporal information at a point, exploring ensemble medium-range flood forecasts for better decisions, etc.).

In order to illustrate how this procedure is implemented in practice, the next sub-section presents a summary of three main steps mentioned above:

- The analysis of feedback questionnaires to EFAS information reports for external flood alerts from July 2005 to June 2006. The analysis is based on the evaluation of the feedback questionnaires received from users and on other communications
- The key conclusions of a workshop held at the JRC on November 2005 and organized to explore together with Member States' forecasters how to deal with uncertainty in operational flood forecasting and decision making
- The discussions developed from the experience gathered on the use of EFAS forecasts concerning the benefits and adverse effects of medium-range forecasts

3.6.1. Feedback to EFAS information reports from partner organizations

During 2005, EFAS reported 10 external flood events and a total of 50 reports were sent out to authorities in Germany (Danube), Austria (Danube), Hungary (Danube), Slovakia (Danube),

Bulgaria (Danube and Evros/Maritsa) and Italy (Po). From March to June 2006, EFAS reported to 10 different organizations and sent out a total of 95 EFAS Information Reports to Germany (Elbe and Rhine), Slovakia (Danube), the Czech Republic (Elbe and Danube), Hungary (Danube and Tisza), Bulgaria (Danube) and Moldova (Prut and Dneestr). After the last EFAS information report, the authorities received a feedback questionnaire to fill in. Although these feedback questionnaires allowed to assess also in a quantitative way the potential impact of EFAS information for the receiving authorities, some authorities preferred to communicate per email or personal communications (this happens, e.g., in the wake of a flood crisis when the authorities have little time for additional work) their 'subjective' assessment. Tables 9-12 summarize the key findings of the analysis of 7 feedback questionnaires received by EFAS on the usefulness, quality and dissemination of EFAS reports and forecasts. Overall, feedback from partner organizations on the EFAS information reports has generally been a very positive and encouraging experience. Partner organizations are generally supporting the project and very keen to gain more experience with EFAS information reports. The analysis has pointed out some strengths and weakness of the project. By processing the answers, it was also noted that some concepts are differently interpreted by the partners. There is thus a need of improving also the feedback questionnaire itself, searching for clearer definitions of technical terms and statements able to conduct to a more homogeneous interpretation of the issues addressed.

3.6.2. Outcome of the 1st EFAS Workshop on the use of EPS in flood forecasting

A workshop (*Thielen et al. 2005*) was organized on the use of EPS in flood forecasting to address two main concerns of EFAS regarding flood forecasting based on EPS: i) how to extract meaningful information from the meteorological EPS for medium-range flood forecasting, and ii) how to communicate the uncertainty in flood forecasting to end-users. The specific objectives of the workshop were to explore together with flood forecasting experts from the Member States the

usefulness of EPS information implemented in EFAS for operational flood forecasting and decision making, and the perception of uncertainty in flood forecasting.

Overall, the workshop's experience was very successful. The participants expressed their interest in the subject and most of them found that the workshop brought their knowledge about ensemble prediction flood forecasting forward. The discussion of four real case studies showed clearly that the use of EPS in flood forecasting has a great potential. Once introduced to the concept of probabilistic flood forecasting and being used to working with ensemble stream flows, the participants missed not having the EPS information during the case studies if they were not provided (i.e., in the case of the control group). The workshop revealed interesting patterns in the use of EPS, e.g. that they were considered positive when confirming the deterministic forecasts whereas they were considered rather disturbing when being contradictory. From the discussions held during the workshop, some important aspects emerged concerning the practice in flood forecasting and the use of EPS. For instance, it was noted that flood forecasters have a tendency to maintain the highest alert issued until they are sure that there is no risk of achieving the alert level anymore. It was also observed that when forecasters issued a high alert level in the first day of forecast, they prefer to keep it through the next days, even if the risk decreased. They would only decrease the alert level they issued if in the third day the situation showed to be no longer severe. Persistency of flood forecasts was taken into account from one day to the other and assisted forecasters in their decisions. Forecasters also highlighted the difficulties of performing flood forecasting over a region where they are not used to work with. Forecasters' local expert knowledge of the river basin and of the prior meteorological and hydrological situations was perceived as a key element in good flood forecasting. The participants felt that training on specific case studies for their own river basins is necessary to properly understand the value of EPS. Providing training material or daily access to EFAS results was considered an important

aspect for the successful use of EFAS forecasts. The importance of visualizing results in a useful and concise way was stressed. It was generally confirmed that the understanding of using EPS increased with subsequent case studies. This also illustrates a “training effect” that arises when using EPS on a daily basis and highlights the importance of providing training on forecasting products to end users.

3.6.3. On the benefits/adverse effects of medium-range forecasts

Generally speaking, within the EFAS users’ community, the local flood forecasting centers do not activate an emergency procedure only based on the early flood alerts forecasted by the medium-range forecasting system. For this specific action, the national authorities take a decision based mainly on their local information, which is more accurate than the medium-range information due to the use of higher resolution and locally calibrated models, as well as specific expert knowledge. Medium-range forecasts are mainly used as a *pre-alert* to allow the receiving authorities to be aware of the possibility of a flood to take place. In other words, with medium-range forecasts in hands, local forecasters assess a number of different scenarii (“what to do if”) and, as the event approaches and its location and magnitude become more certain, can advise national authorities more timely and accurately, thus increasing the economic value of their short-range forecasts. It is worth to mention that users consider the impact of false alarms to be, comparatively, small. In fact, false alarms play a significant role only when if they happen too often to start generating a systematic “distrust” of the earlier forecasts issued. The case where EFAS information could have an important adverse effect on early preparedness is a totally “missed” event if this incurs mistrusting the local forecasts. Since EFAS, however, by definition, covers the early warning range while the local forecasting systems pre-dominantly cover the short-range, and that national forecasting centers have usually greater confidence in their local forecasts, there is little chance for this to happen.

4. From functional and technical quality to forecast value

As could be sensed during the discussion of the six examples discussed in section 3, it is in general very difficult to quantify the functional quality of a forecast, and thus to combine it in an objective way with a measure of the technical to quantify an overall forecast value. But this quantification could be achieved in some cases (see, e.g., the use of specific concepts like NUSAP, as reported by *van der Sluijs et al. 2005*). If the functional quality can be quantified, then the forecast value can be expressed as a ‘generalized’ product \otimes of functional quality (FQ) and technical quality (TQ):

$$(1) \quad FV = FQ \otimes TQ .$$

In this section, this approach is used in some idealized cases of users interested in forecasting the occurrence of precipitation events who can take a protective action with cost C to avoid a potential loss L. The examples, based on analytically-prescribed forecast and observation fields defined by 2-dimensional Gaussian functions on a regular grid, as discussed in *Buizza (2001)*, should help the reader to identify ways to assess the forecast value in real-time applications.

Since the user’s main requirement is to decide whether to take a protective action in case a forecast occurs, the metric that is going to be used to assess technical quality is the potential economic value of the forecast based on a simple and static cost/loss model (*Murphy 1977, Richardson 2003*), a metric that depends on the forecast’s capability to discriminate between events’ occurrences/non-occurrences. The two attributes of functional quality that are taken into considerations are the ‘availability’ and ‘content and format’ ones, since in this case it is extremely important that the forecasts are correctly interpreted and that they are available as soon

as generated (i.e. not in delayed mode). Hereafter, first the approach used to assess the potential economic value is briefly reviewed, and then the forecast value of reliable/un-reliable, on-time and delayed synthetic (i.e. idealized) forecasts is discussed.

4.1. The ‘Potential Economic Value’ metric to assess technical quality

If an extreme event is predicted, the user of the forecast system has to decide whether an alarm has to be given or not (although a warning may consist of several discrete levels). This decision is always related with four possible outcomes of the decision making process (Table 3). A forecast system can produce two right decisions, a hit and an inverse hit: in these cases, an event (or no event) which is forecast is also observed (or not observed). However, the system can also produce two false decisions, a false alarm and a failure. A failure means that no event has been forecast, but the event occurs (in this article we will consider only binary systems for simplification). According to this simple decision model, the economic value of a forecast system can be calculated by combining the outcomes of the decision making process with an economic decision model like the static cost-loss model approach. If this approach is applied, then a hit and a false alarm are related with a cost C , since an alarm causes the user to protect his environment against the event at a cost C . By contrast, if no alarm is given, no protective action is done: if the event is not observed, no loss occurs, but if the event occurs the user has to face a loss L .

A pool of users can be discriminated on the basis of their cost-loss ratio C/L , which is the ratio between a cost and a loss. For instance, if a forecast system is installed for a town, than the commune and the people who live in this town can represent two different users. Both users have their own specific cost-loss ratio. The commune might have a specific cost-loss ratio of 0.01, because an alarm costs 100 thousand euros and a loss of 10 million euros. It has to be pointed out

that monetary costs (as a derivate of vulnerability) are only one way to specify the cost-loss ratio. It is worth reminding the reader that many ways to compute this cost (or vulnerability) have been postulated (see vast amount of literature on Risk assessment). A full assessment should include factors such as environmental and social impacts, pricing methods or, for example, preparedness. The quantification of the costs should be part of a public decision making process in a social scientific framework. In particular the uncertainty in quantifying the component such as cost (what is the value of a life?) can be paramount. In this paper, we will neglect the uncertainty in quantifying these costs as the methodology presented is general and the neglected factors could be included.

