Hydrol. Earth Syst. Sci. Discuss., 5, S688–S696, 2008
www.hydrol-earth-syst-sci-discuss.net/5/S688/2008/

Hydrology and Earth System Sciences Discussions

Interactive comment on "The European Flood Alert System EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts" *by* J. C. Bartholmes et al.

J. C. Bartholmes et al.

Received and published: 17 July 2008

We would like to thank the Reviewer for providing interesting remarks and suggestions to improve the scientific quality of our paper. In the following we will respond to these suggestions one by one.

R2: "Since comparable results have been published for other regional forecasting system it might be interesting to compare the skills obtained by EFAS with those obtained by other systems (cf. list of studies given in the introduction)"

Answer: A rigorous comparison of EFAS results with results obtained from regional hydrological forecasting systems is not straightforward, mainly due to the large differences between the way inputs (weather forecasts) and outputs (discharge forecasts

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

and alert level exceedances) are processed. The differences in space and time scales, as well as in the setup of the hydrological model, need also to be assessed before any statistical comparative analysis. Currently, no scientific project has been dedicated to such an assessment, which implies the set up of a rigorous protocol for a sounded based comparison (i.e., same input data, same time period of analysis, same verification skill scores, etc.). Within the EFAS project, preliminary comparative analysis of the alert levels issued during a flood event has only been performed for individual case-studies and been recently published (Kalas et al., 2008; Younis et al., 2008). For a statistical skill assessment, we have checked the literature trying to find results against which the EFAS performance could be compared. However, the EFAS skill score analysis is unique in the sense that it has been calculated for entire Europe, for each river pixel, over a period of 2 years, with a minimum of input data, and with exclusively probabilistic skill scores. We have found papers in literature refering to the Nash coefficient, which we did not want to use for the reasons we stated on p.294 l.20-22. Roulin et al (2005) calculated the BSS but for a much smaller catchment and with high-resolution input data available, which would make a comparison very difficult with EFAS. Further, typically skill scores such as the Brier Skill Score depend to a large degree on the chosen climatology. Without having comparable input data, the comparison with skill scores from other studies remains difficult. Within the scope of our study, what can be eventually compared, and this is an important finding of the paper, is that the BSS from EFAS hydrological forecasts is higher than the BSS calculated for the rainfall inputs. We will make this clearer in the paper.

R2: "I feel that some of the terminology that stems from meteorology needs to be better explained in order to make the reading a little bit more straightforward for the hydrological community (…)."

Answer: The differences in terminology between hydrology and meteorology are a well known problem (see Pappenberger et al., 2008). The authors will go carefully again through the paper and better explain those terms where misunderstanding may arise.

HESSD

5, S688–S696, 2008

R2: "The authors argued that the Nash Sutcliffe criterion is not a suitable criterion for evaluating probabilistic forecast performances. They are of course right but due to the wide spread of this particular performance measure it would have been nonetheless interesting to present the Nash values that were obtained – for example – by the ensemble mean."

Answer: The use of Nash Sutcliffe performance criteria in the evaluation of hydrologic simulations and model performance in calibration is indeed widely recognized in the literature. However, the authors strongly do not recommend the use of Nash Sutcliffe for probabilistic forecast assessment. In flow forecasting, other authors have also discouraged its use and instead promoted the use of a persistence index (see for instance Anctil et al., 2004). In the paper, we have also tried to explain that probabilistic forecasts require different approaches than deterministic forecasts and that the Nash coefficient is a deterministic skill score and therefore should not be applied for EPS based forecasts. Additionally, collapsing a probabilistic forecasting into the ensemble mean is a deterministic way of thinking and the author believe should not be promoted because it entirely ignores the spread of the ensembles and the value of individual EPS members.

R2: "Is it really necessary to present all skill tests given that many of them provide very similar results? If you want to keep all of them, please specify the added value of each one."

Answer: The authors thank the reviewer for this comment that allows us to make an important point clearer in our manuscript. Different skill scores address different parts of the contingency table and therefore weight the information and performance from the EPS based forecasts. It can happen that one score would give a positive result although the forecasts themselves are not very good. A good example for this is illustrated by Pappenberger et al. in “ Medium range multi model weather forecast ensembles in flood forecasting (a case study), (Technical Memorandum Nr 557, ECMWF). In figure 7, page 15 of the Technical Memorandum, the Rank Probability Score for discharges above a severe threshold is shown for several EPS based fore-

HESSD

5, S688–S696, 2008

Interactive Comment

casts. The way the score is applied in the study, the best score is achieved by an EPS system that in fact systematically predicted low rainfalls, so that flood thresholds, even low ones, were never achieved. As most of the discharge measurements were below the extreme thresholds, the score counted positively for the EPS system analyzed. This is one example in the literature that illustrates that it is therefore important to cross check results with different scores. This is what we have chosen to do in our study. We will make the reasons for this methodological choice more explicit and check again for any redundancies which could be avoided for an easier reading.

