Hydrol. Earth Syst. Sci., 15, 2327–2347, 2011
www.hydrol-earth-syst-sci.net/15/2327/2011/
doi:10.5194/hess-15-2327-2011
© Author(s) 2011. CC Attribution 3.0 License.

Hydrology and
Earth System
Sciences

An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios

N. Addor[1,2,*], **S. Jaun**[1], **F. Fundel**[1], **and M. Zappa**[1]

[1]WSL Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf, Switzerland
[2]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[*]now at: Department of Geography, University of Zurich, Zurich, Switzerland

Received: 30 September 2010 – Published in Hydrol. Earth Syst. Sci. Discuss.: 20 January 2011
Revised: 20 May 2011 – Accepted: 4 July 2011 – Published: 20 July 2011

Abstract. The Sihl River flows through Zurich, Switzerland's most populated city, for which it represents the largest flood threat. To anticipate extreme discharge events and provide decision support in case of flood risk, a hydrometeorological ensemble prediction system (HEPS) was launched operationally in 2008. This model chain relies on limited-area atmospheric forecasts provided by the deterministic model COSMO-7 and the probabilistic model COSMO-LEPS. These atmospheric forecasts are used to force a semi-distributed hydrological model (PREVAH), coupled to a hydraulic model (FLORIS). The resulting hydrological forecasts are eventually communicated to the stakeholders involved in the Sihl discharge management. This fully operational setting provides a real framework with which to compare the potential of deterministic and probabilistic discharge forecasts for flood mitigation.

To study the suitability of HEPS for small-scale basins and to quantify the added-value conveyed by the probability information, a reforecast was made for the period June 2007 to December 2009 for the Sihl catchment (336 km^2). Several metrics support the conclusion that the performance gain can be of up to 2 days lead time for the catchment considered. Brier skill scores show that overall COSMO-LEPS-based hydrological forecasts outperforms their COSMO-7-based counterparts for all the lead times and event intensities considered. The small size of the Sihl catchment does not prevent skillful discharge forecasts, but makes them particularly dependent on correct precipitation forecasts, as shown by comparisons with a reference run driven by observed meteorological parameters. Our evaluation stresses that the capacity of the model to provide confident and reliable mid-term probability forecasts for high discharges is limited. The two most intense events of the study period are investigated utilising a novel graphical representation of probability forecasts, and are used to generate high discharge scenarios. They highlight challenges for making decisions on the basis of hydrological predictions, and indicate the need for a tool to be used in addition to forecasts to compare the different mitigation actions possible in the Sihl catchment. 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 floods associated to return periods of several decades, centuries, is encouraged.

1 Introduction

1.1 Decision-making based on atmospheric and hydrological forecasts

To effectively anticipate and mitigate weather-related impacts, strategies that take into account climatological records or meteorological forecasts have been developed in recent decades. Scientific studies published in the 1960s–1970s already showed that an efficient use of weather and climate information could provide an added-value in diverse fields and

Correspondence to: N. Addor
(nans.addor@geo.uzh.ch)

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

decision-making situations (e.g. the review by Katz and Murphy, 1997). Novel opportunities and challenges were provided in the 1990s by the introduction of global atmospheric ensemble prediction systems (EPS, e.g. Molteni et al., 1996) and more recently by their downscaled limited-area derivatives (e.g. COSMO-LEPS, Marsigli et al., 2005). Ensembles are composed of several members, starting from slightly perturbed initial conditions, and aim to reflect the predictability of atmospheric conditions through the amount of spread among their members. Reliable EPS enable the estimation of the probability of local weather events, and are expected to deliver a more trustworthy basis for quantifying risk and providing early warnings than their deterministic counterparts. Among others, Richardson (2000) and Zhu et al. (2002) have investigated their benefits in the domain of decision-making using a simple cost-loss ratio decision model. They reported that, in comparison with deterministic forecasts, probabilistic forecasts offer an added-value for a wider range of end-users and present a higher economic value for the majority of end-users and lead-times. However, the ability of the standard two-action, two-event cost-loss ratio scheme to effectively assist with real decision-making situations is disputed (e.g. Murphy, 1985).

When coupled to a hydrological model, an EPS forms a hydrological ensemble prediction system (HEPS). HEPS have developed rapidly in the last few years (see the review by Cloke and Pappenberger, 2009). They have been adopted by several flood forecast centres and are, for example, routinely run by the European Flood Alert System (EFAS) of the European Commission Joint Research Centre (Thielen et al., 2009a). An evaluation of two years of EFAS forecasts for Europe suggests the results are promising, especially when accounting for forecast persistence (Bartholmes et al., 2009). On a smaller scale, the Swiss Federal Office for the Environment (FOEN) operationally runs deterministic and probabilistic hydrological forecasts for the Swiss part of the Rhine basin (Zappa et al., 2008). In the framework of the Mesoscale Alpine Programme (MAP), a demonstration of probabilistic hydrological and atmospheric simulation of flood events in the Alpine region (D-PHASE) was performed. The feasibility of a real-time hydrological forecast system that combines radar-based, high-resolution and ensemble hydrological forecasts was shown, and examples illustrating the usefulness of the probability information were provided (Zappa et al., 2008; Rotach et al., 2009). End-user feedback has so far been positive.

Considerable efforts have been made to demonstrate and quantify the added-value provided by HEPS, as illustrated by the following examples. Verbunt et al. (2007) analysed qualitatively two severe discharge events, in the upper Rhine basin and in central Europe. These were missed by deterministic runoff predictions but adequately forecast by probabilistic models. The authors report good probabilistic forecast guidance up to 48 h lead time for the two investigated cases. For the August 2005 flood event in the upper Rhine basin, Jaun et al. (2008) highlighted that forecast uncertainty, as reflected by ensemble dispersion, provides additional guidance in comparison to deterministic forecasts. This is in particular supported by higher Brier skill scores. Velázquez et al. (2009) compared, for a rainfall event in Quebec, the continuous ranked probability score of a hydrological ensemble with the absolute error of a deterministic forecast and concluded that the probability information led to a performance gain. First attempts to evaluate probabilistic discharge forecasts from an economic perspective (Roulin, 2007) relied on a cost-loss ratio-based decision model and showed that hydrological ensemble predictions have greater skills than deterministic ones. Laio and Tamea (2007) proposed new tools for economic evaluation of probability discharge forecasts, but emphasize that the choice of the therefore necessary cost-loss functions is subjective and may be disputed. Reggiani et al. (2009) suggested a stirring approach consisting of the combination of calibrated probabilistic forecasts to cost-loss functions to estimate economic risk.

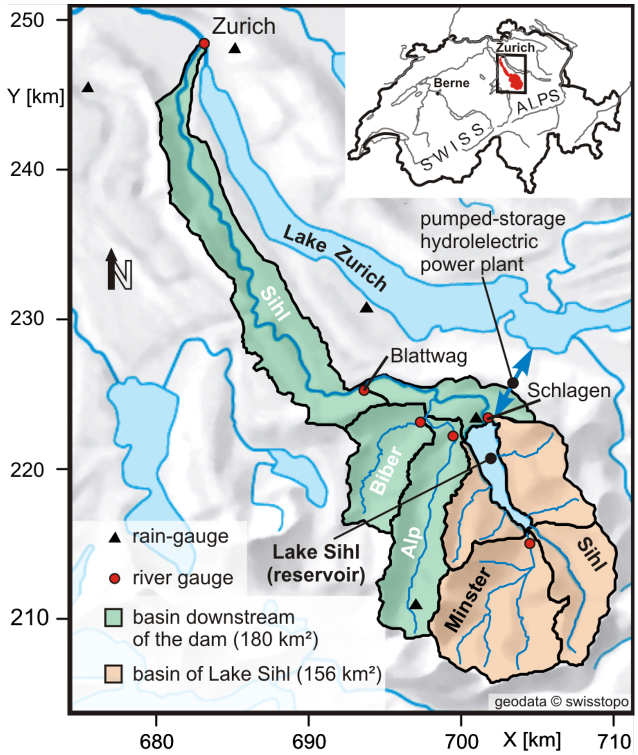

Fig. 1. Map of the Sihl catchment. The nine sub-catchments (including the one formed by the Lake Sihl) and the available measuring stations are shown. Courtesy of J. Schwanbeck, University of Bern.

1.2 Discharge monitoring and flood mitigation in the Sihl catchment

The Sihl catchment originates in the Swiss Alps (Fig. 1) and drains basins which are particularly prone to flash floods.

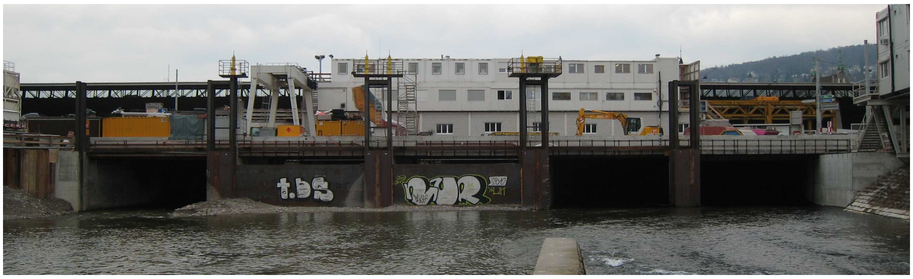

Fig. 2. The Sihl River flows beneath Zurich central railway station. Two of the five channels are sealed to provide dry conditions for the construction site located under the river level. The mean discharge on the day of the picture was 8.92 $m^3 s^{-1}$. Photo courtesy of A. Badoux, WSL.

During wintertime snow accumulates in the headwaters, while snow melt governs runoff generation in late spring and early summer. The Sihl River flows through Zurich, Switzerland's largest city. Shortly before joining the Limmat River, the Sihl flows beneath the main railway station of Zurich (Fig. 2) located in the city centre.

Although this area was comparatively little affected by the devastating floods of August 2005 (Bezzola and Hegg, 2007, 2008; Jaun et al., 2008), they prompted a series of studies assessing the flood risk of the catchment (Schwanbeck et al., 2007). Floods are especially threatening during the construction period of a new underground railway station, located below the Sihl River bed (Bruen et al., 2010). The Sihl passes through the construction site by five culverts, two of which are sealed alternately for the duration of the project (2008–2011, Fig. 2). This provides dry construction areas, but therefore reduces the section available for the river by around 40 %.