For users with a cost-loss ratio C/L , the mean expense E_f that they face using a forecast system can be calculated, if costs and losses are summarized and divided by the number of forecast n :

$$(2) \quad E_f = \frac{a+b}{n}C + \frac{c}{n}L$$

where a , b and c are defined in Table 3.

This average expense can be compared with the average expense of a reference forecast E_c :

$$(3) \quad E_c = \min(E_a, E_n)$$

where E_c is the minimum expense of the two following decisions: either the user always protects if the climatological base rate of an event s is smaller than the cost-loss ratio $E_a = C$, thus incurring an expense E_a , or never protects if s is greater than the cost-loss ratio $E_n = s \cdot L$, thus incurring an expense E_n . Since for some variables, e.g. the discharge, the forecasts are highly auto-correlated, a persistence forecasts might be successful in predicting an event, especially for the next forecast time-step. In this case, the average expense of a persistence forecast E_p can be used as reference:

$$(4) \quad E_p = \frac{a_p + b_p}{n_p} C + \frac{c_p}{n_p} L$$

where a_p is the number of hits, b_p is number of false alarms, c_p is the number of failures and n_p is the number of forecasts. In this case, the reference expense can be defined in the following way:

$$(5) \quad E_0 = \min(E_a, E_n, E_p)$$

Note that the average expense sustained by using a perfect forecast system E_1 is given by:

$$(6) \quad E_1 = s \cdot C$$

The average reference expense associated with the forecast E_f can be transformed into its potential economic value PEV_f using the average expenses of the reference forecast E_c and of the perfect forecast E_1 :

$$(7) \quad PEV_f = \frac{E_0 - E_f}{E_0 - E_1} .$$

PEV_f ranges between minus infinity to 1: a forecast that is better than the reference has positive PEV, and a perfect forecast has $PEV_f=1$. Equation (7) is very similar to the formulation of *Richardson (2003)*, but the reference forecast is extended by the persistence forecast. Note that, despite its name, this metric depends only on the technical quality of a forecast and does not take into account any of the attributes that characterize functional quality (Fig. 2), and thus cannot be considered as a measure of forecast value.

4.2. Technical quality of synthetic probabilistic forecasts

Two sets of synthetic probabilistic forecasts have been generated: the first set is based on reliable (un-biased) forecasts, defined by 51 Gaussian forecasts of precipitation amounts, where the 51 synthetic forecasts slightly differ from each other, and the synthetic observation field coincides with one of the forecast field. This set represents the perfect case of an ensemble of forecasts capable to always include the verification inside the probability distribution of forecast states. The second set is based on un-reliable (biased) forecasts, defined by 51 Gaussian forecasts of precipitation amounts, where the 51 synthetic forecasts slightly differ from each other, and the synthetic observation field differs from the forecast field in its shape (higher maximum, different width) and positioning (it is shifted from all the forecasts). This second set represents the case of an ensemble of forecasts that underestimate the precipitation amount, and fail to properly propagate the precipitation patten to the correct location.

Figure 11 shows the average technical quality of 5-day synthetic forecasts, generated by considering 90 different cases (i.e. realizations): the top panel shows that the forecasts of 10 mm/d are unbiased, the forecasts of 40mm/d have a ~15% positive bias, and both forecasts have positive skill when measured using the threat and the Kuipers score (*Wilks* 1995). The middle and bottom panels of Fig. 11 show the PEV_f for these two thresholds for different cost/loss ratios. Note that the PEV_f depend on the cost/loss ratio C/L . Consider the two categories of users characterized by a 10% and a 60% C/L ratios: the middle panel of Fig. 11 shows that the PEV_f (blue line) of the prediction of 10mm/d is ~60% for the users with a 10% C/L , and is ~20% for the users with a 60% C/L . Note that these two PEV_f are obtained if the users use as probability threshold to predict occurrence 10% (red line) and 60% (green line), respectively, i.e. a threshold that coincide with their C/L ratio, as it is expected from theory since this set of forecasts are unbiased (*Richardson* 2003). The bottom panel of Fig. 11 shows the corresponding results for the

prediction of 40mm/d: note that in this case that forecasts are biased, and each user PEV_f is not achieved by considering a probability threshold that coincide with his/her C/L ratio.

Figure 12 shows the average technical quality of 7-day synthetic forecasts, generated by considering 90 different cases (i.e. realizations). The top panel of Fig. 12 shows that all forecasts underestimate the precipitation amount (have a bias of 0.4 instead of 1), and have lower threat and Kuipers scores. The middle and bottom panels of Fig. 12 show that for each user, the PEV_f is lower than for the first set of forecasts, and that due to the forecast unreliability there is no correspondence between each user C/L ratio and the probability threshold that leads to the maximum PEV_f (blue line).

4.3. Functional quality of synthetic probabilistic forecasts

First, consider two users, 'A' and 'B', who has to take a decision now to protect against events that may occur in 5 days, i.e. who use a 5-day forecast: user 'A' has C/L=10%, was perfectly trained and is interested in the prediction of 10mm/day rainfall, while user 'B' has C/L=60%, was not properly trained and is interested in the prediction of 40mm/day. Then, consider two other users, 'Ad' and 'Bd', who were given exactly the same training as users 'A' and 'B', respectively, but who receive the forecasts in delay mode (i.e. they can access them 2 days later). In other words, users 'Ad' and 'Bd' can only use 7-day forecasts generated 2-days earlier to decide whether o take a protective action.

Case 1 - User 'A' with C/L=10% and interested in 10mm/day forecasts, and with perfect training

User 'A' knows that the 10mm/d forecasts are un-biased and, since he/she had perfect training, and knows exactly how to interpret probabilistic forecasts: in particular, he/she knows that the system is reliable, and thus that he/she would get the highest PEV_f if he/she uses a probability threshold that coincide with C/L to determine whether to spend C to protect against losses). In other words, for user 'A':

- the availability score of the forecast is 1 (there is no delay in getting the forecast)
- the 'content and format' score is 1 (there is no forecast misinterpretation)
- the 'support maintenance and training' score is 1 (the user had support and training and knows how to use the probabilistic forecast)
- the 'communication' score is 1 (the user has been informed of the quality of the forecast).

In this case, the functional score of the forecast is 1, and the forecast value is equal to the technical quality of the forecast, i.e. to 0.6, which is the PEV_f of the probabilistic forecast for a user with a 10% C/L (see Fig. 11, middle panel).

Case 2 - User 'B' with C/L=60% interested in 40mm/d forecasts, and with poor training

User 'B' knows that 40mm/d forecasts are biased, but since he did not have good training he/she does not know how to properly use probabilistic forecasts; in particular, he/she knows that since the forecasts are biased he should NOT use a probability threshold that coincide with C/L to determine whether to spend C to protect against losses. Suppose that he/she uses the 60% probability threshold to predict the occurrence of 40mm/d of precipitation: the bottom panel of Fig. 11 shows that the user would get a PEV_f of ~0.2 instead of ~0.4. In other words, for user 'B':

- the availability score of the forecast is 1 (there is no delay in getting the forecast)

- the ‘content and format’ score is 1 (the forecast format and content is the same as for the first user, i.e. has the right content and format)
- the ‘support, maintenance and training’ score is 0.5 (i.e. 0.2/0.4: the user did NOT have the same training as the first user, and its forecast misinterpretation lead to a reduction of the forecast value)
- the ‘communication’ score is 1 (the user has been informed of the quality of the forecast).

In this case, the functional score of the forecast is 0.5, and the forecast value is equal to the technical quality of the forecast, i.e. to 0.4, which is the PEV_f of the probabilistic forecast for a user with a 60% C/L that uses the correct probability threshold (see Fig. 11, bottom panel).

Case 3 - Users ‘Ad’ and ‘Bd’ with delayed access

Consider now the case of two users with the same attributes as users ‘A’ and ‘B’, plus a further handicap of been given access to forecast products only 2 days after they have been generated. In this case, the users have to use day-7 forecasts issued 2-days before to generate their products. This means that the actual quality of the forecasts is the one shown in Fig. 12 rather than Fig. 11, i.e. the value of the 10mm/d forecast issued by the first user ‘Ad’ is 0.4 instead of 0.6 (red lines in the middle panels of Figs 11 and 12), and the value of the 40mm/d forecasts issued by the second user ‘Bd’ is 0.05 instead of 0.2.

This reduced forecast value has to be attributed to a lower functional quality:

- for user ‘Ad’, the availability score of the forecast is 0.66 (i.e. 0.4/0.6: the delay reduces the forecast value), and the overall functional quality score is 0.66

- for user 'Bd', the availability score of the forecast is 0.25 (i.e. 0.05/0.2: the delay reduces the forecast value), the 'support maintenance and training' score is 0.5 (i.e. 0.2/0.4), and the overall functional quality score is 0.125

Table 13 summarizes the results of this discussion, and illustrates how the overall 'forecast value' can be interpreted in terms of 'technical quality' and 'function quality'.

