Hydrol. Earth Syst. Sci. Discuss., 8, C1098–C1111, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C1098/2011/

Hydrology and Earth System Sciences Discussions

Interactive comment on "An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios" *by* N. Addor et al.

N. Addor et al.

massimiliano.zappa@wsl.ch

Received and published: 22 April 2011

We would like to thank the referee for her valuable comments which clearly contributed to improve both the structure and the content of this manuscript.

The outline of this reply is as follows. Several remarks from the three referees concerned two important aspects of the manuscript: the assessment of the uncertainty sources and the under-sampling. In the first part of this answer, we propose to dedicate two new Sections (2.3 and 3.1) as well as a few new paragraphs to address these concerns. In the second part, we submit our replies to each point raised by the referee.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

Concerning uncertainty sources, we propose to create a Section 2.3 with the following content:

2.3 Uncertainty sources

It is important to stress that the operational ensemble system only quantifies and propagates the atmospheric uncertainty by ingesting atmospheric ensembles from COSMO-LEPS. In other words, the spread of the hydrological ensemble solely reflects the uncertainty associated to the atmospheric boundary conditions. In particular, with the current setting, the spread of the forecast does not account for the uncertainties associated to the formulation of the atmospheric models, to the stations measurements, to the interpolation errors, to the estimation of the hydropower production and to the hydrological and hydraulic modelling. Note that Zappa et al. (2011) used a model configuration similar to the one presented here, but considered several of the mentioned uncertainty sources. They identified in particular that the uncertainty of the hydrological model is about ten times smaller than that stemming from COSMO-LEPS in case of severe flood events.

For the Sihl catchment, a more comprehensive a priori (before the occurrence of the forecast event) assessment of the uncertainty is still necessary. In the present study, we nevertheless quantify the a posteriori (after the occurrence of the event) error stemming from the atmospheric part of the model chain and from the hydraulic/hydrologic part. Therefore, HREF is compared with OBS and the forecasts, to differentiate between two sources of prediction errors.

Comparing HREF to a forecast highlights the first source of error, a divergence between the interpolated meteorological surface observations and the meteorological forecast. When the whole model chain is considered (Fig. 3), the single difference between producing HREF and producing a standard forecast is the type of meteorological data. Interpolated surface observations are used for the former, while a forecast is used for the latter. Hence, if these two datasets correspond, the match between HREF and the

HESSD

8, C1098–C1111, 2011

forecast should be perfect.

Note that the interpolated meteorological observations do not necessarily correctly reflect the true meteorological situation. There are uncertainties, mainly related to measurement errors at the meteorological stations and to the interpolation process. In this study, we assumed that the combined uncertainty of these two effects is usually smaller than that between the forecast and the interpolated observations. Hence, we interpreted differences between HREF and an associated forecast as the divergence between the actual and the forecast meteorological situation, i.e. as an imperfect meteorological forecast.

Comparing HREF to OBS reveals the second source of prediction error. A difference between the two parameters is the consequence of approximations in one or both of the following steps: a) the meteorological measurements and their interpolation, b) the simulation using PREVAH and FLORIS. In a few cases, the interpolated data were clearly erroneous, e.g. because a very local event had been missed by the measuring network. Nevertheless, we consider that measurement and interpolation errors are usually comparatively small. We thus interpret in continuation divergences between HREF and OBS as resulting mainly from hydrological/hydraulic errors.

We propose to add this paragraph before the second paragraph of page 724 (line 8):

The hydropower production is not set by the dam operators themselves, but is determined by a control centre, which considers in particular the previsions of the electricity demand and the market prices. As no access to the details of this procedure was granted for this study, estimations of the hydropower production were used. They were obtained by means of a multiple regression based on a 31-month record of the daily hydropower production. The explanatory variables considered were the hydropower production of the previous day, the day of the week (less electricity is produced during the week-ends), the month of the year and the level of the Sihl Lake.