To cope with the resulting increased flood risk, the Department of Waste, Water, Energy and Air (AWEL) of Canton Zurich requested the Swiss Federal Railway (SBB) to organise a panel of experts. This panel is in charge of monitoring the Sihl discharge, of representing the interests of the stakeholders concerned by the river, and of setting up an emergency procedure to mitigate flooding events. A first mitigation measure is the preventive controlled water release (drawdown) from the Lake Sihl, which collects the waters from a 156 km^2 large headwater. This reservoir is operated by a private company for hydropower production. The water used therefore is not released into the Sihl River, but is diverted through a penstock to a hydropower station and flows into the Lake Zurich (Fig. 1). In contrast, for a preventive drawdown, three gates located at the top of the dam are gradually lowered and water is directly released into the Sihl River, without passing through the power plant, i.e. without producing electrical energy. The water is thus lost from the point of view of the dam operators, but it enables to increase the buffering capacity of the lake. Secondly, should the Sihl discharge exceed 300 $m^3 s^{-1}$, the gates sealing the two channels beneath the main railway station can be opened, giving the river bed its full capacity. This would result in the inundation of the construction site, but it would reduce the risk of flooding the areas around the Zurich main railway station. To improve decision support for the panel of experts, the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) was mandated to implement an ensemble flood forecasting system. This constitutes the corner stone of this study. More details on the Sihl catchment, the flood-warning procedures and the mitigation measures can be found in Badoux et al. (2010).

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

Discussions with the stakeholders helped us understand the 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 Lake Sihl 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 profiles (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.

1.3 Scope of the present paper

This study concentrates on the evaluation of the operational streamflow forecasting system of the Sihl catchment for a 31-month reforecast period. The evaluation was in particular tailored to cope with the weaknesses identified in the review paper by Cloke and Pappenberger (2009) and strives to achieve the following objectives: (1) to compare quantitatively probabilistic and deterministic hydrological forecasts to assess whether the former has added-value over the latter, (2) to discriminate between errors originating from atmospheric forecasts and those stemming from the hydrological/hydraulic components of the model chain, (3) to analyse forecasts reliability and eventual under- or overforecasting bias, under- or overdispersion of the ensemble members, proneness to false-alarms as well as forecasts ability to capture observed events, (4) to obtain some insights into the performance of the flood prediction chain by analysing the two most severe events during the study period (one of which being a forecast failure), and (5) to investigate extreme discharge scenarios.

This paper furthermore differs from similar studies in the nature of the catchment investigated. It has a total area of 336 km^2, which is considerably smaller than most catchments, and even sub-catchments, referred to in the current literature on ensemble streamflow forecasting (e.g. Dietrich et al., 2008; Jaun and Ahrens, 2009; Thielen et al., 2009b; Reggiani et al., 2009). Earlier studies have shown that forecast skill depends on temporal and spatial scales. For instance, the current state of knowledge for larger basins suggests that the skill of ensemble prediction systems improves with increasing catchment size (Renner et al., 2009). Furthermore, forecast uncertainty is reported to decline with increasing catchment size (Jaun et al., 2008). The usefulness of such systems in small mesoscale areas has not yet been investigated. We therefore concentrate on the skill of an operational HEPS for a comparatively small catchment. Concerning the temporal scales, while the focus of this study is on high flows (spanning over a few days), we additionnaly assess several aspects of the performance of the model for average discharge situations (on the basis of the entire 31-month reforecast period). Moreover, the reservoir lake required the implementation of a module to account for the consequences of possible lake drawdowns, overflow and hydropower production on the discharge in Zurich. Finally, in this basin, the correct assessment of events leading to snow accumulation and snow melt is crucial for obtaining correct forecasts from October to May. The pre-alpine topography of the region and the presence of sub-catchments prone to flash floods triggered by summer thunderstorms complicate correct meteorological and hydrological modelling.

This paper also explores how flood mitigation measures could be triggered on the basis of the presented streamflow forecasts. The operational setting of the Sihl catchment enables the illustration of the complexity of such a decision process involving imperfect forecasts (Bruen et al., 2010).

2 The hydrological ensemble prediction system

2.1 Probabilistic and deterministic atmospheric models

As the operational hydrological forecasts for the Sihl catchment are not systematically archived, a reforecast from June 2007 to December 2009 was completed to proceed with model evaluation. Runs in hindcast mode were issued from the November 2009 version of the model, using operationally available information only. The prediction chain is sketched out in Fig. 3 and described below.

Probabilistic atmospheric forecasts are based on the global Ensemble Prediction System (EPS, Molteni et al., 1996) of the European Centre for Medium-Range Weather Forecasts (ECMWF), which is issued twice daily. 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 populations 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). These 16 members are dynamically downscaled on a daily basis from their original ∼50 km horizontal resolution to a ∼10 km resolution (∼7 km since December 2009). This is performed by the Limited-area Ensemble Prediction System developed and run by the COnsortium for Small-scale MOdelling (COSMO-LEPS, Molteni et al., 2001). COSMO-LEPS relies on the non-hydrostatic COSMO model (Steppeler et al., 2003) run with initial and boundary conditions provided by the representative members. Hydrological forecasts for the Rhine catchment have been shown to improve after this dynamic downscaling (Renner et al., 2009).

Jaun et al. (2008) investigated the influence on the forecast skill of a decrease from 51 to 10 dynamically downscaled ensemble members. They therefore focussed on the August 2005 flood event in the Swiss part of the Rhine 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

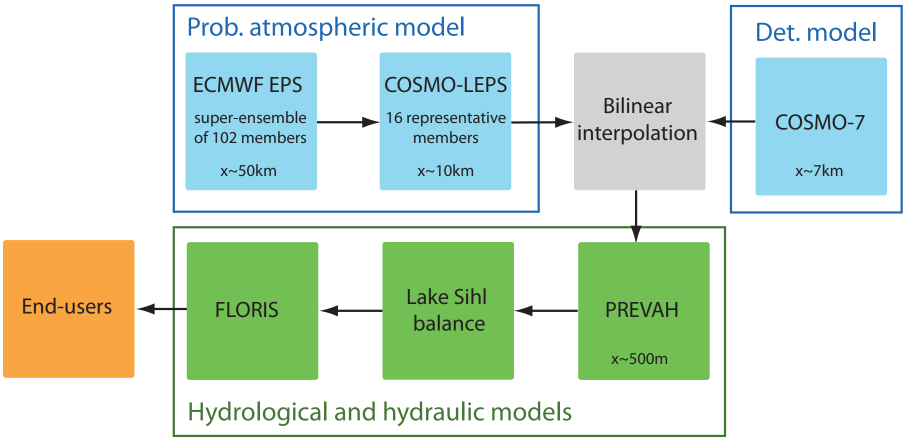

Fig. 3. Flowchart of the prediction chain, illustrating in particular that probability (prob.) and deterministic (det.) atmospheric forecasts are used to force the hydrological model. The horizontal grid spacing of the models is denoted by x and indicates that atmospheric forecasts are downscaled throughout the HEPS. A detailed description of the model chain can be found in Sects. 2.1 and 2.2.

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 demand on computational resources.

The lead time of COSMO-LEPS is 132 h, with three-hourly output intervals. Forecasts are initialized at 12:00 UTC and are completed and delivered approximately 10 h later. As the hydrological model requires initialization at 00:00, the first 12 h of the atmospheric ensemble forecasts are disregarded and 120 h of hydrological forecasts are computed. This cutoff is consistent with the temporal data availability in operational mode. Note that, while errors in streamflow forecasts have multiple sources, the present ensemble principally aims to capture and cascade the uncertainty arising from boundary atmospheric conditions, as they are commonly regarded as the most important error factor in hydrological forecasts. To account for some of the model uncertainty, the convection scheme is randomly chosen at each COSMO-LEPS integration (Marsigli et al., 2005). Similarly, the value of two model parameters (the maximal turbulent length scale and the length scale of thermal surface patterns) is randomly chosen from a set of two reasonable values for each variable.

Deterministic atmospheric forecasts are obtained from the operational model COSMO-7, the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) implementation of COSMO model. COSMO-7 is nested in the ECMWF deterministic global model. Its horizontal grid spacing is of $\sim$7 km and it offers a total time horizon of 72 h. In contrast to COSMO-LEPS, no random selection of a parametrization or a parameter value is applied to reflect model uncertainty. During the study period, COSMO-7 forecasts were issued twice a day (at 00:00 UTC and 12:00 UTC). However, only the forecasts from the 00:00 UTC run are here considered. As mentioned for COSMO-LEPS, PREVAH requires initialization at 00:00. This could favour COSMO-7, as the first 12 h of COSMO-LEPS are disregarded but the full COSMO-7 forecast is considered. However, despite truncating the first 12 h of COSMO-LEPS, this model performs better than COSMO-7 for the large majority of the metrics and events considered, as shown in Sect. 4. Note that COSMO-7 driven forecasts are faster computed than the ones relying on COSMO-LEPS, and are available as early as 04:00 UTC.

Forecasts of temperature, precipitation, wind, relative humidity, sunshine duration and global radiation are downscaled for both atmospheric models to a resolution of 500 m to meet the grid-size requirements of the hydrological model, as described in Jaun et al. (2008).

2.2 Hydrological and hydraulic models

The downscaled atmospheric forecasts are used to force the semi-distributed hydrological modelling system PREVAH (PREecipitation-Runoff-EVApotranspiration HRU Model). PREVAH is a conceptual hydrological model and clusters raster grids of similar hydrological properties into hydrologic response units (HRU, Gurtz et al., 1999). For the Sihl catchment, one HRU averages about 7 raster cells of $500 \times 500\,\text{m}^2$. Details on PREVAH input data, structure, parametrisation and tools can be found in Viviroli et al. (2009b).

The hydrological model calibration and evaluation for the Sihl catchment were performed by Schwanbeck et al. (2007), with the catchment split into nine sub-catchments (Fig. 1). This discretisation enables in particular a simple representation of water management of the Sihl reservoir.

The sub-catchments of Alp, Biber, Minster and the sub-catchment downstream of the gauge Blattwag were calibrated on the basis of the observed runoff time series. The parameters for the other sub-basins were regionalised on the basis of the five most similar calibrated catchments out of a database of 140 successfully calibrated Swiss catchments (Viviroli et al., 2009a,c). The chosen calibration method is an intermediate solution between a flood-oriented and average discharge-oriented optimization. Validation revealed a tendency towards volume overestimation, but an overall satisfying peak discharge representation (Schwanbeck et al., 2007).

Initial conditions for PREVAH are provided by a hydrological reference simulation (HREF) driven by interpolated observations from weather stations (see Fig. 1 for the location of the rain-gauges). Note that HREF was also used to identify the origin of forecasts errors, as presented in Sect. 2.3.

PREVAH forecasts were combined with observations of the level of the Lake Sihl for a sound water balance of the artificial lake. This module accounts principally for (1) the estimation of the hydropower production, (2) the eventual triggering of the dam emergency regulation (water is released from the lake into the Sihl if its level rises more than two centimetres within 30 min) and (3) water overspills if the lake operation limit (889.34 m a.s.l.) is exceeded. See Badoux et al. (2010) for more details on dam regulation.

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 Lake Sihl.

Because of the elongated shape of the basin between Blattwag and Zurich (Fig. 1), a hydraulic model was used to propagate the flood wave. Routing is carried out by the hydraulic model FLORIS, a commercial 1-D simulation program developed in the 1990s by the Laboratory of Hydraulics, Hydrology and Glaciology (VAW) of the ETH Zurich. FLORIS computes possible Lake Sihl overflows and delivers forecasts of the timing and discharge of the flood wave originating from the PREVAH sub-catchments, combined to eventual water release from the Lake Sihl.