5. Suggestions and recommendations

PREVIEW is a multi-dimensional project, drawing on the most advanced research and technological development to provide innovative geo-information services in different thematic applications at the European scale. One component of this project is the analyses of the overall forecast value of probabilistic meteo- and hydro-logical forecasts. This report is a contribution to this discussion.

First, it has been recognized that a forecast is valuable only if it has both a high *technical quality* and a high *functional quality*, and two frameworks that can be applied to assess the forecast technical and the functional qualities have been introduced. The need for both a technical and a functional quality assessment stems from the fact that, for example, a technically-perfect forecast does not necessarily have any value to its final user if it is not communicated timely and in an understandable way (i.e. in a way that the user can easily decode). The technical quality of a forecasting system is directly related to its technical specifications, it gives information on the correctness, the accuracy, the scientific maturity of the products. For example, the technical quality of a deterministic or probabilistic forecast is a measure of the accuracy of the forecast statement, with accuracy measured using a range of metrics that quantify how close the forecast

was to the observed value. The technical quality should indicate how well the predicted precipitation, water level, etc, fit to the later measured ones, it may be so called the skill of the forecast. However, it has been highlighted that the technical quality is comprehended in quite different ways by different users. The requirement on the technical quality could be quite different when the user looks over a year's performance or when he looks at a specific hazardous event (Edlund 2007). The functional quality of a forecasting system depends on the efficiency of the service to meet the user's needs, resulting in user satisfaction and productivity. It is directly related to the capability of the service to be understood, delivered and used in accordance to the user's expectations.

Second, two frameworks have been introduced to assess the technical and functional quality of a forecasting system:

- The 'Technical Quality Framework' (Fig. 1), based on the assessment of four key technical attributes: 'forecast characteristics', 'validation characteristics', 'metric' and 'significance test types'
- The 'Functional Quality Framework' (Fig. 2), been based on the assessment of four key functional attributes: 'availability and means of distribution', 'content and format', 'support, maintenance and training' and 'communication of product's technical quality'

Third, these frameworks have been applied to six real-time verification problems, to illustrate how they can be used to structure in a more appropriate way the problem of the validation of a forecasting system.

Finally, the issue of the quantification of functional quality has been discussed in more details, and the possibility to estimate the forecast value as a generalized product of a measure of

technical quality and functional quality has been discussed. Four cases of ‘synthetic’ (idealized) users have also been analyzed. These examples have indicated that if the functional quality can be measured in an objective way, then it is possible to quantify the overall forecast value. If this is not possible, users should at least try to assess the functional quality of their forecasts, and should not misinterpret technical quality as a measure of forecast value.

It is suggested that the approach introduced in this work is applied to real-case examples, in particular to some of the different aspects of the PREVIEW project.

References

- Beven, K.J., 2006. A manifesto for the equifinality thesis. *Journal of Hydrology*, **320**(1-2), 18-36.
- Bougeault, P., 2003. The WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. *Technical report, WMO*. Available at <http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Bougeault/BougeaultVerification-methods.htm>.
- Buizza, R., 2001: Accuracy and economic value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev.*, 129, 2329-2345.
- Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., & Vitart, F., 2007: The new ECMWF VAREPS. *Q. J. R. Meteorol. Soc.*, under revision.
- De Roo, A., Wesseling, C. G. & Van Deurssen, W. P. A., 2000: Physically based river basin modelling within a GIS: the LISFLOOD model. *Hydrological Processes*, Vol. 4, Issues 11-12, p.1981-1992.
- De Roo, A., Thielen, J. & Gouweleeuw, B., 2002: LISFLOOD, a distributed water balance, flood simulation and flood inundation model. User manual. Version 1.0. Report of the European Commission, Joint Research Centre, Special Publications No. I.02.131.
- De Roo, A., Gouweleeuw, B., Thielen, J., Bates, P., Hollingsworth, A. et al., 2003: Development of a European Flood Forecasting System. *International Journal of River Basin Management*, Vol. 1, No. 1, p.49-59.
- Edlund, C., 2007: PREVIEW Flood platform: a general framework to measure “functional quality” of products/services. *SMHI report*, in press.
- Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., & Butts, M. B., 2004: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, **298**, 222–241 (doi:10.1016/j.jhydrol.2004.03.037, 2004. 2146, 2147, 2163).

Buizza et al, 2006: EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts (version 06/03/07, 14:13)

Gouweleeuw, B., Thielen, J., Franchello, G., De Roo, A., & Buizza, R., 2005: Flood forecasting using medium-range probabilistic weather prediction. *Hydrological and Earth System Sciences*, **9**(4), p.365-380.

Janssen, P.H.M., Petersen, A.C., van der Sluijs, J.P., Risbey, J.S. & Ravetz, J.R., 2004. Towards Guidance in Assessing and Communicating Uncertainties. In: K.M. Hanson and F.M. Hemez (Editors), *Sensitivity Analysis of Model Output*, (Los Alamos National Laboratory, Los Alamos, 2005, <http://library.lanl.gov/ccw/samo2004/>), pp. 201-210.

Katz, R. W., Murphy, A. H., & Winkler, R. L., 1982: Assessing the value of frost forecasts to orchardists: a dynamic decision-making approach. *J. Appl. Meteorol.*, **21**, 518-531.

Konikow, L.F. and Bredehoeft, J.D., 1992. Groundwater Models Cannot Be Validated. *Advances in Water Resources*, **15** (1), 75-83.

Mason, I, 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S, Molteni, F., & Buizza, R., 2001. A strategy for high-resolution ensemble prediction. Part II: limited-area experiments in four Alpine flood events. *Q. J. R. Meteorol. Soc.*, **127**, 2095-2115.

Marsigli C., Boccanera F., Montani A., & Paccagnella, T., 2005: The COSMO-LEPS ensemble system: validation of the methodology and verification. *Non-linear Processes in Geophysics*, Vol. **12**, pp. 527-536.

Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F., & Paccagnella, T., 2001. A strategy for high-resolution ensemble prediction. Part I: definition of representative members and global-model experiments. *Q. J. R. Meteorol. Soc.*, **127**, 2069-2094.

Montani A., Capaldo, M., Cesari, D., Marsigli, C., Modigliani, U., Nerozzi, F., Paccagnella, T., Patruno, P. & Tibaldi, S., 2003. Operational limited-area ensemble forecasts based on the Lokal Modell. *ECMWF Newsletter*, 98, pp.2-7.

Buizza et al, 2006: EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts (version 06/03/07, 14:13)

Murphy, A. H. 1977. The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803-816.

Murphy, A. H., 1985: Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Mon. Wea. Rev.*, **113**, 362-369.

Oreskes, N., Shrader-Frechette, K. & Belitz, K., 1994. Verification, Validation, and Confirmation of Numerical-Models in the Earth-Sciences. *Science*, **263** (5147), 641-646.

Pappenberger, F. & Beven, K.J., 2006. Ignorance is bliss - or 7 reasons not to use uncertainty analysis. *Water Resources Research*, 42(5): doi: 10.1029/2005WR004820.

Pappenberger, F., Harvey, H., Beven, K.J & Hall J., 2006, Choosing an uncertainty analysis: a Wiki experiment. *Hydrological processes*, **20** (17), 3793-3798.

Ramos, M-H, Bartholmes, J., Thielen, J., Kalas, M., & de Roo, A., 2006: The additional value of ensemble weather forecasts to flood forecasting: first results on EFAS forecasts for the Danube River Basin. Proceedings of the *XXIII Conference of the Danubian Countries on the Hydrological Forecasting and Hydrological Bases of Water Management*, 28-31 August 2006, Belgrade, Republic of Serbia, CD-ROM, 12p.

Richardson, D. S., 2000: Skill and economic value of the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **126**, 649-668.

Richardson, D. S. 2003. Economic Value and Skill. In: *Forecast Verification - A Practitioner's Guide in Atmospheric Science* (eds. I. T. Jolife and D. B. Stephenson). John Wiley and Sons Ltd, 165 – 188.

Theis S. E, A. Hense & U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications*, **12** (3), 257-268.

Buizza et al, 2006: EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts (version 06/03/07, 14:13)

Thielen, J., 2004: Flood forecasting results: Europe. In: Gouweleeuw, B., Reggiani, P. & De Roo, A. (ed.): *A European Flood Forecasting System EFFS*. Full Report, European Commission, EUR 21208 EN, p.187-197.

Thielen J, Ramos, M.H., Bartholmes, J., De Roo, A., Cloke, H., Pappenberger, F., & Demeritt, D., 2005: Summary report of the *1st EFAS workshop on the use of Ensemble Prediction System in flood forecasting*, 21-22nd November 2005, Ispra. European Report EUR 22118 EN, European Commissions 2005, 23p.

Thielen, J., Bartholmes, J., Ramos, M-H., Kalas, M., van der Knijff, J. & De Roo, A., 2006a: Added value of ensemble prediction system products for medium-range flood forecasting on European scale. In: Proceedings of the workshop “Ensemble Predictions and Uncertainties in Flood Forecasting”, International Commission for the Hydrology of the Rhine Basin (CHR), Bern Switzerland, 30-31 March 2006, p.77-82.