Specific comments: R2: "Introduction: I found the introduction very much focussed on different performance criteria. Since it is the aim of the authors to assess the skill of the EFAS with different input data rather than to provide a new indicator for assessing the performance of a forecast, it would be more interesting to report the skills of forecasting systems presented in literature. This would help to put the skills that they computed for EFAS in a more general context. I suggest to specify the results of skill studies that the authors mention on p.291 (l. 3-8) . These could serve as a kind of benchmark for EFAS to which the skills found for EFAS could be compared. This would enable a true assessment of the performances achieved by EFAS."

Answer: As stated on p. 291, l.3-4 and 11-13, the study aims at assessing the skill of EFAS forecasts. It addresses indeed different input data, but also, and not least important, it focuses on a broad assessment using a relatively large number of skill measures (the importance of such assessment is commented in our reply above). Therefore, we believe that introducing the main scores one can find in the literature is essential to understand the choice of skill measures made by the authors, as well as to guide the reader to other practices if s/he needs so. Additionally, as we stated earlier in this reply, regional statistics computed at a specific point are not straightforwardly comparable to those obtained from the statistical analysis performed in our study on a European scale (where several pixels are computed together). One can hardly objectively compare systems with different aims, modelling characteristics, input data and output objec-

HESSD

5, S688–S696, 2008

tives, since forecast verification depends, at different degrees, of all such factors. That explains why we focused in the results reported in the literature concerning the use (and misuse) of skill scores, their strengths and limitations, instead of reporting values of skill scores of other forecasting systems (which would be useless in the scope of the study). The authors will make it clearer in the introduction and, whenever appropriate, specify the score values and main results of the studies referenced.

R2: "In the introduction I would also give some more details on the meteorological input data that were used. What are the skills of these products with respect to predicted rainfall amounts? I think it is necessary to better evaluate the quality of the input data before using them in hydrological forecasting. This is important in order to evaluate the rankings that they establish for the hydrological forecasts using different input data sets."

Answer: The authors fully agree with this comment. Part I of this paper, dealing with the system's development and concept, addresses the main features (temporal and spatial characteristics) of input data (including weather forecasts). If the journal accepts publishing Part I and Part II of this paper, the skill scores for meteorological data will be detailed in Part I and briefly reminded in Part II when contributing for the analysis of the results on the quality of the hydrological forecasts. Otherwise, if the journal excludes Part 1, we will assess this point in this paper.

R2: "p. 293 l. 21 please clarify what the contribution of this paper is compared to the EFAS skill study of Bartholmes et al. 2006 The study of 2006 was a precursor to this paper."

Answer: It was a preliminary study (technical report) and not a scientific paper. It thus allowed to identify the main features that needed some deeper scientific investigation, which was only performed in the study presented in this paper submitted to HESSD. Also, the study of 2006 was based only on a statistical analysis of the year 2005.It concluded on the need of having more flood events to be included in a statistical skill

HESSD

5, S688–S696, 2008

Interactive Comment

assessment than the ones which occurred in Europe during 2005. We will clarify this point in the text.

R2: "p.293 l. 27: using the same kind of meteorological input data any regional or national service could provide medium-range flood forecasts. Hence, provided the same data sets are available (and normally they are), one could argue that national or regional hydro-meteorological services could also provide better medium-range forecasts than EFAS (since they should have more complete data sets for model calibration)."

Answer: Indeed, any national forecasting center running the ECMWF EPS through their forecasting system can also report medium-range flood forecasts, at least within their administrative boundaries. We also agree that by using higher resolution data sets in their systems, they will probably achieve a better quality in their forecasts than the one EFAS can possibly achieve. However, as it is explained in Part I of this paper, and in response to the referees comments from 9th June 2008, there are a number of issues that not all national or regional flood forecasting centres in Europe can currently address. For example, national or regional flood forecasting centres do not necessarily run their models on a catchment-based mode, but rather for specific administrative units, which do not always enclose the whole hydrologic catchment area (this is specifically the case for large and transboundary catchments like the Elbe, the Danube or the Rhine river basin). Therefore, one of the most significant added values of EFAS is exactly the spatial overview over the whole catchment it can provide, including information from neighbouring catchments and upstream forecasted flood situation, which surely national or regional systems would not be able to provide or to obtain elsewhere. Also, the units modelled by local systems can often be very small, so that serious work on downscaling of EPS forecasts would be necessary. Finally, in order to do proper probabilistic forecasting a sufficient number of events need to be present (i.e., to have been observed and/or forecasted by the system). Single catchments may not achieve the sufficient number of occurrences for a flood frequency analysis, which is specially the case when we consider that meteorological models change frequently

HESSD

5, S688–S696, 2008

R2: "p. 296 l.7 I feel that more explanations are needed with respect to ECMWF’s EPS products."