2.3 Uncertainty sources affecting the HEPS output

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.

The performance of the system was assessed for its current operational setting. Hence, although calibrated COSMO-LEPS precipitation forecasts exist (Fundel et al., 2010), they were not used. No calibration (e.g. Reggiani et al., 2009) or correction for underdispersion or bias was applied to the output of the atmospheric forecasts before their use in PREVAH. Similarly, no bias correction or statistical post-processing of the hydrologic and hydraulic ensemble forecasts was done. 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 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 errors, a divergence between the interpolated meteorological surface observations and the meteorological forecast. Let us remind that the single element that differentiates HREF from a standard discharge 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, HREF should match the forecast.

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 a 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 errors. 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 assumed that measurement and interpolation errors are usually comparatively small. We thus interpret in continuation divergences

Table 1. Thresholds for the Sihl discharge in Zurich considered for the evaluation of the model chain.

	Quantile	Discharge ($m^3\,s^{-1}$)	Average frequency
$Q0.75$	75th	9.10	Every four days
$Q0.9$	90th	21.18	Three times a month
$Q0.99$	99th	73.13	∼ every three months

between HREF and OBS as resulting mainly from hydrological/hydraulic errors.

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 300 $m^3\,s^{-1}$, i.e. with a return period of about $T = 70$ years. None of the events of the 31-month reforecast period exceeded this threshold, as the maximum discharge peaked at 229 $m^3\,s^{-1}$. This of course made the assessment of the model 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 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 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 construction of the Sihl dam in 1937, the 300 $m^3\,s^{-1}$ threshold has never been exceeded?

Although in our view these questions deserve further investigation, they will not be answered in this study. We nevertheless 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). These quantiles 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 $m^3\,s^{-1}$) are significantly lower than the threshold considered to decide the flooding of the construction site (300 $m^3\,s^{-1}$). 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 these results can highlight those deficiencies in the model chain that may also affect extreme discharge forecasts.

Second, COSMO-7- and COSMO-LEPS-based hydrological 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 as neither of these events reached the 300 $m^3\,s^{-1}$ 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 subsections.

3.3 Evaluation of low to high discharge forecasts

For the evaluation of the year-round model performance, the Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) and the volume error (VOL, as formulated in Zappa and Kan, 2007) were selected because they enable fast comparison with other studies due to their widespread use in the evaluation of hydrological models. The mean absolute error MAE (Wilks, 2006) was also chosen as it is easily interpretable and enables the contribution of some error sources within the forecast chain to be assessed quantitatively. As these three indices are designed for deterministic forecasts, COSMO-LEPS forecasts were reduced to their median to be evaluated. Note that a large part of the ensemble information is thereby disregarded and hence these scores do not capture the information content of the full ensemble. Attention was focused on the level of the Lake Sihl and on the Sihl discharge in Zurich, as these two variables are of most interest to the end-users.

For the computation of VOL, hourly values were used. In contrast, for NSE, MAE and all the scores discussed below, evaluations were based on daily maxima. Different lead times from 1 day (1–24 h) to 5 days (97–120 h), referred to as LT1 to LT5 subsequently, were considered. For COSMO-7, LT1 to LT3 were assessed while for COSMO-LEPS, LT1 to LT5 were evaluated to take into account the respective time horizons of the models.

To further compare the performance of deterministic and probabilistic forecasts, the Brier score (BS, Eq. 1) was chosen (e.g. Wilks, 2006). This score can be seen as a mean squared error of probability forecasts, and has the advantage that it can be applied to both deterministic and probabilistic forecasts, without requiring the transformation of the ensemble forecast into a deterministic one (e.g. by considering the median only). While a ranked probability score would enable to assess the overall quality of the ensemble, the BS permits to focus on specific warnings and thresholds meaningful for this case study. BS reads:

$$\mathrm{BS} = \frac{1}{n}\sum_{d=1}^{n} (o_\mathrm{d} - y_\mathrm{d})^2 \tag{1}$$

where n is the number of days of the reforecast. o_d (resp. y_d) indicates whether the daily maximum of the observation (resp. of COSMO-7) exceeded the threshold considered (1 = yes, 0 = no). For COSMO-LEPS, y_d is the probability of threshold exceedance. In this study, such probabilities were computed as the number of ensemble members exceeding the threshold divided by the total number of members (16, i.e. ensemble members were not weighted). Forecasts show a null BS if they are perfect and a positive BS otherwise. The combination of the obtained BS with the BS associated to a systematic forecast of the climatological frequency of the event (BS_ref) and with the BS of a perfect forecast ($\mathrm{BS}_\mathrm{perf} = 0$) yields to the Brier skill score (BSS) which reads:

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_\mathrm{ref}}{\mathrm{BS}_\mathrm{perf} - \mathrm{BS}_\mathrm{ref}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_\mathrm{ref}}. \tag{2}$$

A BSS of 1 designates perfect forecasts, while positive BSS correspond to forecasts with more skill than the reference. As we are interested in investigating the actual and not the potential model performance, the negative BSS bias associated with small ensemble size was not removed (Weigel et al., 2007). To evaluate the influence of the limited number of high intensity events on BSS, confidence intervals were derived by bootstrapping (Efron, 1992). 500 random samples of 940 daily maxima pairs of forecast-observation were drawn with replacement from the 940 days of the study period. The BSS was then computed for each bootstrap sample, enabling an estimation of a confidence interval formed by the 5 % and 95 % quantiles. To improve the estimation of the sampling uncertainty, block bootstrapping could be considered to account for temporal dependency (see Lahiri, 2003). The BSS was used to estimate the skill of each of the forecasts in comparison to a reference forecast. Using COSMO-7 as a reference to compute COSMO-LEPS BSS would have been a direct way to assess the added-value of the probabilistic forecasts in comparison to the deterministic ones. This assessment is however also possible with the chosen reference, and permits in addition an individual evaluation of the two models. To analyse how much gain the COSMO-LEPS ensemble brings to the hydrologic ensembles, one could generate hydrologic ensembles based on climatological forcing inputs using the same hydrologic model chain and same initial conditions (see Demargne et al., 2010, for such analysis).

When using a probability forecast, a common way to decide whether or not to issue a warning is based on threshold exceedance. This requires the definition of a probability threshold P (e.g. 60 %) and a weather or hydrological threshold Q (e.g. a discharge of 300 $\mathrm{m}^3\,\mathrm{s}^{-1}$ in Zurich). If the forecast probability of exceeding Q is greater than P, a decision to implement protection measures may be taken. A challenge here consists in finding a balance between a risk-adverse strategy (e.g. a low P might frequently lead to unnecessary preventive measures) and a risk-friendly strategy (e.g. a high P might lead to missing an extreme event). This dilemma is illustrated by the variation in the hit rate H and false alarm rate F (Eqs. 3 and 4) with P, as summarized by relative operating characteristics curves (ROC, Mason, 1982).

$$H = \frac{h}{h + m} = \frac{\text{hits}}{\text{observed events}}, \tag{3}$$

$$F = \frac{f}{f + c} = \frac{\text{false alarms}}{\text{observed non}-\text{events}}, \tag{4}$$

where h is the number of hits, m the number of misses, f the number of false alarms and c the number of correct rejections during the study period. h, m, f and c are defined using a contingency table (e.g. Zhu et al., 2002).

In forecasting of extreme events, false alarms are considerably less frequent than correct rejections, as highlighted by the well known Finley case for evaluating tornadoes (Murphy, 1996). Therefore, even false alarm prone systems can benefit from a low F. In contrast, the false alarm ratio FAR (Eq. 5) does not reward correct rejections, and hence can be considered as a more informative metric on the frequency of false alarms for severe events (Ambühl, 2010). Therefore, in this study, H-FAR curves were preferred to ROC curves (depicting H versus F).

$$\mathrm{FAR} = \frac{f}{f + h} = \frac{\text{false alarms}}{\text{forecast events}} \tag{5}$$

HREF and COSMO-7 forecasts were considered as binary forecasts (i.e. exceedance or not of the discharge threshold by the daily maximum) to compute corresponding H and FAR.

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

Reliability diagrams enable in particular the assessment of model reliability, i.e. of the correspondence between the forecast probability and the observed relative frequency (e.g. Wilks, 2006). The diagram associated with reliable forecasts follows the plot diagonal.

Rank diagrams show the rank of OBS (resp. of HREF) within the ensemble members (Anderson, 1996). They highlight whether the ensemble includes OBS (resp. HREF) being predicted as an equiprobable member. If it does, the rank histogram has an uniform distribution. Other histogram shapes indicate over- or underdispersion tendencies and model biases (see Wilks, 2006, for examples).

3.4 Visualisation of case studies using continuous persistence plots

Uncertainty in probabilistic forecasts is not solely depicted by the spread of the ensemble members, but is also reflected by the persistence of the forecast, i.e. the consistency with which an event is forecast by successive model runs. For instance, a model showing great variability from one run to the next will be interpreted by the end-user as uncertain. This has significant consequences when the forecasts are used for decision-support, for example to decide on flood mitigation actions. In presence of a forecast showing great variability, the end-user might prefer not to base her/his decision on this forecast and to wait for the release of the next model run, delaying thus the decision process and taking the risk to end in an emergency situation, with a limited range of generally sub-optimal actions at choice. Forecast consistency is therefore greatly valued by end-users (Lashley et al., 2008).

For the evaluation of the forecasts for the two most intense events of the study period, a novel representation of probabilistic forecasts is proposed. Similarly to persistence plots (e.g. Thielen et al., 2009b), this new type of plot shows how the predictions of a given event evolve over time, by displaying the output of several model runs on the same graph.

However, while for each realisation of the model, persistence plots usually display one forecast threshold exceedance per day of forecast, hourly values of selected quantiles are depicted by these plots. In this article, the ensemble minimum and maximum, as well as the 25 % and the 75 % quantiles, were chosen. We argue that this kind of plot enables a finer and more quantitative comparison of the model runs because (1) it is based on hourly instead of daily values and (2) we consider that the four quantiles chosen reflect with more details the distribution of the HEPS output than threshold exceedance information.

For readability reasons, only a selection of lead times are shown and transparent colours are used to depict the inter-quartile ranges (IQR). We acknowledge that the interpretation of such a graphical representation might require an adaptation time. While we believe that traditional persistence plots are an efficient way to provide a global overview of the situation in a first place, we found these "continuous persistence plots" useful to obtain complementary and deeper insights into the forecasts, for operational work and verification exercises as in this paper. Indices to quantify forecasts consistency (Kay, 2004; Lashley et al., 2008) would constitute a helpful complement to the graphic representations.