Thielen, J., Bartholmes, J., Ramos, M.-H., Franchello, G. & de Roo, A., 2006b: European Flood Alert System (EFAS): Evaluation of EFAS Results 2005/2006. EUR Report, in press.

van der Sluijs, J., van Eijndhoven, J., Shackley, S. and Wynne, B., 1998: Anchoring devices in science for policy: The case of consensus around climate sensitivity. *Social Studies of Science*, **28** (2), 291-323.

van der Sluijs, J.P., Risbey, J. and Ravetz, J., 2005: Uncertainty assessment of VOC emissions from paint in the Netherlands using the NUSAP system. *Environmental Monitoring and Assessment*, **105** (1-3), 229-259.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences*, Academic Press, pp 467 (ISBN 0-12-751965-3).

Wilks, D., & Hamill, T. M., 1995: Potential economic value of ensemble-based surface weather forecasts. *Mon. Wea. Rev.*, **123**, 3564-3575.

Buizza et al, 2006: EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts (version 06/03/07, 14:13)

Yates, E., S. Anquetin, V. Ducrocq, J.-D. Creutin, D. Ricard, K. Chancibault, 2006 : Point and areal validation of forecast precipitation fields. *Meteorol. Appl.*, 1-20.

Forecast characteristics	Fc variables	Which is/are the forecast variables of interest?
	Area	Which is/are the area of interest?
	Resolution	What is the resolution of interest (spatial, temporal)?
	Calibration	Which data set has been used for calibration? Which type of calibration framework has been used? Which performance measure or error model?
Verification characteristics	Verification	Which verification field is used to assess the fcs? Data set? Data origin?
	Verification uncertainty	Has ‘observation’ uncertainty been taken into account?
Metric	Average or ‘unique’	Is the focus of the investigation the average performance or the prediction of ‘rare’ events?
	Skill single fcs	Which metrics are used to assess single fcs? If multiple metrics are combined to one – which ones and how have they been combined
	Spread of ensemble	Which method/metrics are used to assess whether the ensemble spread has the right level of spread?
	Skill prob fcs	Which metrics are used to assess probabilistic fcs?
	Occurrence / non-occurrence	Which metrics are used to assess whether the fc can distinguish occurrence and non-occurrence of events?
	Other metrics	Is there any other ‘metric’ that should be considered?
Statistical significance	Signif tests	Which significance test is going to be used?

Table 1. List of some of the key attributes of the framework (Fig. 1) used to assess technical quality.

Availability and means of distribution	Availability	How often are forecasts updated? When are they available on the forecast bench?
	Means of distribution	Which communication channels are used to receive the forecasts? Are these means reliable? Are they secure?
Content and format	Content	Are forecasts understandable? Can they be easily interpreted?
	Format	Which format is used? Text? Images? Coded language?
Support maintenance and training	Support	Is there any form of support given to users? (e.g. helplines, web information service, ..)
	Maintenance	Are products maintained? How often are they changed? If they are generated using post-processing software, how often is this software maintained/updated?
	Training	Are training sessions organized to inform the users on products' format? Do users know which forecast should be used for their application?
Communication of product's technical quality	Communication in general	Are users informed on changes in the product generation system? Is there an open/established communication channel between users and forecast generators?
	Communication of forecast technical quality	Are users informed about the quality of each forecast product?

Table 2. List of the key attributes of the framework (Fig. 2) used to assess functional quality.

Cost/Loss model		Observation			
		Yes		No	
Forecast	Yes	a	Cost C	b	Cost C
	No	c	Cost L	d	Cost 0

Table 3. Possible Outcomes (hit a, false alarm b, failure c and inverse hit d) and expense matrix (protection cost C, loss L) of a decision making process for a decision maker that takes a protective action or not. A hit and a false alarm a related with a cost C, whereas a failure of the system causes a loss L.

Forecast characteristics	Fc variables	12-hour accumulated precipitation
	Area	Northern Hemisphere and Europe
	Resolution	2.5 degree, regular lat/long grid
	Calibration	None
Verification characteristics	Verification	Proxy defined by 24-hour forecasts
	Verification uncertainty	No
Metric	Average or 'unique'	Average performance
	Skill single fcs	RMSE, MAE
	Spread of ensemble	STD versus error of the ensemble-mean forecast
	Skill prob fcs	RPS, RPSS, BS, BSS
	Occurrence / non-occurrence	Area under a ROC curve
	Other metrics	No
Statistical significance	Signif tests	No

Table 4. List of some of the key entries of the general framework used to assess the technical quality of the ECMWF EPS system.

Forecast characteristics	Fc variables	24-hour accumulated precipitation
	Area	Europe
	Resolution	1.5 degree, regular lat/long grid, forecasts up to t+120h
	Calibration	None
Verification characteristics	Verification	Observations
	Verification uncertainty	No
Metric	Average or 'unique'	Average performance
	Skill single fcs	No
	Spread of ensemble	No
	Skill prob fcs	BSS
	Occurrence / non-occurrence	No
	Other metrics	Cost/loss analysis
Statistical significance	Signif tests	No

Table 5. List of some of the key entries of the general framework used to assess the technical quality of the COSMO-LEPS system.

Forecast characteristics	Fc variables	6- and 24-hour accumulated precipitation
	Area	Europe, river catchments
	Resolution	7 km regular lat/long grid, forecasts up to t+50h
	Calibration	None
Verification characteristics	Verification	Observations
	Verification uncertainty	No
Metric	Average or 'unique'	Unique events
	Skill single fcs	Mean error
	Spread of ensemble	No
	Skill prob fcs	No
	Occurrence / non-occurrence	No
	Other metrics	No
Statistical significance	Signif tests	No

Table 6. List of some of the key entries of the general framework used to assess the technical quality of the COSMO-LME system.

Forecast characteristics	Fc variables	18-hour accumulated precipitation
	Area	South-East France (42-46°N, 2-8°E)
	Resolution	2-3 km grid, forecasts up to t+18h
	Calibration	None
Verification characteristics	Verification	Observations (rain gauges)
	Verification uncertainty	No
Metric	Average or 'unique'	Unique events (flood events)
	Skill single fcs	Bias
	Spread of ensemble	No
	Skill prob fcs	No
	Occurrence / non-occurrence	POD and FAR for 10 mm/d events
	Other metrics	No
Statistical significance	Signif tests	No

Table 7a. List of some of the key entries of the general framework used to assess the technical quality of the kilometric scale deterministic forecast from COSMO-ALMO2, COSMO-LAMI, Meso-NH/AROME, MM5 models.

Forecast characteristics	Fc variables	1-hour forecast discharge forced by 1-hour observed or forecast precipitation
	Area	Cévennes-Vivarais watersheds
	Resolution	1 km grid, forecasts up to t+18h
	Calibration	None
Verification characteristics	Verification	Observations (discharges)
	Verification uncertainty	No
Metric	Average or 'unique'	Unique events (flash-flood events)
	Skill single fcs	Nash coefficient
	Spread of ensemble	No
	Skill prob fcs	No
	Occurrence / non-occurrence	No
	Other metrics	No
Statistical significance	Signif tests	No

Table 7b. List of some of the key entries of the general framework used to assess the technical quality of the hydrological models forced by deterministic high-resolution 1-h accumulated precipitation forecast.

Forecast characteristics	Fc variables	River discharge
	Area	River basins
	Resolution	5.5 km regular lat/long grid, forecasts up to t+240h
	Calibration	None
Verification characteristics	Verification	Observations
	Verification uncertainty	No
Metric	Average or 'unique'	1-year average
	Skill single fcs	MAE, Bias
	Spread of ensemble	No
	Skill prob fcs	No
	Occurrence / non-occurrence	POD and FAR for single forecasts
	Other metrics	No
Statistical significance	Signif tests	No

Table 8. List of some of the key entries of the general framework used to assess the technical quality of the EFAS system.

Question	yes	no	Don't know
Do you find EFAS information reports useful	7		
Flood ensemble prediction system information is given in the form of maps counting the number of ensemble forecasts generating discharges exceeding critical flood level thresholds. Do you find this information useful?	7		
Were the EFAS reports used in some way by the flood forecasting team?	6	1	
		<i>(the reports arrived too late because of technical problems with the receiving mail server)</i>	
Did the EFAS reports effectively help you?	6	1	
		<i>(no earlier information than from local sources)*</i>	

* with the new rules of dissemination (send reports to all authorities within the catchment and not only when the area of MoU is affected) this authority would have been informed 5 days earlier.

Table 9. Usefulness and impact of EFAS information reports.

Question: How were the EFAS information reports used?	Category*		
Answers	Early warning	Additional information	Decision making
It was the first warning that focused our attention to the Drava river	X	X	
The reports were useful for the estimation of peak discharge		X	
We are using the reports as indication (the local 48h forecasts from Meteo-France Aladin and DWD HRM are used in a quantitative respect)	X	X	
We get an overview of the situation in the whole catchment, e.g. which tributaries are affected		X	
It is good to know which general development is predicted by EFAS	X	X	
EFAS reports are used to present the hydrological situation in the near days to institutes responsible for flood protection	X	X	X
EFAS reports were used as orientation information	X	X	
EFAS reports were used as support to create statement of development of flood situation		X	X

*The category was estimated a posteriori and not ticked by the partner organizations

Table 10. Type of use of EFAS information reports.