Finally, we propose to replace the end (starting from line 11) of the first paragraph of

HESSD

8, C1098–C1111, 2011

page 730 by:

We found that in standard conditions, approximations stemming from the regression explain typically half of the absolute error of the lake level forecasts. This error is not represented by the spread among the members. Substantial forecast improvements could be achieved if the planned hydropower production could be integrated in real-time model operations. Usually, the only forecast variable affected by this approximation is the Sihl Lake level. In the rare occasions when dam overflows occur, this error propagates downstream of the dam, but its influence is small in comparison to the volume of water involved in a dam overflow.

Concerning the under-sampling, we propose to reformulate and enhance the two first paragraphs of Section 3 (page 724) and to replace them by the following Sections 3.1 and 3.2:

3 Evaluation of the model chain

3.1 A longer reforecast to cope with the under-sampling of extreme events?

We focused on the flooding of the construction site of the Zurich railway station which would occur for a Sihl discharge of 300m3/s, i.e. with a return period of about T=70 years. Without surprise, none of the events of the 31-month reforecast period exceeded this threshold, as the maximum discharge peaked at 229m3/sec. This of course made the assessment of model's performance for events endangering the construction site delicate.

Intuitively, one could think that this issue can be solved by producing a longer reforecast. A model run over a longer period would generate more examples of intense events. It would hence enable to build robust statistics on flood forecasting and to provide a clear guidance for mitigation measures. But how long should such a reforecast be? Let us assume at least several times T. This would mean two orders of magnitude longer than the reforecast presented in this study. This quick reasoning illustrates

HESSD

8, C1098–C1111, 2011

that producing a longer reforecast is not a straight-forward task and leads to delicate questions such as: Should the HEPS be run in hindcast mode as far back in time as possible, or only for selected events? Is it necessary to run the full ensemble, or could running a reduced number of the ensemble members enable to decrease the computing expenses while preserving most of the original ensemble representativity? How to cope with the fact that since the beginning of the measurements of the Sihl discharge in Zurich in 1938, the 300m3/sec threshold has never been exceeded?

Although in our view these questions deserve further investigation, they will not be answered in this study. Nevertheless, we argue that a longer reforecast would better support evaluation studies of extreme events (Hamill et al., 2004), constitute a useful basis to develop post-processing corrections and enable to design efficient decision support rules or systems for hydrologic applications (Alfieri et al, 2011). However, it is at present unclear which form such a reforecast should take.

3.2 Three perspectives on the reforecast

To cope with under-sampling of extreme events for the present dataset, three complementary perspectives were chosen. First, HEPS skills to forecast low to high discharges were evaluated using several metrics and graphical representations. A large part of this evaluation is based on discharge thresholds: the 75th, 90th and 99th quantiles of the daily maximum distribution estimated from records of hourly measurements from 1974 to 2007 in Zurich (Table 1). They represent a trade-off between low thresholds (e.g. the average discharge) and very high thresholds (e.g. associated with a return period of 100 years or more). The former would lead to an evaluation largely irrelevant for flood forecasting purposes and the latter to weak statistics, given the duration of the present reforecast. Note that the discharges associated to these thresholds (9.10, 21.18 and 73.13 m3/sec) are significantly lower than the threshold considered to decide the flooding of the construction site (300m3/sec). Hence, the metrics computed on the basis of these thresholds are not used to draw definitive conclusions on the model capacity to correctly forecast flooding events. Nevertheless, it is argued that

HESSD

8, C1098–C1111, 2011

these results can highlight those deficiencies in the model chain that may also also affect extreme discharges forecasts.

Note that the text from line 27 page 725 to line 5 page 726 will be removed as it was included in the previous paragraph.