3.5 Scenarios based on an artificially increased Lake Sihl level

Under normal conditions, the areas upstream and downstream of the Lake Sihl dam can be considered as uncoupled. About 88 % of the inflows to Lake Sihl are used for energy production and released directly into Lake Zurich (Fig. 1). Only low water amounts are necessary to guarantee the residual water discharge in the downstream part of the basin, as required by the Swiss environmental law. However, during heavy precipitation events, the application of the dam emergency regulation may result in significant water releases into the Sihl. In such situations, the catchment area upstream of the dam contributes greatly to the discharge in Zurich. To explore the consequences of coupling the upstream and downstream areas of the catchment during an extreme event, scenarios were considered. For the two discharge events investigated using continuous persistence plots, the lake balance and the hydraulic model were initialised using an artificially increased Lake Sihl level. For each event, two simulations were started: one forced by interpolated observed meteorological data (HREF-SCEN) and one using the COSMO-LEPS forecast initiated about one day before the peak discharge observed in Zurich (CLEPS-SCEN).

4 Results and discussion

4.1 Quantification of the added-value of probabilistic forecasts

For the Lake Sihl level, COSMO-LEPS- and COSMO-7-based forecasts show almost equal scores for the shortest lead time (LT1), but performance differences in favour of COSMO-LEPS increase with the time horizon (Table 2). In particular, the COSMO-LEPS LT5 median shows an equally elevated NSE value as the COSMO-7 LT3 forecasts. The COSMO-LEPS median is also associated with slightly better results in terms of MAE, but this should not draw attention from the relatively high amplitude of absolute error in both models. It is as high as several centimetres and even reaches 15.9 cm for the COSMO-LEPS LT5 median. A non-negligible source of errors for the Lake Sihl forecasts is the multiple regression used to approximate the hydropower production. We found that in standard conditions, it explains

Table 2. Nash Sutcliffe efficiency NSE (−) and mean absolute error MAE (cm) for the daily maximum of the Lake Sihl level. HREF reflects the skill of the hydrological/hydraulic part of the model chain. Forecasts based on COSMO-LEPS median (CLm) and COSMO-7 (C7) are evaluated for lead times (LT) of 1 to 5, and 1 to 3 days, respectively.

	HREF	LT1	LT2	LT3	LT4	LT5
NSE CLm	0.84	0.65	0.65	0.58	0.52	0.46
NSE C7		0.64	0.56	0.46	–	–
MAE CLm	1.7	2.4	5.9	9.2	12.5	15.9
MAE C7		2.4	6.3	10.1	–	–

Table 3. Nash Sutcliffe efficiency NSE (−), mean absolute error MAE ($m^3 s^{-1}$) and volume error VOL (%) for the Sihl discharge in Zurich. The notation conventions are the same as in Table 2.

	HREF	LT1	LT2	LT3	LT4	LT5
NSE CLm	0.87	0.70	0.44	0.25	0.20	0.10
NSE C7		0.55	0.20	−0.09	–	–
MAE CLm	3	3.9	5.2	5.6	6	6.3
MAE C7		4.1	5.7	6.2	–	–
VOL CLm	17	18	18	12	9	3
VOL C7		12	7	12	–	–

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

For forecasts of the Sihl discharge in Zurich, the added-value conveyed by the probability information can be appreciated by comparing the COSMO-LEPS median to the COSMO-7 NSE and MAE (Table 3). For these two scores, using the COSMO-LEPS median instead of the COSMO-7 forecast corresponds to a performance gain of 1 to 2 days lead time (COSMO-LEPS LT3 and LT5 are equivalent or better than COSMO-7 LT2 and LT3, respectively). MAE amplitude could be reduced by tuning PREVAH to better simulate low flows. However, this would probably be at the expense of the flood forecasting performance (Viviroli et al., 2009c). Positive VOL values indicate a discharge overestimation for both atmospheric models and all lead times. For the metrics considered, the overall performance of the COSMO-LEPS median is better than that of COSMO-7.

Some key forecast characteristics are summarized by BSS variations with the lead time and the discharge threshold (Fig. 4):

- the COSMO-LEPS scores are higher than those achieved forcing PREVAH with COSMO-7 forecasts. This is valid for all lead times and thresholds, and constitutes a quantitative proof of the benefits of running a probability model for the Sihl catchment;
- as expected, BSS reflects the difficulty of correctly forecasting intense events and the decline in weather predictability with increasing lead time. This is depicted by better scores for lower discharge thresholds and generally decreasing values for longer lead times;
- the decrease in performance is faster for COSMO-7 than for COSMO-LEPS, i.e. the "loss of BSS per day of lead time" is smaller for COSMO-LEPS than for COSMO-7. For instance, this appears clearly for the threshold $Q0.9$. It could indicate greater robustness of mid-term probabilistic forecasts thanks to the sampling of the initial atmospheric uncertainties;
- the mid-term forecasts (LT3 to LT5) for the $Q0.9$ and $Q0.99$ thresholds have little skill, and sometimes no skill. This reflects the limited predictability of high discharge events in the small Sihl catchment with the currently available forecasting chain;
- the size of the error bars underlines that the uncertainty in evaluating model performance increases significantly with event intensity. This emphasizes under-sampling resulting from the rarity of extreme events.

4.2 Dispersion of the HEPS members

Rank histograms depicting the OBS rank for all days of the time series (first column in Fig. 5) show overpopulation of the lowest bin. This denotes recurrent discharge overestimation in Zurich, although this tendency is slightly dampened by increasing lead times. When the HREF rank is depicted (second column), rank uniformity is improved. This suggests that the overestimation originates at least partially from the hydrological model, and affects HREF and COSMO-LEPS forecasts similarly.

When HREF is considered instead of OBS, the histograms switch from an "L-shape" to a "W-shape" (LT1) or a "U-shape" (LT3 and, to a lesser extent, LT5). The population of the two extreme ranks is higher than average, which indicates that COSMO-LEPS atmospheric forecasts are globally affected by underdispersion. However, as the ensemble spread usually increases with lead time, this tendency is weaker for LT4 and LT5. There seem to be several reasons for this underdispersion. In particular, the ensemble is coerced by the deterministic initial conditions, so that the spread for the first few hours of the forecast is too narrow.

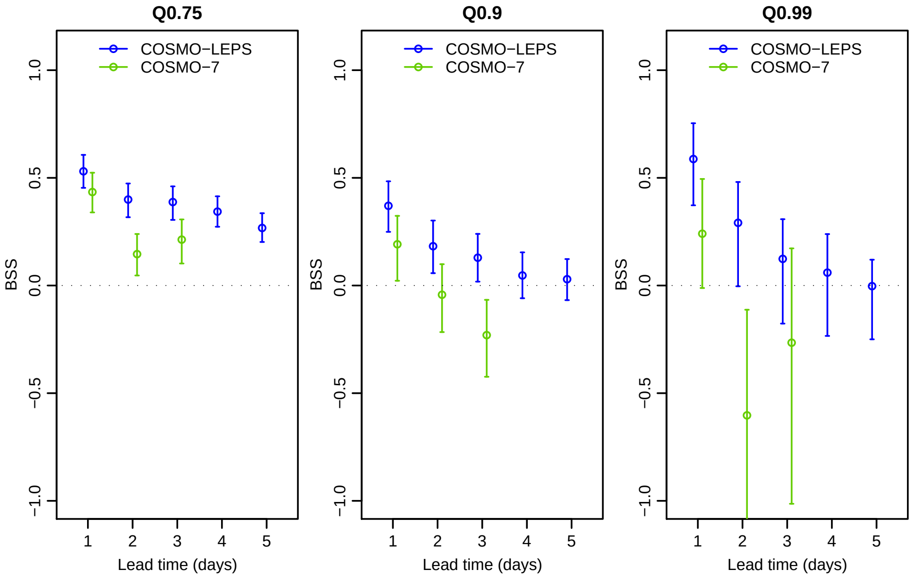

Fig. 4. Brier skill scores (BSS) for COSMO-LEPS- and COSMO-7-based forecasts for the daily maximum Sihl discharge in Zurich. Scores for the discharge thresholds $Q0.75$, $Q0.9$ and $Q0.99$ are shown from left to right. The circles exhibit the raw BSS, while the extremities of the confidence intervals consist of the 5th and 95th percentiles derived by bootstrapping.

This overconfidence for short-term forecasts is also due to the ECMWF EPS setup, which maximizes the growth of the perturbation total energy in the first 48 h of the forecast (Buizza and Palmer, 1995). As COSMO-LEPS relies on a combination of the two youngest EPS runs (Marsigli et al., 2005), its spread probably needs around two days to develop and reflect atmospheric uncertainty. But note that underdispersion is still pronounced in LT3 forecasts.

The overconfidence of COSMO-LEPS-based flow forecasts has also been reported by Renner et al. (2009). Marsigli et al. (2008) found that the percentage of outliers for 66-h COSMO-LEPS precipitation forecasts reached about 30 % (around 2.5 times the theoretical percentage) when considering the maximum values over boxes of 1.0 × 1.0°. 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.

The “W-shape” of the histogram depicting the HREF rank for LT1 can be explained as follows. The comparatively high population of rank 9 is due to the initialization of the ensemble using HREF. If the initialization discharge is the highest discharge of the day for HREF and the 16 members, these 17 simulations will have the same daily maximum. This results in the value 9 (the mean rank among 17 elements) being assigned as the HREF rank. It does not reflect ensemble overdispersion but results from the model setup. It disappears for lead times exceeding 1 day.

To focus on events of more interest from a flood perspective, only forecasts for days with a maximum discharge higher than a selected threshold are considered. Note that the higher the threshold, the greater the under-sampling. When a threshold selection is applied, the histograms showing OBS rank lose their “L-shape”. This suggests that low and middle discharges are overestimated and that such overestimation is probably due to calibration of the hydrological/hydraulic model setup. It is also possibly related to an overestimation of precipitation occurrence and very light rain events, as it is common with numerical weather prediction model outputs. All this helps to explain the positive VOL values noted earlier (Table 3).