Question	Yes	no	Don't Know
Is EFAS information clearly stated	6	1	
Is all information necessary	7		
Would you suggest improvements of the EFAS reports	1	6	

Table 11. Editorial aspects of EFAS Information Reports.

Timeliness and dissemination of Report			
You consider 16:00 a good time to receive forecasts	Yes: 1	Too late: 3	Does not matter: 3
The forecasters receive EFAS reports	Directly: 5	Via technical contact: 2	
The forecasters receive EFAS reports	At the time it was sent: 5	Within the following 24 h: 1	Later than 24 h: 1

Table 12. Timeliness and dissemination of reports.

Forecast type	User	Scores FQ attributes				Func Qual (FQ)	Tech Qual (TQ)	Forecast value FV=FQ*TQ
		Avail	Cont	Train	Comm			
Rel 10	A	1.0	1.0	1.0	1.0	1.0	0.6	0.6
Biased 40	B	1.0	1.0	0.5	1.0	0.5	0.4	0.2
Delayed rel 10	Ad	0.66	1.0	1.0	1.0	0.66	0.6	0.4
Delayed bias 40	Bd	0.25	1.0	0.5	1.0	0.125	0.4	0.05

Table 13. Functional quality, technical quality and forecast value of the (synthetic) forecasts issued by users who are given access to reliable 10mm/d forecasts (user A, 2nd row), biased 40mm/d forecasts (user B, 3rd row), and to corresponding delayed forecasts (user 'Ad', 4th row, and user 'Bd', 5th row).

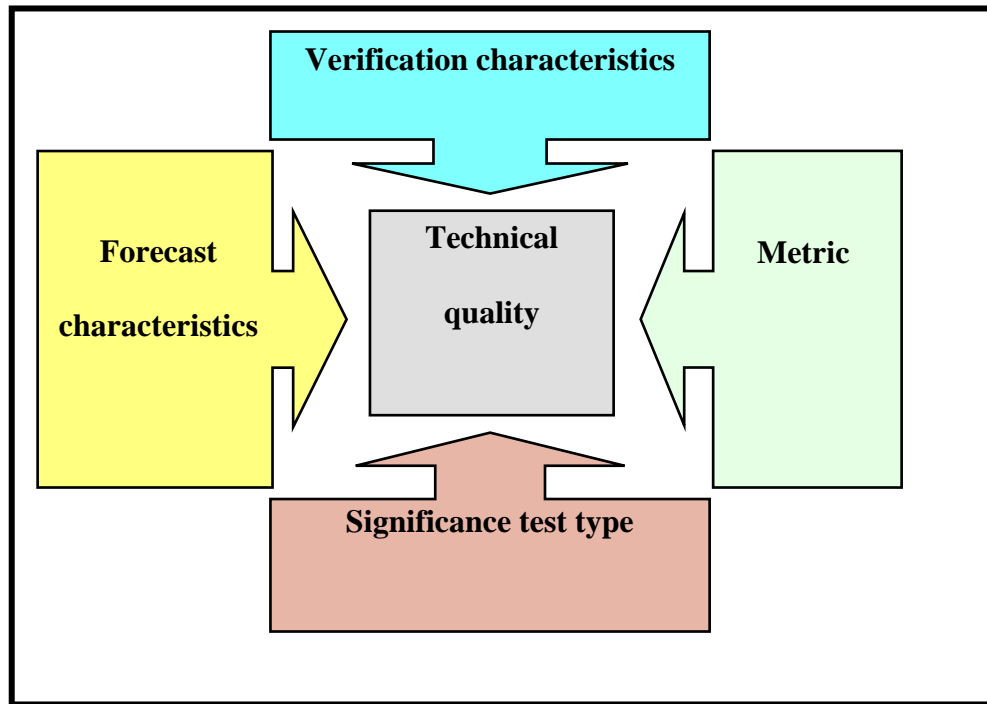


Figure 1. Schematic of the framework used to assess the technical quality of a forecast.

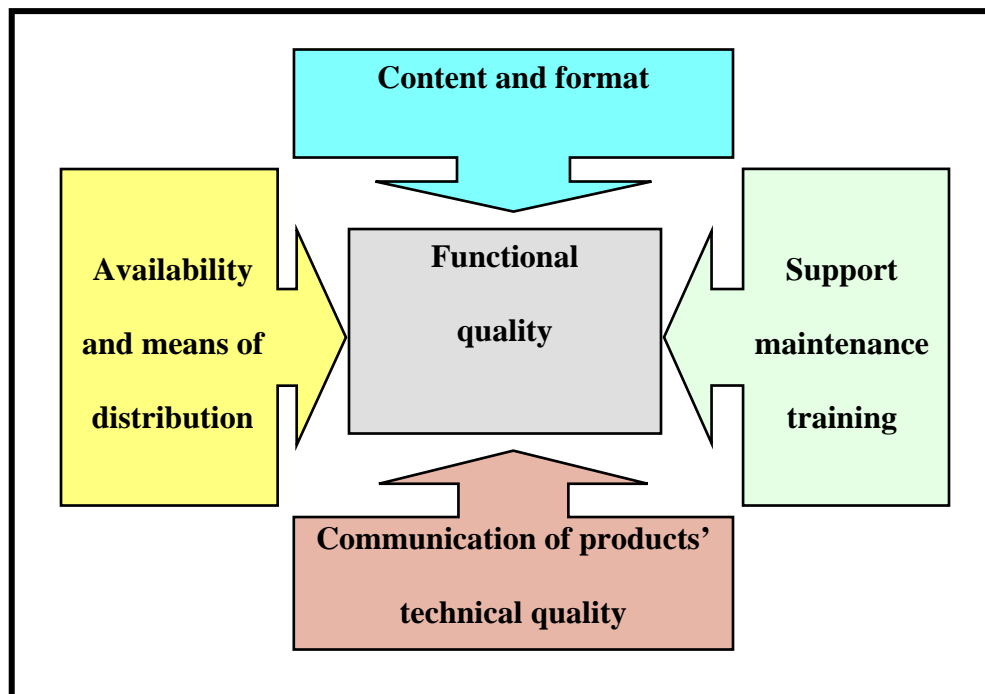


Figure 2. Schematic of the framework used to assess the technical quality of a forecast.

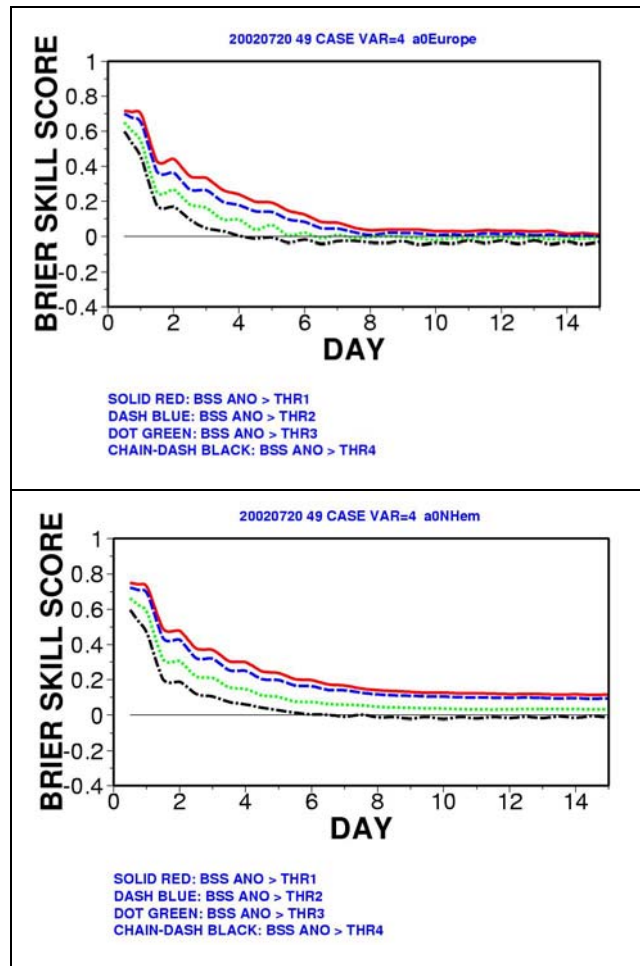


Figure 3. Average Brier skill score of the probabilistic precipitation prediction of VAREPS forecasts over Europe (top panel) and Northern Hemisphere (bottom panel), for total precipitation in excess of 1 mm (solid red lines), 2 mm (dashed blue lines), 5 mm (dotted green lines) and 10 mm (chain-dashed black lines). Averages have been computed over the PREVIEW special period (20 July to 31 August 2002), a 0-24h $T_L511L60$ forecast has been used as a proxy for verification, and the skill score has been computed using the sample climatology as reference.