Second, COSMO-7- and COSMO-LEPS-based forecasts for the two most intense events of the study period were analysed and compared. The insights provided by this case-by-case analysis are limited these events did not reach the 300m3/sec threshold. To overcome this, a third perspective was explored. Two scenarios with increased lake level were computed. They led to overspills of the Lake Sihl, resulting in increased discharges in Zurich. These three approaches are described in the next three sub-sections.

We finally propose to add the following sentences to the end of the abstract:

No definitive conclusion on the model chain capacity to forecast flooding events endangering the city of Zurich could be drawn because of the under-sampling of extreme events. Further research on the form of the reforecasts needed to infer on floodings associated to return periods of several decades, centuries, is encouraged.

We now reply to the points raised by the referee:

1 - Can you provide more information on what 2-way feedback you have undertaken with your stakeholders to date (not including the new module – page 739)? This is not entirely clear from the manuscript and would be interesting information to have to place your study in context.

We propose to add the three following paragraphs at the end of Section 1.2 (page 720).

Evaluation of the socio-economic consequences of floods and investigation of protection measures were performed via collaboration with stakeholders concerned by the management of the Sihl discharge. They include (1) the Department of Waste, Water, Energy and Air (AWEL) of Canton Zurich in charge of protection against floods, (2)

HESSD

8, C1098–C1111, 2011

the operators of the Sihl reservoir, (3) the consortium in charge of the construction of the railway station located beneath the Sihl bed, (4) the engineers responsible of the present flood warning system and emergency plan and (5) the company insuring this construction site against flood damage.

Discussions with the stakeholders helped for the comprehension of challenges inherent to the Sihl discharge management. For instance, as we visited the construction site of the railway station, the head of the works confided us that he "would have slept more peacefully" if the level of the Sihl Lake had been a few meters lower during the building period. As we mentioned that to the operators of the dam, we were told that, in contrast, they would have probably slept better if the lake level had been a few meters higher. Despite this expected divergence, both parties agreed that "win-win" situations could be found. In particular, for a high lake level, hydropower production before a heavy precipitation event can decrease both the water losses for the dam operators and the risk of flooding in Zurich.

All the stakeholders showed interest in a system to support decision-making in the Sihl catchment based on the forecasts corresponding to their needs, and which would account for their individual profile (e.g. for their respective economic risks and room to manoeuvre in case of flood risk). In this study, we provide an overview of the form that such a system could take and of the information that would therefore be required.

This paragraph could be added at the end of Section 4.8 (page 740):

Finally, on the basis of our contacts with the stakeholders, we can report that in the Sihl catchment, several actions rely on observations only. For example, the emergency regulation of the Sihl Lake depends on the actual lake level and on the rate of its actual increase. We argue that giving less weight to the observations of the parameters most relevant from a flooding perspective, and more to their forecasts could lead to a sounder management (less forced, uncontrolled water releases). However, reliable forecasts are therefore needed. When we presented the simulations of the August

HESSD

8, C1098–C1111, 2011

2007 event to the AWEL, we showed that, in the associated scenario, the forecasts anticipated the Lake overspill which hence could have been reduced by a higher hydropower production. However, this positive element was clearly occulted by the miss of the event by the model chain. A better performing HEPS would certainly enhance the stakeholders' confidence in the system, and encourage them to give more weight to the forecasts when making decisions.

2 - How are the 16 representative members selected? Do you have any evidence regarding to what extent these 16 preserve the information contained in the 102? If this is contained in a reference somewhere then please make this clearer. (page 721, line 22). What are the implications for your results of the underdispersivity noted (page 732)? The discussion of COSMO-LEPS could do with some more depth here. Is any of the underdispersivity transferred from the original ensemble?

We propose to replace the sentence starting on line 20, page 721 by:

The two youngest runs of this model are combined to form a super-ensemble of 102 members. To reduce the computational burden that the downscaling of the full super-ensemble would represent, only 16 "representative members" are downscaled. To select them, the super-ensemble is distributed in 16 groups of different population with help of a cluster analysis. The geopotential height, specific humidity and horizontal wind are considered to identify similar weather patterns and establish the groups (Marsigli et al., 2005). A single representative member is finally selected from each cluster. It is defined as the member with the smallest ratio between the average distance from its cluster members and the distance from the remaining members (Molteni et al., 2001).