The third column (observed discharges exceeding $Q0.9$) in Fig. 5 indicates frequent underdispersion for LT1 forecasts. For LT3 and LT5, the ensemble members tend to underestimate the intensity of observed larger events, as illustrated by high ranks being more populated than the low ranks. This tendency is also reflected in the histograms of the fourth column (observed discharge exceeding $Q0.99$). This implies that the discharges of the most intense events during the study period were associated with relatively low probabilities three to five days before their occurrence. Although

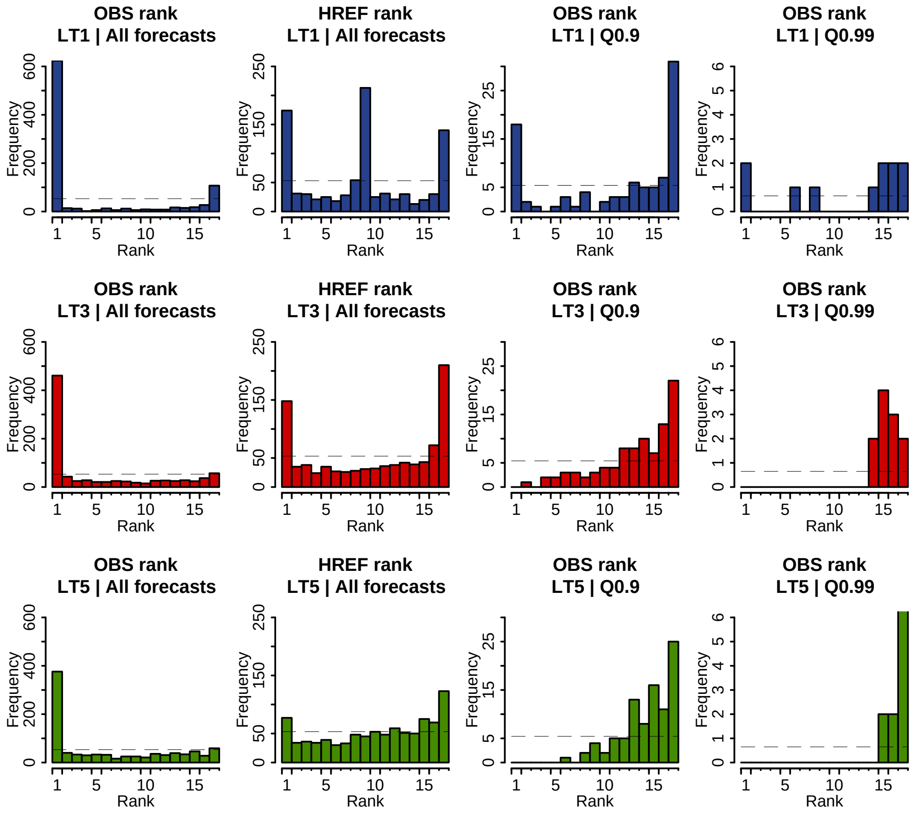

Fig. 5. Rank histograms for the daily maximum Sihl discharge in Zurich. The rank of OBS or HREF, within the 16 daily maxima forecast by the ensemble members, is depicted for lead times of 1, 3 and 5 days (LT1, LT3 and LT5). Histograms of the two first columns are based on the whole time series. For the two last columns, only days with an observed discharge exceeding $Q0.9$ and $Q0.99$, respectively, are included. Perfect rank uniformity is indicated by the horizontal dashed line.

this may hinder the effective anticipation of high discharges, more intense events might show earlier warning signs (e.g. Jaun et al., 2008; Thielen et al., 2009b).

For the forecasts of the Lake Sihl level, the two extreme ranks of the histograms are overpopulated for all lead times (not shown). This means that the atmospheric uncertainty, as propagated by HEPS, underestimates the full system uncertainty. As already mentioned, approximations of the hydropower production using a multiple regression represent a large source of errors, but they are not represented by the ensemble spread.

4.3 Hit rate and false alarm ratio

Figure 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 monetary 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 ($H \sim 0.35$).

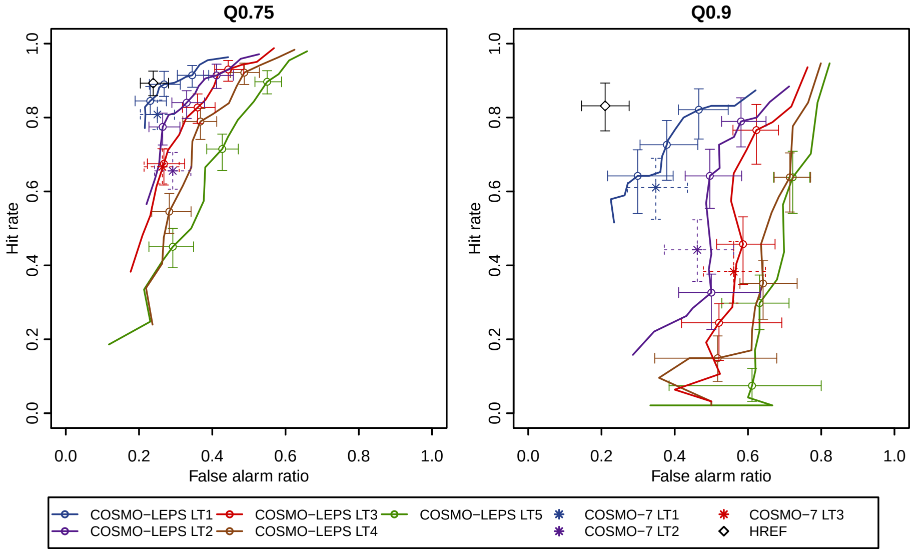

Fig. 6. False alarm ratio and hit rate for the daily maximum Sihl discharge in Zurich for the discharge thresholds $Q0.75$ (left) and $Q0.9$ (right) and for lead times (LT) of 1 to 5 days. The lines refer to COSMO-LEPS-based forecasts and their circles correspond to the probability thresholds 25 %, 50 % and 75 % (from right to left). The diamonds and stars indicate the performance of HREF and COSMO-7-based forecasts, respectively. The raw scores are exhibited by the symbols (circles, diamonds and stars), while the extremities of the confidence intervals consist of the 5th and 95th percentiles derived by bootstrapping.

$Q0.9$ corresponds to a discharge of about $21\,\mathrm{m}^3\,\mathrm{s}^{-1}$, 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.

4.4 Forecasts reliability

It was not possible to consider the forecast reliability for each probability threshold because of the limited number of events (see the small effective shown by the sub-plots in Fig. 7 and the important dispersion of the circles). Instead, we tried to capture the dominant tendency using linear regressions. Regression lines for all lead times and both thresholds are mostly located under the diagonal of the plots, with a slope lower than 1. This denotes overforecasting (forecast probability overestimation), which seems to be more accentuated for $Q0.9$ than for $Q0.75$. As a consequence, caution is required when using raw model outputs to assess flood risk in its most basic definition ("probability times consequence") as it might to lead to biased (overestimated) risk estimates for the Sihl catchment.

This overforecasting tendency probably does not come solely from PREVAH calibration, but also stems from COSMO-LEPS tendency to overforecast precipitation (Marsigli et al., 2008). The production of too wet forecasts has been in particular demonstrated for Switzerland using a single-member reforecast of 30 years (Fundel et al., 2010). The mean amplitude of this bias depends on the intensity of the event considered, the region and the season, but in the large majority of the country (including the Sihl catchment) precipitation amounts are generally overestimated. This bias can however be reduced consistently by post-processing calibration, which leads to more reliable forecasts (Fundel et al., 2010). Although some experience exists at the European scale (Thielen et al., 2009b), the influence of this bias on the reliability of discharge forecasts still needs to be assessed.

4.5 Insights from the event on 8 August 2007

For both the events chosen, precipitation forecasts were compared to hourly rainfall, measured by the stations shown in Fig. 1. The data were analysed using continuous persistence plots. Theses figures are not included in this paper, but some of the findings are enumerated here in order to better understand the impact of precipitation forecasts on the predicted discharge. Two intense precipitation events on 8 August 2007 triggered the generation of two distinct peak flow events (Fig. 8). A first peak discharge in Zurich was recorded at 09:00 UTC and a second at 23:00 UTC ($229\,\mathrm{m}^3\,\mathrm{s}^{-1}$, return period of ∼18 years). The first intense precipitation event was missed by COSMO-LEPS and COSMO-7 for all time horizons. Both models performed better in forecasting the second precipitation peak, although they underestimated it for all lead times.

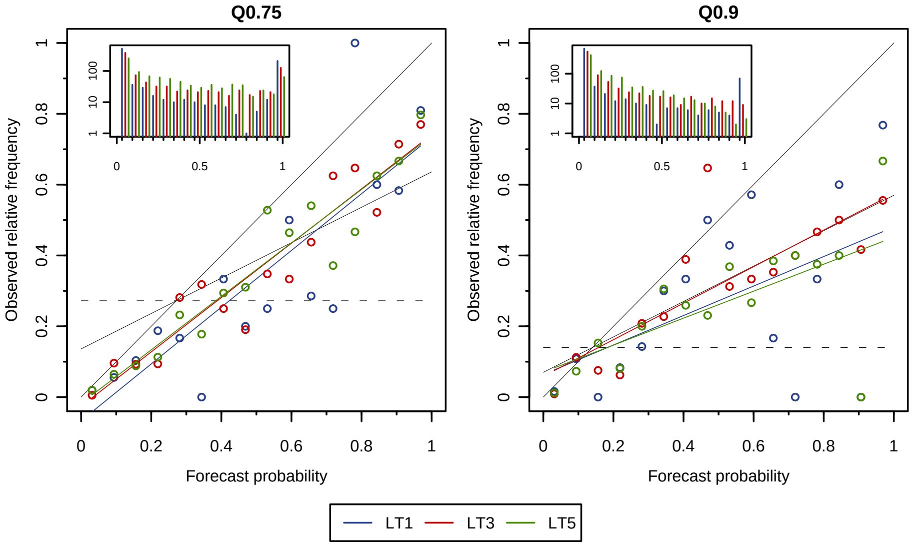

Fig. 7. Reliability diagrams for the daily maximum Sihl discharge in Zurich for the discharge thresholds $Q0.75$ (left) and $Q0.9$ (right). The circles indicate the observed frequency of each forecast probability class for lead times of 1, 3 and 5 days (LT1, LT3 and LT5). A linear regression is depicted for each lead time. The sub-plots show the associated refinement distributions.

The COSMO-LEPS-based hydrological forecast on the day of the event (LT1) showed underdispersion and its spread did not envelope the two discharge peaks (Fig. 8a). The peaks were exceeded by the daily maxima of a single member of the LT2 forecast, indicating an underestimation of the observed peak discharge by most ensemble members. The LT3 forecast missed both peaks. On 9 August, model initialisation using HREF (very close to the observed value) explained the good performance of the LT1 forecast. The LT2 and the LT3 forecasts showed higher discharges on 9 August than on the day of the event, but remained overall lower than the observed maximum. COSMO-7-based hydrological forecasts also reflected the rather conservative precipitation forecasts. The first peak amplitude was underestimated by at least a factor 3 and the second by at least a factor 2 for all time horizons (Fig. 8b).

The poor hydrological forecasts of this event can be mainly attributed to the atmospheric components of the model chain. This is confirmed by the satisfying agreement between OBS and the HREF run, which captured the timing and the magnitude of both peaks correctly.

4.6 Insights from the event on 15 August 2008

The Sihl discharge in Zurich reached 136 m^3 s^{-1} at 18:00 UTC on 15 August 2008 (return period of ~3 years). The correspondence between the COSMO-LEPS precipitation forecasts for 15 August and the observed rainfall increased when going from LT3 to LT1. The forecast initiated the day of the event provided a good approximation of the 24 h cumulated rainfall. However, the forecasts suffered from a rainfall overestimation in the morning. The COSMO-7 cumulated precipitation amounts were lower for LT1 than for the antecedent forecasts (LT2 and LT3). Observed amounts were underestimated for the afternoon when the main precipitation event was recorded.