Answer: Reliability diagrams are standard representations in probabilistic forecast skill assessment and can be found in numerous studies in the literature. The authors will add standard references for readers desiring to go more into details on this topic.

R2: "p. 298 l. 4 please explain in more detail what you mean by “climatology”"

Answer: The referee has pointed out an important point. Climatology is in fact a meteorological term that may mean something different for hydrologists. Here, "the climatology" means the "climatological forecast" used as a benchmark, i.e., the forecast associated with a climatological frequency (sample mean frequency of the event computed using long-term statistics). As indicated for a previous comment on the use of meteorological terms, the authors will revise this point to make sure that the definition is clearly stated..

R2: "p. 304 in Fig. 3 you give the absolute numbers of the contingency table. But could you also briefly mention how many HAL and SAL per pixel were observed in average."

Answer: This is described in detail in Part I of this paper where the definition of thresholds are explained. Actually, the thresholds are deduced from the average frequency of events in a pixel. We will make a reference in the Part II paper here to Part I again.

R2: "p. 305 l. 5 I found it surprising that the skill using DWD data is in general smaller than the one obtained with ECMWF data. I was especially surprised by the explanation given by the authors. They claim that the ECMWF resolution is more similar to the JRC-MARS and that due to this, the skill of EFAS forecasts with ECMWF data might be better. Didn’t you do any resampling of the DWD raster to make it compatible with the Lisflood grid? Was there no averaging of the rainfall predictions over each Lisflood cell? Why should the lower resolution product necessarily provide better

HESSD

5, S688–S696, 2008

skills?"

Answer: In fact, there are studies that indicate that the DWD precipitation forecasts tend to have lower scores than the ones from ECMWF (see for instance Pedemonte et al, 2005) It is therefore not surprising that the skill scores of the hydrological forecasts based on DWD data should also be lower. In addition to this, there is the contributing factor of the mismatch in resolutions between the underlying observations density and the forecasting model resolution. For the 2005 & 2006 data, the DWD grid is smaller than the ECMWF grid, meaning that convection is better resolved and higher rainfalls simulated than with ECMWF model. This can then result in higher rainfall totals than sampled by a coarse meteorological station network (like it is in the case of JRC-MARS) and consequently lead to overpredictions as compared to the thresholds. The authors agree that a resampling might be a good way to check this, however, it would be computationally too heavy to perform on a European scale. This point will be made clearer in the paper.

References:

Anctil, F., Michel, C., Perrin, C., Andréassian, V., (2004) A soil moisture index as an auxiliary ANN input for stream flow forecasting. Journal of Hydrology, 286, 155-167.- -

Kalas, M., Ramos M.H., Thielen, J., Babiakova, G. (2008) Evaluation of the medium-range European flood forecasts for the March-April 2006 flood in the Morava River. J. Hydrol. Hydromech., 56, 2008, 2, 116-132.- -

Pappenberger F., K. Scipal, R. Buizza (2008) Hydrological aspects of meteorological verification Atmospheric Science Letters, Volume 9, Issue 2, 43-52 - -

Pedemonte L., M. Corazza, D. Sacchetti, E. Trovatore and A. Buzzi.(2005) VERIFICATION OF LIMITED-AREA MODELS PRECIPITATION FORECASTS DURING THE MAP-SOP. Proceedings ICAM/MAP 2005 Zadar, Croatia, 23rd – 27th May, 2005 - -

HESSD

5, S688–S696, 2008

Roulin, E., Vannitsem, S. (2005): Skill of Medium-Range Hydrological Ensemble Predictions. Journal of Hydrometeorology, Volume 6, Issue 5 (October 2005) pp. 729–744 - -

Younis, J., Ramos, M.H., Thielen, J. (2008) EFAS forecasts for the March-April 2006 flood in the Czech part of the Elbe river basin – a case study. Atmos. Sci. Let. DOI: 10.1002/asl.179, 2008, 7p.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 5, 289, 2008.