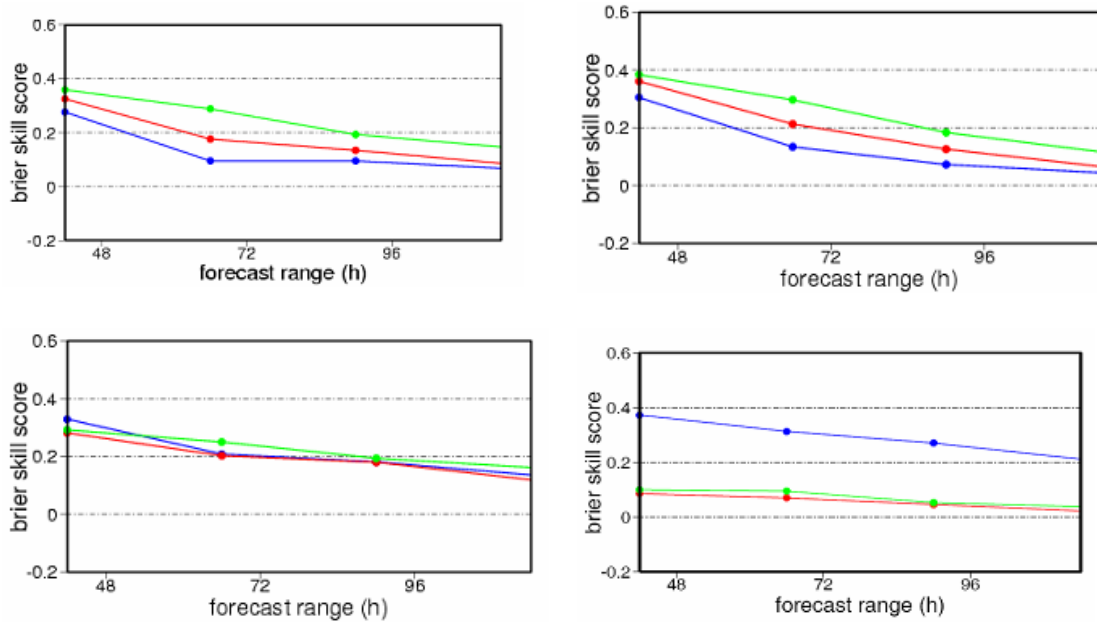


Figure 4. Brier Skill Score relative to the event “precipitation exceeding 10mm/24h” for different forecast ranges. Top left: average values; top right: 50th percentile; bottom left: 90th percentile; bottom right: maximum values. Blue lines are relative to the COSMO-LEPS system, red lines to the small-size EPS and green lines to the full-size EPS.

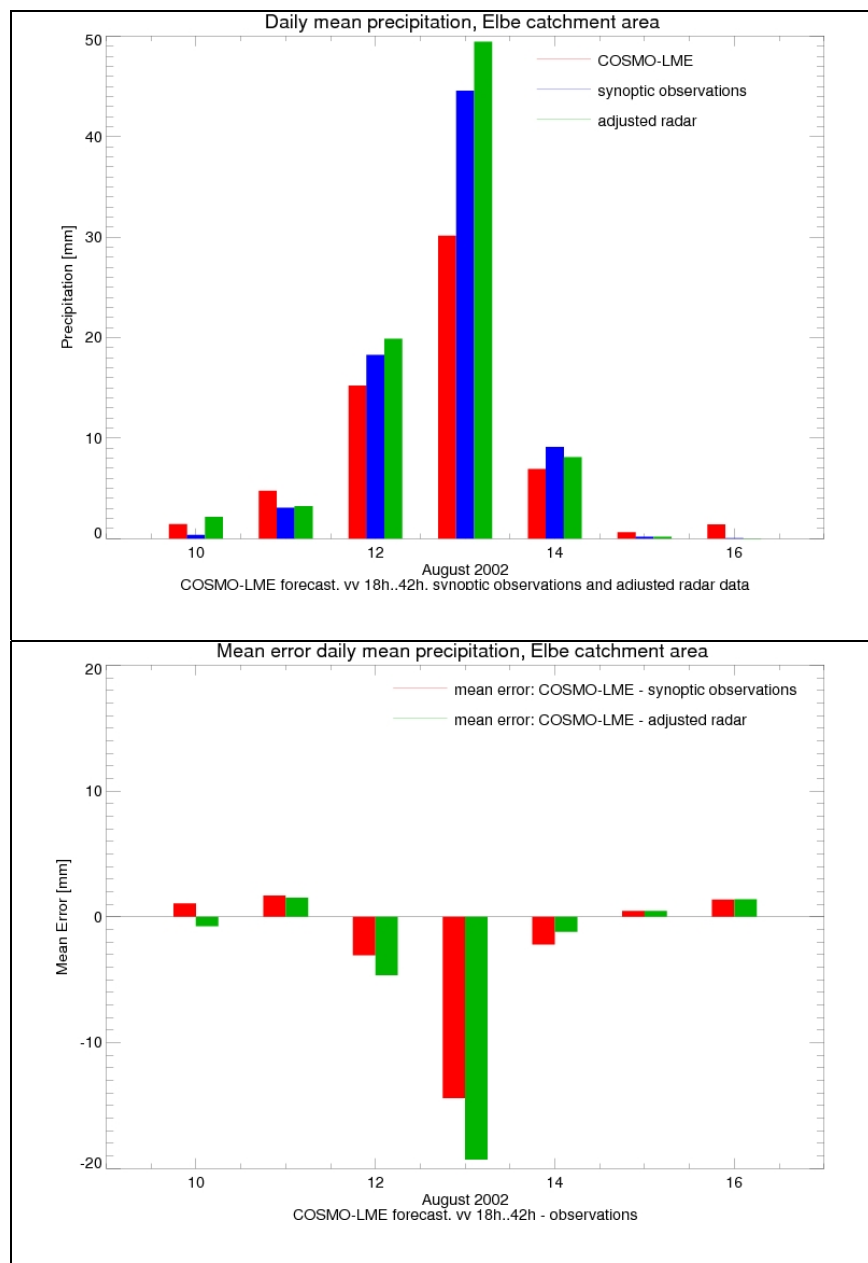


Figure 5: COSMO-LME total precipitation (top panel) and mean error (bottom panel) for the period 10 August 2002 to 16 August 2002 in the Elbe catchment area. See text for details.

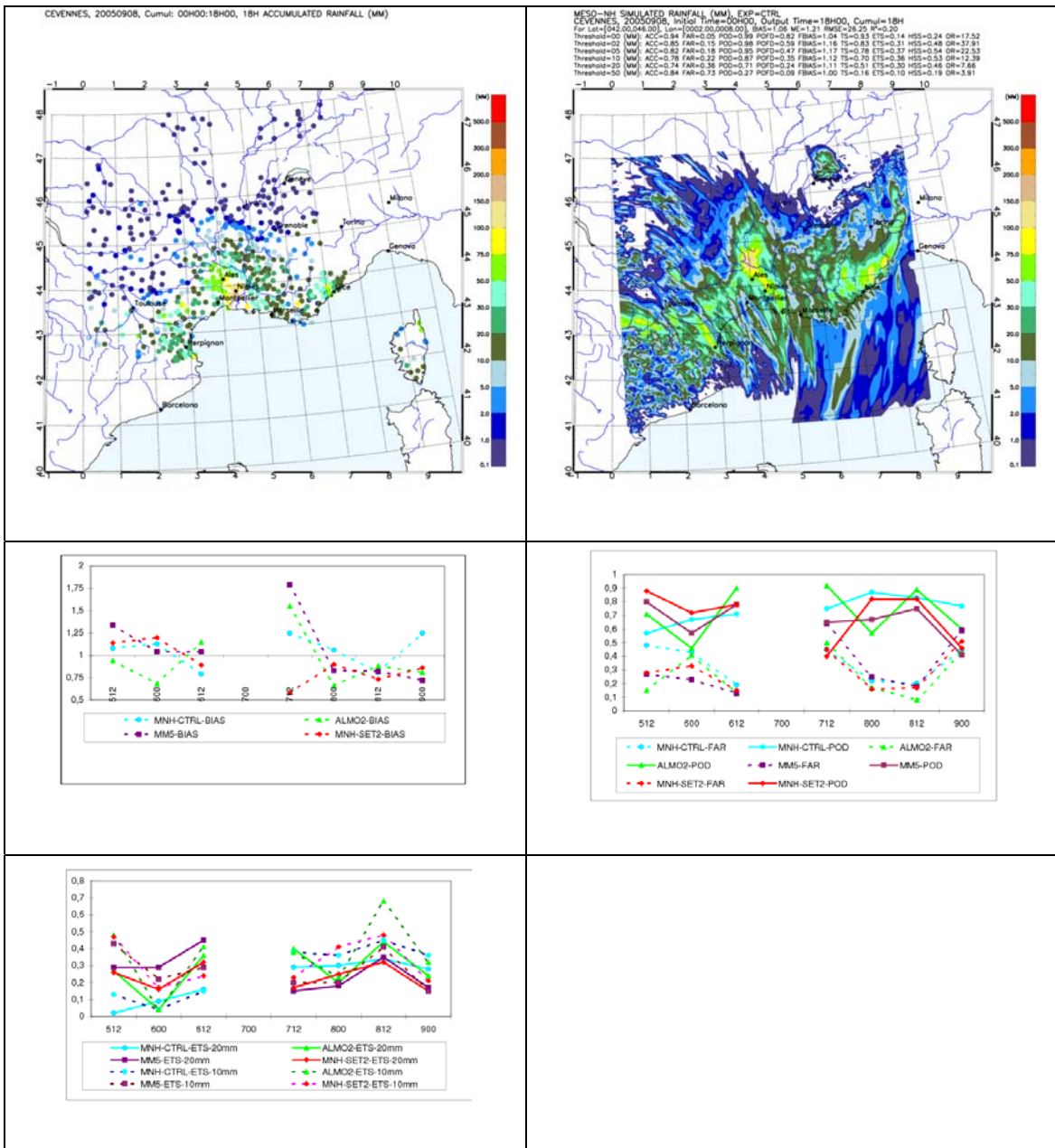


Figure 6: Bias (middle left) and FAR and POD (middle right) for the event “precipitation exceeding 10 mm/18 h” and ETS (bottom left) the events “precipitation exceeding 10 mm/18 h and 20mm/18h” for MESO-NH (2 different configurations tested), COSMO-ALMO2 and MM5 models starting each 12 hours from 5 September 2005, 12 UTC to 9 September 2005, 00UTC (the