The Sihl discharge based on COSMO-LEPS forecasts increased too early on 15 August (Fig. 9a) because of the precipitation overestimation for the morning. For all the depicted quantiles, the peak discharge gradually increased with decreasing lead times. The forecast initiated at 00:00 UTC on 15 August nicely enveloped the amplitude of the peak discharge, although the observed discharge increase was steeper and occurred a few hours later than forecast. The COSMO-7-based forecast with best correspondence to the observed hydrograph is LT3 (Fig. 9b). For this event, COSMO-7-based forecasts worsened with decreasing lead time and were clearly outperformed by the probabilistic forecasts.

The observed peak discharge amplitude on 15 August 2008 was well captured by HREF, although it was simulated a few hours too late.

4.7 Insights from the scenarios

Given the satisfying correspondence between HREF and the observed discharge for the two analysed events (Figs. 8 and 9), HREF-SCEN is assumed to approximate the discharge that would have been observed if the level of the Lake Sihl was of 889.00 m a.s.l. when the forecasts were initialized.

In the case of the August 2007 event, the operation limit of the dam (889.34 m a.s.l.) was exceeded by HREF-SCEN for 52 h (Fig. 10). This caused an emergency water release into the Sihl, coupled with a dam overflow whose peak reached 180 % of the observed peak discharge at the outlet of the dam (river gauge Schlagen, see Fig. 1). The first inflow peak from

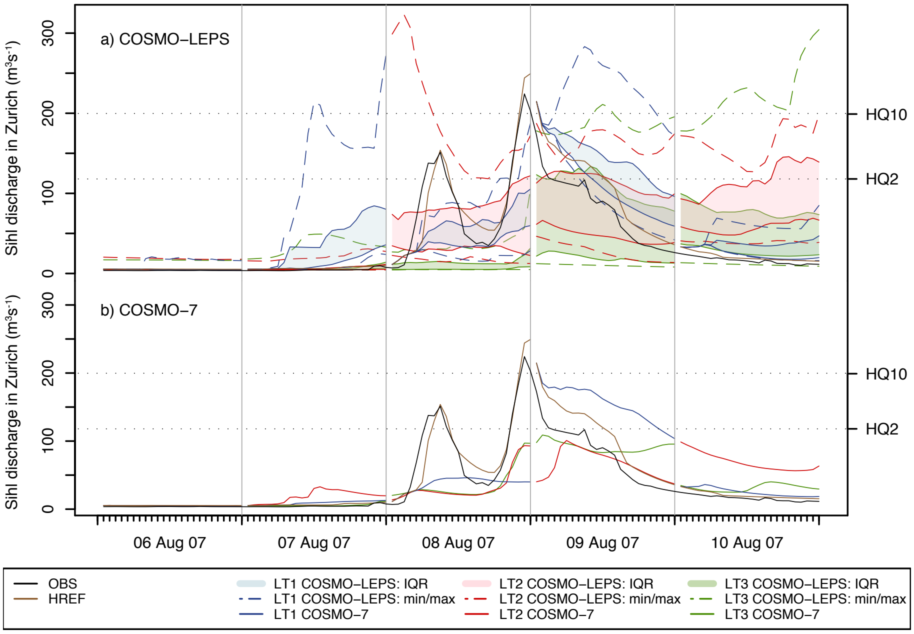

Fig. 8. Continuous persistence plots centred on the 8 August 2007 event depicting the discharge in Zurich. COSMO-LEPS-based **(a)** and COSMO-7-based forecasts **(b)** are shown for lead times of 1, 2 and 3 days (LT1, LT2 and LT3). The vertical grey lines indicate 00:00 UTC. Discharges associated with return periods of 2 and 10 years are depicted by the horizontal dotted lines, HQ2 and HQ10, respectively.

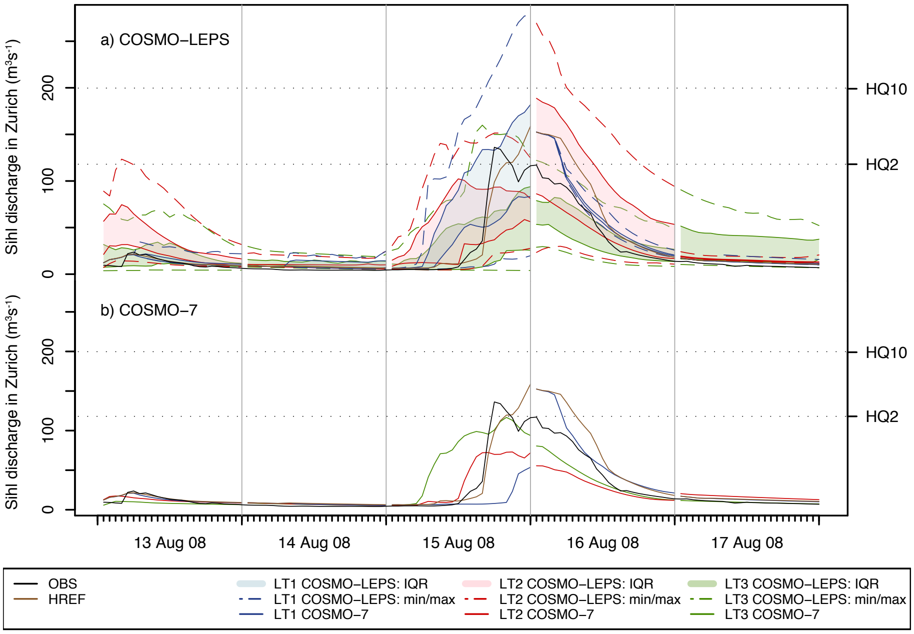

Fig. 9. Continuous persistence plots as in Fig. 8, but centred on the 15 August 2008 event.

the Lake Sihl catchment caused a sharp lake level increase, but most water was retained by the dam. This is indicated by the comparatively low resulting HREF-SCEN discharge at the dam outlet. However, it caused the Lake Sihl to reach and exceed the dam operation limit. As a consequence, most of the inflow generated during the second rainfall-runoff event was released into the Sihl. This is supported by the close match between the curves depicting HREF-SCEN discharge in Schlagen and the sum of the Lake Sihl inflows. The peak discharge of the emergency release into the Sihl occurred two hours before the peak observed in Zurich. As the travel time from the dam to Zurich is around three hours for high discharges (Schwanbeck et al., 2007), this release accentuated the observed peak in Zurich. Hence, a situation like

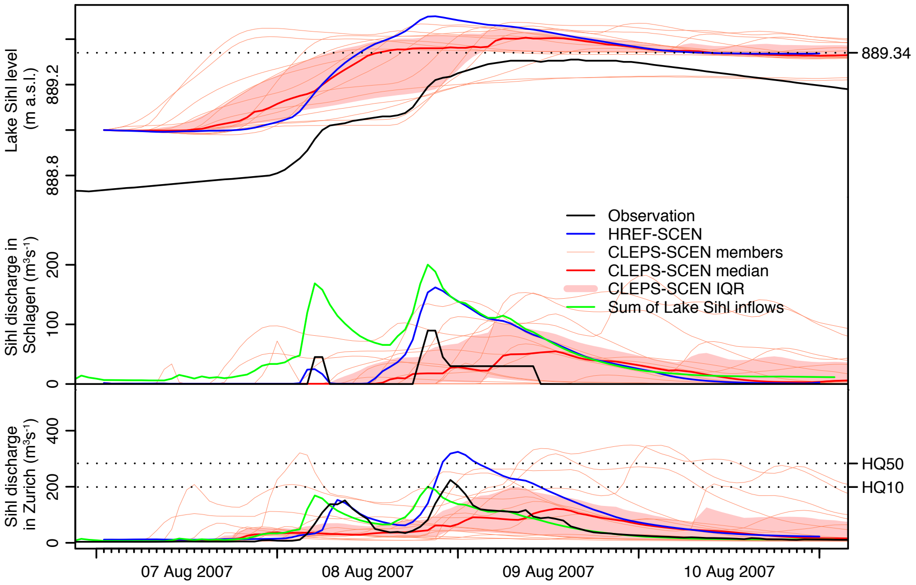

Fig. 10. Lake Sihl level (top), water released at the dam outlet (Schlagen, centre) and discharge in Zurich (bottom) for an artificially increased Lake Sihl level of 889 m a.s.l. on 7 August 2007 at 00:00 UTC. Dotted lines indicate the altitude of the dam operation limit (889.34 m a.s.l., top plot) and the discharges associated with return periods of 10 and 50 years (HQ10 and HQ50, bottom plot).

HREF-SCEN in August 2007 could have led to a peak runoff of about 325 $m^3 s^{-1}$. Such a runoff would have probably caused the flooding of the construction site and large damage to the city of Zurich.

As the forecast evaluation would let us expect (Sect. 4.5), this peak discharge was heavily underestimated by CLEPS-SCEN, principally as a result of the propagation of the precipitation underestimation by COSMO-LEPS. It caused the lake level to be underestimated, which biased the water release in Schlagen and led to an overly conservative forecast for the Sihl discharge in Zurich. It is probable that if only this forecast for the Sihl discharge in Zurich had been considered, no preventive action would have been taken, although it would have been clearly necessary. CLEPS-SCEN did, however, indicate that an exceedance of the dam operation limit and an emergency water release into the Sihl were very probable. In such a situation, producing hydroelectricity at maximum capacity can help to reduce damage by generating additional storage capacity within the reservoir. On 7 August, the daily average discharge for hydroelectricity production was of 6.33 $m^3 s^{-1}$, well below the maximum capacity of 26 $m^3 s^{-1}$. Proceeding to a *controlled* lake drawdown into the Sihl before the event can moreover decrease the risk of superposition of the peak discharge caused by an unfavourable *forced* water release with the peak generated in the downstream part of the basin. These two mitigation measures could probably have been implemented successfully on the basis of this still imperfect CLEPS-SCEN forecast.

In the case of the August 2008 event (Fig. 11), the lake level simulated by HREF-SCEN exceeded the dam operation limit by 14 cm at its maximum. The peak discharge released into the Sihl was 107 $m^3 s^{-1}$ greater than the observation, and took place when the total Lake Sihl inflows were close to their maximum and almost fully released into the Sihl. Although CLEPS-SCEN simulated the maximum lake level and the peak discharge in Schlagen around four hours too early, it captured their amplitude correctly. HREF-SCEN peak discharge in Schlagen occurred around eight hours *after* the observed peak in Zurich. Hence, the two wave peaks were delayed and did not superimpose. HREF-SCEN peak discharge in Zurich (238 $m^3 s^{-1}$) was higher than the observation (136 $m^3 s^{-1}$), but probably would have not caused more serious damage than driftwood. Nevertheless, as seven members of CLEPS-SCEN exceeded 300 $m^3 s^{-1}$, a preventive lake drawdown would probably have been chosen on the basis of this hypothetical forecast. It can in this case be argued that CLEPS-SCEN correctly reflected the flooding risk in Zurich, and that it justified the cost of a preventive drawdown.