The following paragraph could be inserted between the present first and second paragraph (on line 4) of page 722:

Jaun et al., 2008 investigated the influence on the forecast skill of a decrease from 51 to 10 ensemble members for the August 2005 flood event in the Swiss part of the Rhine

HESSD

8, C1098–C1111, 2011

catchment. They pointed out a loss of information, denoted by an overall decrease of Brier skill score (BSS) for precipitation and runoff forecasts. By plotting the BSS for the ensemble sizes of 1 to 51 members, they observed an improvement of the BSS with the ensemble size, but emphasized that the increase of ensemble members beyond 15 had comparatively little impact on the BSS. This led them to conclude that, for the studied event, working with a reduced ensemble constituted a reasonable trade-off between the forecast skill and the computational resources demand.

We propose to replace the two last sentences of the second paragraph of page 732 (lines 18-21) by the following paragraph:

Marsigli et al. (2008) found that the percentage of outliers for 66-h COSMO-LEPS precipitation forecasts reached around 30% (about 2.5 times the theoretical percentage) when considering the maximum values over boxes of 1.0x1.0 degrees. They furthermore revealed that the 51-member global EPS produces even more over-confident 66-h forecasts, with a percentage of outliers of around 40% (about 10 times the theoretical percentage). Hence, part of the underdispersion affecting COSMO-LEPS precipitation forecasts most probably stems from the original EPS. This motivated the introduction of the super-ensemble (Marsigli et al., 2005). However, on the basis of our results, it appears that it was not sufficient to solve the underdispersion issue for the Sihl catchment. From an end-user point of view, a crucial information conveyed by probabilistic forecast is the confidence of the forecast, as represented in particular by the amount of spread between the members. Unfortunately, the tendency to underdispersion means that the correspondence between a narrow spaghetti plot and a confident forecast cannot be guaranteed. Note finally that the overconfidence of COSMO-LEPS-based flow forecasts has also been reported by Renner et al. (2009).

3 - Can you further discuss the implications of using proxy observations (HREF) in lieu of actual observations (e.g. as per discussion in Pappenberger et al (2008)). There are advantages and disadvantages which could be further discussed, especially if you focus on the end product of the operational rules of Lake Sihl. (Page 723, line 24). Have

HESSD

8, C1098–C1111, 2011

you statistics (rather than just the graphs, fig 8 etc) to demonstrate the relationship?

We are not sure about the interpretation of this third point. In fact, in our manuscript, HREF has not really been used “in lieu” of actual observations (OBS), but rather “in addition” to OBS, for selected parameters or graphs.

We agree that if we focus on the end-product, it is decisive for the model chain to forecast correctly the actual value (OBS) of e.g. the level of the Sihl Lake or the discharge of the Sihl River. That is why a large part of our manuscript deals, in a quantitative way, with the correspondence between the forecasts and OBS (see the five last columns of Tables 2 and 3, Fig. 4, the first, third and fourth column of Fig. 5, Figs. 6 to 9).

However, another objective of our study was to discriminate between errors originating from atmospheric forecasts and those stemming from the hydrological/hydraulic components of the model chain (see the new Section 2.3). Therefore, in addition to comparing the forecasts to OBS, they have also been compared to HREF (see the first column of Tables 2 and 3, the second column of Fig. 5, Fig. 6 and Figs. 8 and 9).

4 - The discussion of F and FAR might benefit from more contextual discussion about the requirement for a long time series. For what length of time series, number of events etc. do these indicators stabilise. The crux is – when can they be trusted? (Page 727...)