Buizza et al, 2006: EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts (version 06/03/07, 14:13)

results for the runs starting , 7 September, 00 UTC are not shown as no intense precipitation are observed for that period). Top panels show an example of 18-h accumulated precipitation from raingauge observations (left) and 2.5 km Meso-NH forecast (right) between 00 UTC to 18 UTC, 8 September 2005. Model forecasts are interpolated to raingauge observations before computing scores.

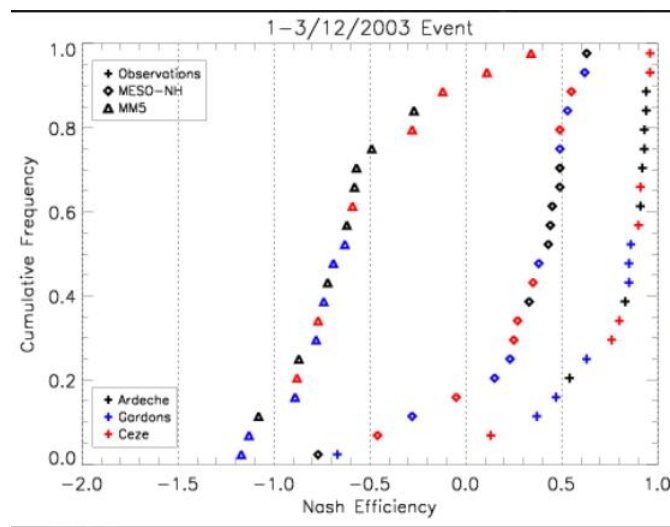


Figure 7. Nash coefficients obtained when the hydrological TOPMODEL model is forced by raingauge observations, Meso-NH forecast and MM5 forecast for the flash-flood of 1 to 3 December 2003. The three main watersheds (Ardèche, Gardon and Cèze) of the Cévennes-Vivarias region are considered and Nash coefficients for different outlets for each watershed are issued.

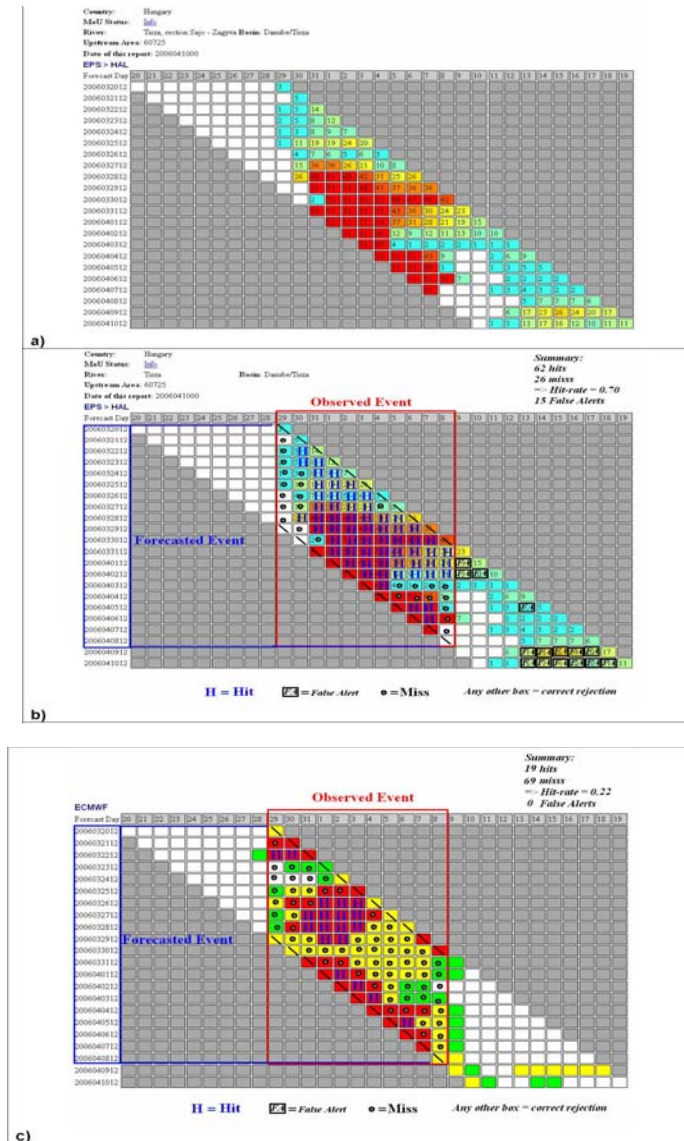


Figure 8. Illustration of the methodology adopted: a) EFAS diagram of forecasted alert levels from 20th March to 10th April 2006 for EPS-based EFAS forecasts; b) indication of hits, misses and false alerts on the criteria that a forecast event needs at least 5 EPS-based simulations on two consecutive simulations above EFAS high alert levels; c) EFAS diagram of forecasted alert levels for EFAS forecasts based on deterministic ECMWF weather forecasts and indication of hits, misses and false alerts on the criteria that a forecast event needs at least two consecutive simulations above EFAS high alert levels.

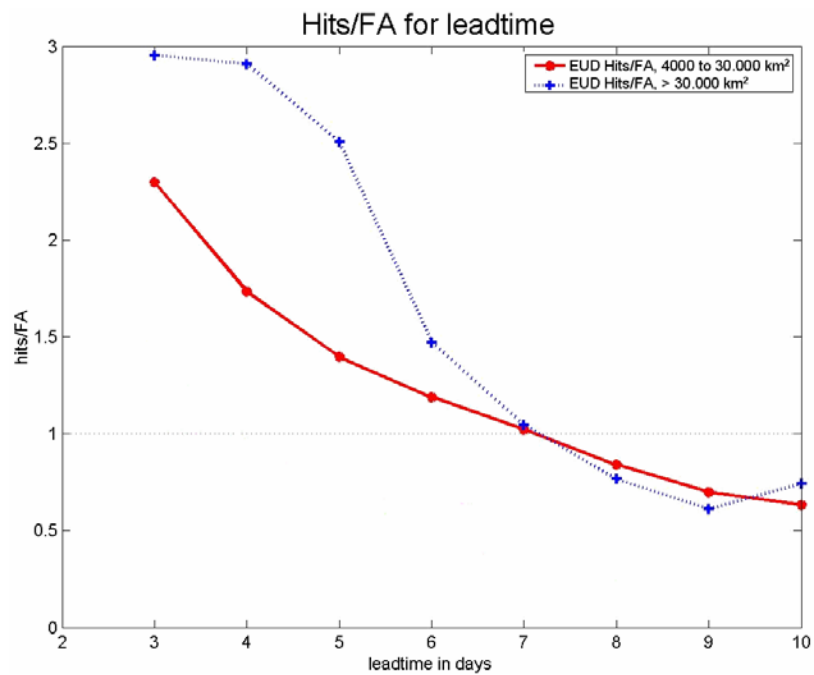


Figure 9. Hits/false-alarms for EFAS deterministic ECMWF forecasts for 2 different upstream area classes plotted over lead time.