4.8 Challenges in decision-making based on hydrometeorological forecasts

The first scenario illustrates how dam overflow can cause serious damage in Zurich and emphasizes the importance of timely controlled water release into the Sihl and modulated hydropower production. Concretely, this implies determining how much and when water should be released to minimize water losses and the flood risk. A new module is currently being developed. It consists of an online interface where decision makers can prescribe several drawdown regimes. Re-running the hydraulic model then enables

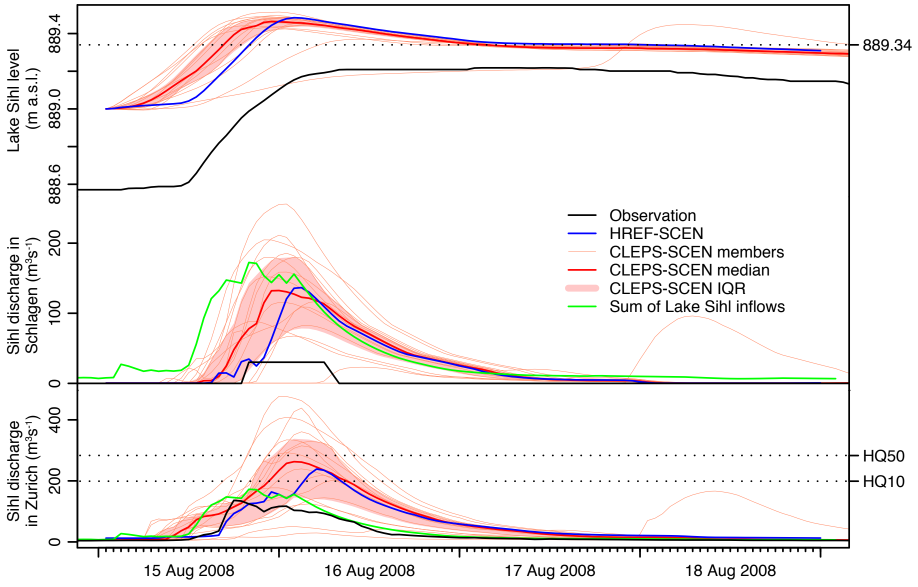

Fig. 11. As in Fig. 10, but for an artificially increased Lake Sihl level of 889 m a.s.l. on 15 August 2008 at 00:00 UTC.

consequences of these scenarios on the peak runoff forecast to be compared. As new atmospheric forecasts become available, the chosen drawdown strategy can be re-evaluated and if necessary adjusted.

Several studies made use of the cost-loss ratio method to interpret the quality of atmospheric and hydrologic ensemble forecasts in terms of forecast value (Richardson, 2000; Roulin, 2007). This method is a static method and is probably too simplistic to provide efficient guidance for a situation like the Sihl catchment, which involves several stakeholders with divergent interests, as well as several interrelated mitigation actions. Multi-purpose dam-management, for example based on dynamic programming, could be envisaged to circumvent these limitations (e.g. Faber and Stedinger, 2001; Yao and Georgakakos, 2001; Turgeon, 2005).

Improvements of the system towards decision support might focus on quantitatively assessing whether taking the risk of performing an unnecessary drawdown is justified by an even higher risk of flooding in Zurich. This would require at least two cost-loss functions: one relating the flooding damage in Zurich to the Sihl discharge and one expressing the costs likely to be incurred by the dam operators (losses in energy production) if water is released into the Sihl. The probabilistic hydrological forecasts could be used as input for these two functions to quantify risk. This procedure still presents at least two difficulties for the present case study. First, the reliability diagrams point towards a possible overforecasting of the discharge in Zurich, and too few events are available to assess whether forecasts for extreme discharges are reliable or not. Hence, it seems risky to use raw ensemble outputs as probabilities. Post-processing of the forecasts before combining them with economic data (e.g. via Bayesian calibration, Reggiani et al., 2009; Raftery et al., 2005) is probably a necessary step towards reliable risk assessment. Second, a quantitative risk estimate and a cost-benefits analysis of the system require the determination of several cost-loss functions for the Sihl catchment. Estimations of flood costs in the city of Zurich are underway and will be available at the earliest in the second half of 2011. Until then, only orders of magnitude are available. It is at present unclear whether precise and accurate risk assessment is necessary for flood mitigation in the Sihl catchment, or if robust protection measures can be implemented without reliable models and rather approximate economic information (Dessai et al., 2009).

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 Lake Sihl 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 relevant from a flooding perspective, and more to their forecasts could lead to a sounder management (less forced, uncontrolled water releases). However, improved forecasts are therefore needed. When we presented the simulations of the August 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.

5 Conclusions and outlook

This study reveals that probability information can be efficiently used by the model chain and delivers useful support for flood mitigation in the Sihl catchment. Multiple deterministic and probabilistic metrics, as well as graphical representations, have been used to evaluate the model chain. The performance of the hydrologic ensemble prediction system is better and decreases less rapidly with lead time than for deterministic forecasts. However, the spread of the weather forecasts is often too low. The hydrological and hydraulic models appear overall to perform well in capturing the amplitude and the timing of the observed peak discharges. The largest source of forecast uncertainty stems from the difficulty of accurately forecasting the intensity, location and timing of intense precipitation events in the relatively small-scale Sihl catchment. Although caution is required because of under-sampling, this seems to limit the ability of mid-term forecasts to confidently and reliably capture observed intense peak discharges. Therefore, although probabilistic forecasts do convey added-value in comparison to deterministic ones, precipitation forecasts must be improved to guarantee sufficiently early flood predictions in the Sihl catchment.

The reliability diagrams and false alarm ratios suggest that medium to high discharges tend to be overforecast. This might as well affect extreme discharge forecasts and impede reliable assessment of flood risk. Furthermore, the first high discharge scenario showed that uncontrolled water releases into the Sihl could lead to dramatic damage in Zurich. This advocates for the development of a dedicated system to support efficiently decision-making based on hydrological forecasts. Correct streamflow forecasts may not be sufficient for efficient flood mitigation if they are not accompanied by a dedicated tool to compare multiple mitigation actions.

As presented in Sect. 3.1, a more robust assessment of the flood forecasting capacity of the system and the further development of an efficient decision-support system imply an enhancement of the reforecast period.

This study focused on the uncertainty related to atmospheric boundary conditions. Probable future developments include the integration of modules to account for other uncertainty sources (e.g. Zappa et al., 2011) such as the formulation of the atmospheric models, the stations measurements, the interpolation errors, the estimation of the hydropower production and the hydrological and hydraulic modelling. For instance, the combination of ensemble forecasts with deterministic forecasts could be explored (e.g. Dietrich et al., 2008) to give more weight to uncertainties stemming from the formulation of atmospheric models. COSMO-7 and COSMO-2 being more frequently updated than COSMO-LEPS, this would furthermore provide time-lagged ensembles of discharge predictions (e.g. Zappa et al., 2008). In addition, an ensemble radar precipitation (Germann et al., 2009) could be used to assess the measurement errors, and an observational precipitation ensemble (Ahrens and Jaun, 2007) could be implemented to study the interpolation uncertainty. In parallel, using calibrated COSMO-LEPS rainfall forecasts (Fundel et al., 2010) to drive the hydrological and hydraulic model is planned. As this calibration method based on quantile mapping improves the reliability of precipitation forecasts, it is expected to improve the discharge forecasts as well. Note that alternative calibration methods for limited-area ensemble precipitation forecasts are currently investigated (Diomede et al., 2011).

This study illustrated the challenge that represent the interpretation, communication and efficient use of probabilistic forecasts for decision-making (Demeritt et al., 2007; Bruen et al., 2010). We would like to emphasize that the framework of this study is a real-life case and not purely experimental. There is a real panel of experts consisting of hydrologists and stakeholders, with the delicate task of making decisions by interpreting the outputs of high-end but nevertheless imperfect models (Badoux et al., 2010). Further real-time experience in dealing with such uncertainties should be gained by the end of the construction of the new railway station below the Sihl River.

Acknowledgements. This study is part of COST-731 (Zappa et al., 2010). The IFKIS-SIHL project is founded by the Department of Waste, Water, Energy and Air (AWEL) of Canton Zurich. The contribution of the WSL to the Sihl experts panel was partly funded by the Swiss Federal Railway (SBB). We are grateful to the Swiss Federal Office of Meteorology and Climatology, MeteoSwiss, and the Swiss Federal Office for the Environment for providing meteorological and hydrological data. C. Schär is thanked for advising the master thesis of N. Addor, on which this article is based. The constructive reviews of H. Cloke, M. Werner and of a third anonymous reviewer are greatly acknowledged. We are grateful to S. Dingwall and E. Gleeson for the linguistic revision of the manuscript. The data analysis and visualisation were performed with R (R Development Core Team, 2011).

Edited by: A. Shamseldin

References

Ahrens, B. and Jaun, S.: On evaluation of ensemble precipitation forecasts with observation-based ensembles, Adv. Geosci., 10, 139–144, doi:10.5194/adgeo-10-139-2007, 2007.

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, doi:10.5194/adgeo-29-69-2011, 2011.

Ambühl, J.: Customer oriented warning systems, Tech. Rep. 84, Veröffentlichung der MeteoSchweiz, http://www.meteosuisse.admin.ch/web/de/forschung/publikationen/alle_publikationen/veroeff_84.Par.0001.DownloadFile.tmp/veroeff84.pdf (last access: July 2011), 2010.

Anderson, J.: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations, J. Climate, 9, 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, 1996.

Badoux, A., Zappa, M., Schatzmann, M., Optlatka, M., Jaun, S., Bösch, M., Gross, M., Steiner, P., Hegg, C., and Rhyner, J.: IFKIS-Hydro Sihl: Beratung, Alarmorganisation und Handlungsmöglichkeiten während dem Bau der Durchmesserlinie beim Hauptbahnhof Zürich, Wasser Energie Luft, 4, 309–320, 2010.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141–153, doi:10.5194/hess-13-141-2009, 2009.

Bezzola, G. R. and Hegg, C.: Ereignisanalyse Hochwasser 2005, Teil 1 – Prozesse, Schäden und erste Einordnung, Bundesamt für Umwelt BAFU, Eidgenössische Forschungsanstalt WSL, http://www.bafu.admin.ch/publikationen/publikation/00044/, last access: 19 January 2011, vol. 707, Umwelt-Wissen, 2007.

Bezzola, G. R. and Hegg, C.: Ereignisanalyse Hochwasser 2005, Teil 2 – Analyse von Prozessen, Massnahmen und Gefahrengrundlagen, Bundesamt für Umwelt BAFU, Eidgenössische Forschungsanstalt WSL, http://www.bafu.admin.ch/publikationen/publikation/00100/, last access: 19 January 2011, vol. 825, Umwelt-Wissen, 2008.