Fig. 6 has been completed (see below) and we propose to add the following sentence on page 727 line 22:

To depict the uncertainty related to under-sampling and affecting H and FAR, confidence intervals have been computed using the same bootstrapping procedure as for BSS.

Accordingly, the caption of Fig. 6 could be:

False alarm ratio vs. hit rate for the daily maximum Sihl discharge in Zurich for the discharge thresholds Q0.75 (left) and Q0.9 (right). The lines refer to COSMO-LEPS forecasts and their circles correspond to the probability thresholds 25%, 50% and 75%

HESSD

8, C1098–C1111, 2011

(from right to left). The stars and diamonds indicate COSMO-7 and HREF performance, respectively. The raw scores are exhibited by the symbols (circles, stars and diamonds), while the extremities of the confidence intervals consist of the 5th and 95th quantiles derived by bootstrapping.

As the observations presented in Section 4.3 are still valid after the computation of the confidence intervals, only minor modifications to this Section were done. In particular, it is proposed to rename the "ROC curves" H-FAR curves in the whole manuscript (see the point 11 raised by the Anonymous referee). The most important change is the addition of a third paragraph discussing the gap between Q0.75 and Q0.9 and the threshold considered for decision-making.

It is hence proposed to reword Section 4.3 as follows:

4.3 H-FAR skill

Fig. 6 indicates that forecast skills in terms of H and FAR tend to decrease with lead time for COSMO-LEPS and COSMO-7. Given the comparatively good scores of HREF, this emphasizes that correct atmospheric forecasts are essential for trustworthy discharge forecasts. The diamonds referring to COSMO-7 are located close to COSMO-LEPS H-FAR curves for the same lead times, which suggests comparable performance. However, probabilistic forecasts allow end-users to optimize the choice of their warning thresholds according to their economic profile (e.g. Roulin, 2007), which is not possible when using deterministic forecasts.

By increasing the discharge threshold from Q0.75 to Q0.9, a performance decrease for all lead times and both models is observed. The scores for the threshold Q0.99 are not shown because of their very high sampling uncertainty. For the Q0.90 threshold, LT2 to LT5 forecasts produce false alarms at a preoccupying rate, as false alarms account for roughly 50 to 70% of the warnings. Although end-users are usually more concerned about missed events than by false alarms, these high FAR should not be neglected or trivialized. Unnecessary preventive drawdowns represent significant mon-

HESSD

8, C1098–C1111, 2011

etary losses for the dam operators, and successive false alarms could undermine end-users' confidence in the flood forecasting system. Furthermore, the almost vertical inclination of the COSMO-LEPS H-FAR curves implies that increasing the probability threshold barely reduces this high FAR, but largely penalises the forecasts in terms of H. Note that for the probability threshold 50% (indicated by the central circle on the H-FAR curves), mid-term (LT3 to LT5) forecasts perform poorly when capturing observed events (H âĹij0.35).

Q0.9 corresponds to a discharge of about 21 m3/sec, which is one order of magnitude smaller than the warning threshold considered to evacuate the railway station construction site. Hence, we cannot extrapolate H-FAR results for Q0.9 to discharges endangering the infrastructure. Nevertheless, we do not expect the scores to improve with increasing thresholds, and we consider that the poor model performance in terms of H and FAR constitutes a real issue for flood mitigation.

5 - The discussion of uncertainty in hydropower would surely be more useful earlier (page 730), alongside a clearer discussion of all uncertainties in your method/data.

We accounted for this comment when formulating the new Section 2.3 and the two paragraphs dedicated to uncertainty sources (see the first part of this reply).

Once again, we would like to thank the referee for her inspiring comments. The typos she mentioned will be corrected and a new linguistic revision of the whole manuscript will be performed.

On behalf of all co-authors | N. Addor

References

Alfieri, L., Velasco, D., and Thielen, J.: Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events, Adv. Geosci., 29, 69–75, 2011.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schär, C.: A probabilistic view on the

HESSD

8, C1098–C1111, 2011

August 2005 floods in the upper Rhine catchment, Nat. Hazards Earth Syst. Sci., 8, 281–291, doi:10.5194/nhess-8-281-2008, 2008.