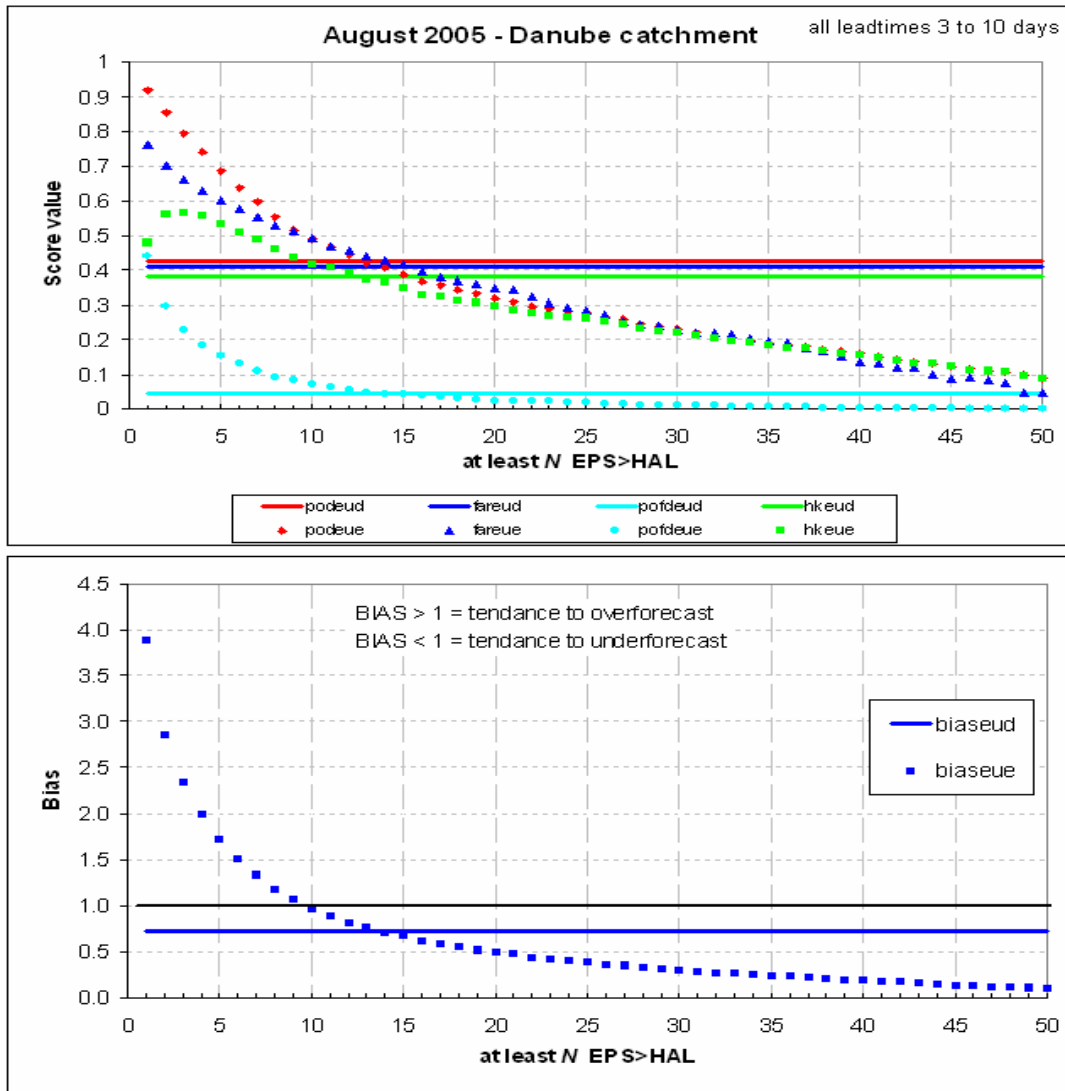


Figure 10. Probability of detection (hit rate) - pod, False alarm ratio-far, Probability of false detection (false alarm rate)-pof, Hanssen and Kuipers discriminant (true skill statistic, Peirces's skill score)-hk and Bias for EUD and EUE.

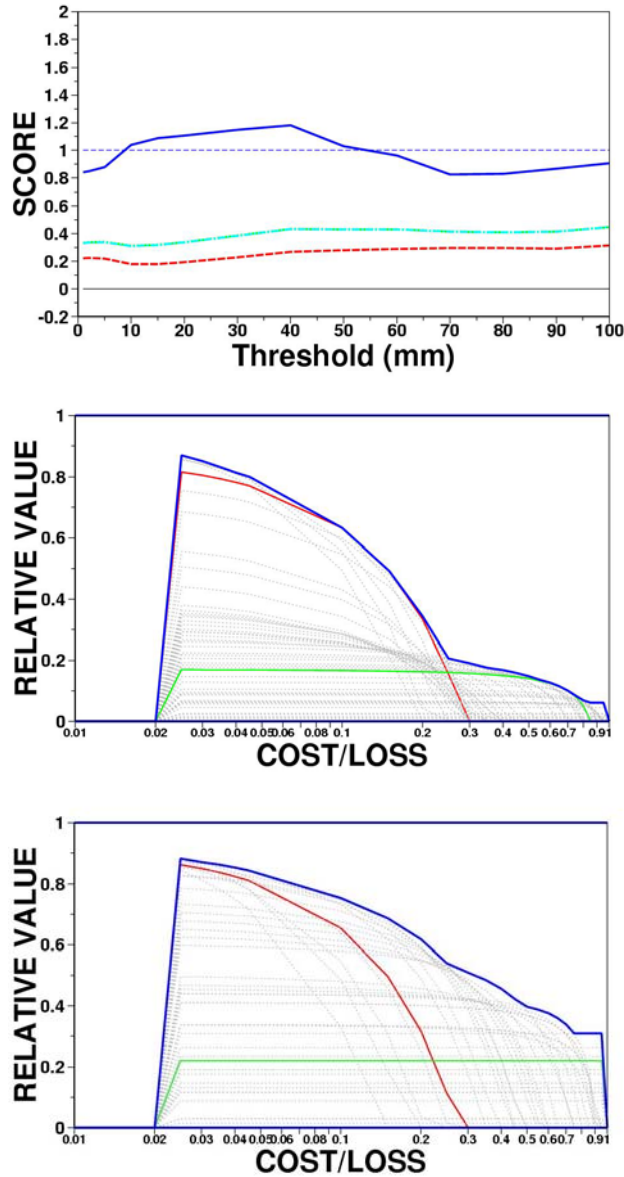


Figure 11. Reliable synthetic probabilistic forecasts. Top panel: average bias (blue line), threat score (red dashed line) and Kuipers skill score (cyan chain-dashed line) of the 51 single synthetic forecasts. Middle panel: potential economic value of the synthetic probabilistic prediction of precipitation in excess of 10 mm/d of different probabilistic thresholds (grey lines), of the 10% (red line) and the 60% (green line) probabilistic thresholds, and of the whole ensemble (defined as the envelop of all probabilistic curves, blue line). Bottom panel: as middle panel but for the synthetic probabilistic prediction of precipitation in excess of 40 mm/d.

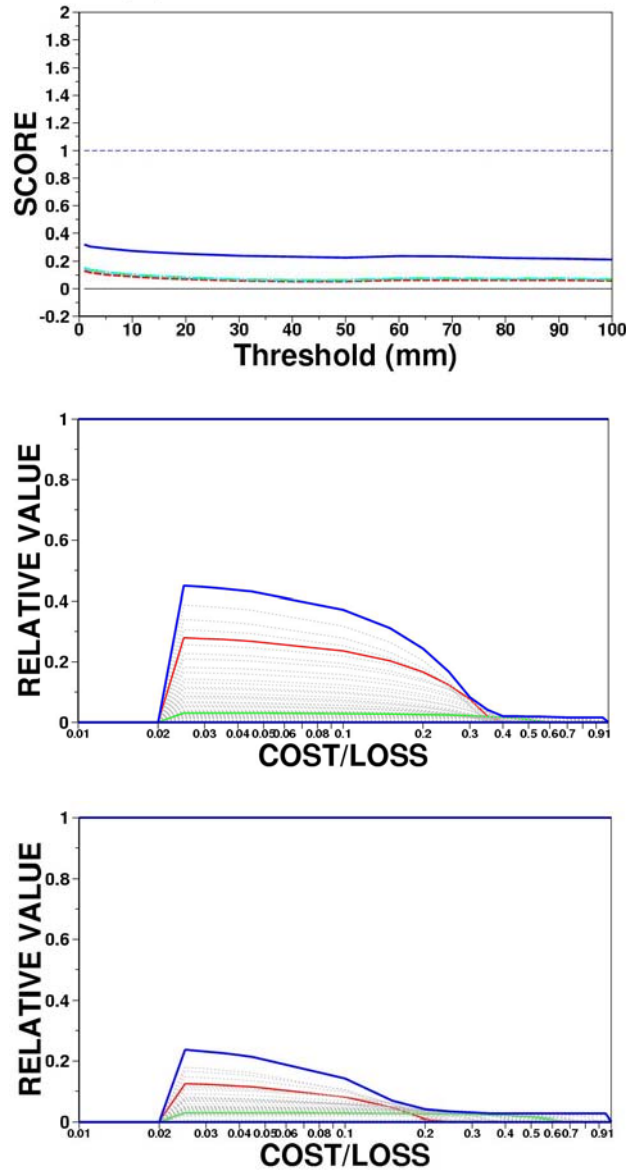


Figure 12. Unreliable (biased) synthetic probabilistic forecasts. Top panel: average bias (blue line), threat score (red dashed line) and Kuipers skill score (cyan chain-dashed line) of the 51 single synthetic forecasts. Middle panel: potential economic value of the synthetic probabilistic prediction of precipitation in excess of 10 mm/d of different probabilistic thresholds (grey lines), of the 10% (red line) and the 60% (green line) probabilistic thresholds, and of the whole ensemble (defined as the envelop of all probabilistic curves, blue line). Bottom panel: as middle panel but for the synthetic probabilistic prediction of precipitation in excess of 40 mm/d.