Hamill, T., M. Whttaker, J. S., Wei, X.: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts, Mon. Weather Rev., 132, 1434-1447, 2004.

Marsigli, C., Boccanera, F., Montani, A., and Paccagnella, T.: The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification, Nonlin. Processes Geo- phys., 12, 527–536, doi:10.5194/npg-12-527-2005, 2005.

Marsigli, C., Montani, A., and Paccangnella, T.: A spatial verification method applied to the evaluation of high-resolution ensemble forecasts, Meteorol. Appl., 15, 125–143, doi:10.1002/ met.65, 2008.

Molteni, F., Buizza, R., Marsigli, C., Montani, A., Nerozzi, F., and Paccagnella, T.: A strategy for high-resolution ensemble prediction. I: Definition of representative members and global- model experiments, Q. J. Roy. Meteor. Soc., 127, 2069–2094, doi:10.1002/qj.49712757612, 2001.

Renner, M., Werner, M. G. F., Rademacher, S., and Sprokkereef, E.: Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376, 463–475, doi:10.1016/j.jhydrol.2009.07. 059, 2009.

Zappa, M., Jaun, M., Germann, U., Walser, A., Fundel, F., 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. Atmospheric Research 100, 246–262.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 8, 715, 2011.

HESSD
8, C1098–C1111, 2011

Interactive Comment

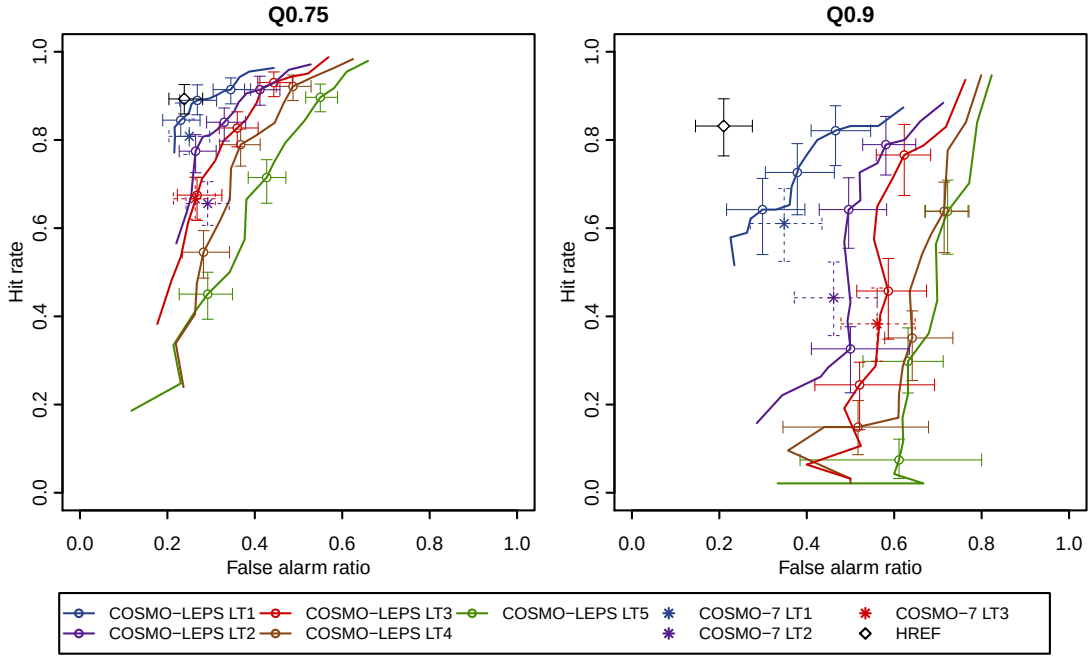

Fig. 1.

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