Bruen, M., Krahe, P., Zappa, M., Olsson, J., Vehvilainen, B., Kok, K., and Daamen, K.: Visualizing flood forecasting uncertainty: some current European EPS platforms – COST731 working group 3, Atoms. Sci. Lett., 11, 92–99, doi:10.1002/asl.258, 2010.

Buizza, R. and Palmer, T. N.: The Singular-Vector Structure of the Atmospheric Global Circulation, J. Atmos. Sci., 52, 1434–1456, doi:10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2, 1995.

Cloke, H. L. and Pappenberger, F.: Ensemble flood forecasting: A review, J. Hydrol., 375, 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.

Demargne, J., Brown, J., Liu, Y., Seo, D., Wu, L., Toth, Z., and Zhu, Y.: Diagnostic verification of hydrometeorological and hydrologic ensembles, Atmos. Sci. Lett., 11, 114–122, doi:10.1002/asl.261, 2010.

Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., and Ramos, M.: Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting, Environ. Hazard., 7, 115–127, doi:10.1016/j.envhaz.2007.05.001, 2007.

Dessai, S., Hulme, M., Lempert, R., and Pielke, R. J.: Adapting to climate change: thresholds, values, governance, chap. Climate prediction: a limit to adaptation?, Cambridge University Press, 64–78, 2009.

Dietrich, J., Trepte, S., Wang, Y., Schumann, A. H., Voß, F., Hesser, F. B., and Denhard, M.: Combination of different types of ensembles for the adaptive simulation of probabilistic flood forecasts: hindcasts for the Mulde 2002 extreme event, Nonlin. Processes Geophys., 15, 275–286, doi:10.5194/npg-15-275-2008, 2008..

Diomede, T., Marsigli, C., Montani, A., and Paccagnella, T.: Comparison of calibration techniques for a limited-area ensemble precipitation forecast using reforecasts, EGU2011-7261, EGU General Assembly, Geophys. Res. Abstr., 13, 2011.

Efron, B.: Jackknife-after-bootstrap standard errors and influence functions, J. Roy. Stat. Soc. B Meth., 54, 83–127, 1992.

Faber, B. A. and Stedinger, J. R.: Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, J. Hydrol., 249, 113–133, doi:10.1016/S0022-1694(01)00419-X, 2001.

Fundel, F., Walser, A., Liniger, M. A., Frei, C., and Appenzeller, C.: Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts, Mon. Weather Rev., 138, 176–189, doi:10.1175/2009MWR2977.1, 2010.

Germann, U., Berenguer, M., Sempere-Torres, D., and Zappa, M.: REAL – Ensemble radar precipitation estimation for hydrology in a mountainous region, doi:10.1002/qj.375, Q. J. Roy. Meteorol. Soc., 135, 445–456, 2009.

Gurtz, J., Baltensweiler, A., and Lang, H.: Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins, Hydrol. Process., 13, 2751–2768, doi:10.1002/(SICI)1099-1085(19991215)13:17<2751::AID-HYP897>3.0.CO;2-O, 1999.

Hamill, T., Whitaker, J., and Wei, X.: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts, Mon. Weather Rev., 132, 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2, 2004.

Jaun, S. and Ahrens, B.: Evaluation of a probabilistic hydrometeorological forecast system, Hydrol. Earth Syst. Sci., 13, 1031–1043, doi:10.5194/hess-13-1031-2009, 2009.

Jaun, S., Ahrens, B., Walser, A., Ewen, T., and Schär, C.: A probabilistic view on the August 2005 floods in the upper Rhine catchment, Nat. Hazards Earth Syst. Sci., 8, 281–291, doi:10.5194/nhess-8-281-2008, 2008.

Katz, R. W. and Murphy, A. H.: Economic value of weather and climate forecasts, Cambridge University Press, 1997.

Kay, M.: The design and evaluation of a measure of forecast consistency for the Collaborative Convective Forecast Product, Preprints, in: 11th Conference on Aviation, Range and Aerospace Meteorology, 4–8, 2004.

Lahiri, S.: Resampling methods for dependent data, Springer Verlag, 2003.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.

Lashley, S., Fisher, L., Simpson, B., Taylor, J., Weisser, S., Logsdon, J., and Lammers, A.: Observing verification trends and applying a methodology to probabilistic precipitation forecasts at a National Weather Service forecast office, in: 19th Conference on Probability and Statistics, 2008.

Marsigli, C., Boccanera, F., Montani, A., and Paccagnella, T.: The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification, Nonlin. Processes Geophys., 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.

Mason, I.: A model for assessment of weather forecasts, Aust. Meteorol. Mag., 30, 291–303, 1982.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble prediction system: methodology and validation, Q. J. Roy. Meteorol. Soc., 122, 73–119, doi:10.1002/qj.49712252905, 1996.

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. Meteorol. Soc., 127, 2069–2094, doi:10.1002/qj.49712757612, 2001.

Murphy, A. H.: Decision Making and the Value of Forecasts in a Generalized Model of the Cost-Loss Ratio Situation, Mon. Weather Rev., 113, 362–369, doi:10.1175/1520-0493(1985)113<0362:DMATVO>2.0.CO;2, 1985.

Murphy, A. H.: The Finley Affair: A Signal Event in the History of Forecast Verification, Weather Forecast., 11, 3–20, doi:10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2, 1996.

Nash, J. E. and Sutcliffe, J. V.: River forecasting using conceptual models, 1. A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

R Development Core Team: R: A language and environment for statistical computing, http://www.R-project.org, last access: 4 July 2011, ISBN 3-900051-07-0, 2011.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, doi:10.1175/MWR2906.1, 2005.

Reggiani, P., Renner, M., Weerts, A. H., and van Gelder, P. A. H. J. M.: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system, Water Resour. Res., 45, W02428, doi:10.1029/2007WR006758, 2009.

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.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Q. J. Roy. Meteorol. Soc., 126, 649–667, doi:10.1002/qj.49712656313, 2000.

Rotach, M. W., Ambrosetti, P., Appenzeller, C., Arpagaus, M., Fontannaz, L., Fundel, F., Germann, U., Hering, A., Liniger, M. A., Stoll, M., Walser, A., Ament, F., Bauer, H.-S., Behrendt, A., Wulfmeyer, V., Bouttier, F., Seity, Y., Buzzi, A., Davolio, S., Corazza, M., Denhard, M., Dorninger, M., Gorgas, T., Frick, J., Hegg, C., Zappa, M., Keil, C., Volkert, H., Marsigli, C., Montaini, A., McTaggart-Cowan, R., Mylne, K., Ranzi, R., Richard, E., Rossa, A., Santos-Muñoz, D., Schär, C., Staudinger, M., Wang, Y., and Werhahn, J.: MAP D-PHASE: Real-Time Demonstration of Weather Forecast Quality in the Alpine Region, B. Am. Meteorol. Soc., 90, 1321–1336, doi:10.1175/2009BAMS2776.1, 2009.

Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725–737, doi:10.5194/hess-11-725-2007, 2007.

Schwanbeck, J., Viviroli, D., Röser, I., Trosch, J., and Weingartner, R.: Prozessbasierte Abschätzung von Extremhochwassern im Einzugsgebiet der Sihl, Schlussbericht zur Studie im Auftrag des Amtes für Abfall, Wasser, Energie und Luft des Kantons Zürich (AWEL), Tech. rep., Veröffentlichung der Universität Bern, 2007.

Steppeler, J., Doms, G., Schättler, U., Bitzer, H., Gassmann, A., Damrath, U., and Gregoric, G.: Meso-gamma scale forecasts using the nonhydrostatic model LM, Meteorol. Atmos. Phys., 82, 75–96, doi:10.1007/s00703-001-0592-9, 2003.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13, 125–140, doi:10.5194/hess-13-125-2009, 2009a.

Thielen, J., Bogner, K., Pappenberger, F., Kalas, M., Del Medico, M., and de Roo, A.: Monthly-, medium-, and short-range flood warning: testing the limits of predictability, Meteorol. Appl., 16, 77–90, doi:10.1002/met.140, 2009b.

Turgeon, A.: Solving a stochastic reservoir management problem with multilag autocorrelated inflows, Water Resour. Res., 41, W12414, doi:10.1029/2004WR003846, 2005.

Velázquez, J. A., Petit, T., Lavoie, A., Boucher, M.-A., Turcotte, R., Fortin, V., and Anctil, F.: An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting, Hydrol. Earth Syst. Sci., 13, 2221–2231, doi:10.5194/hess-13-2221-2009, 2009.

Verbunt, M., Walser, A., Gurtz, J., Montani, A., and Schär, C.: Probabilistic flood forecasting with a limited-area ensemble prediction system: selected case studies, J. Hydrometeorol., 8, 897–909, doi:10.1175/JHM594.1, 2007.

Viviroli, D., Mittelbach, H., Gurtz, J., and Weingartner, R.: Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part II: Parameter regionalisation and flood estimation results, J. Hydrol., 377, 208–225, doi:10.1016/j.jhydrol.2009.08.022, 2009a.

Viviroli, D., Zappa, M., Gurtz, J., and Weingartner, R.: An introduction to the hydrological modelling system PREVAH and its pre-and post-processing-tools, Environ. Modell. Softw., 24, 1209–1222, doi:10.1016/j.envsoft.2009.04.001, 2009b.

Viviroli, D., Zappa, M., Schwanbeck, J., Gurtz, J., and Weingartner, R.: Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part I: Modelling framework and calibration results, J. Hydrol., 377, 191–207, doi:10.1016/j.jhydrol.2009.08.023, 2009c.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: The Discrete Brier and Ranked Probability Skill Scores, Mon. Weather Rev., 135, 118–124, doi:10.1175/MWR3280.1, 2007.

Wilks, D.: Statistical Methods in the Atmospheric Sciences, vol. 91 of International geophysics series, Elsevier, New York, 2006.

Yao, H. and Georgakakos, K. P.: Assessment of Folsom lake response to historical and potential future climate scenarios: 2. Reservoir management, J. Hydrol., 249, 176–196, doi:10.1016/S0022-1694(01)00418-8, 2001.

Zappa, M. and Kan, C.: Extreme heat and runoff extremes in the Swiss Alps, Nat. Hazards Earth Syst. Sci., 7, 375–389, doi:10.5194/nhess-7-375-2007, 2007.

Zappa, M., Rotach, M. W., Arpagaus, M., Dorninger, M., Hegg, C., Montani, A., Ranzi, R., Ament, F., Germann, U., Grossi, G., Jaun, S., Rossa, A., Vogt, S., Walser, A., Wehrhan, J., and Wunram, C.: MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems, Atoms. Sci. Lett., 9, 80–87, doi:10.1002/asl.183, 2008.

Zappa, M., Beven, K., Bruen, M., Cofino, A., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., Schroeter, K., Seed, A., Stzurc, J., Vehvilaeinen, B., Germann, U., and Rossa, A.: Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2, Atmos. Sci. Lett., 11, 83–91, doi:10.1002/asl.248, 2010.

Zappa, M., Jaun, S., Germann, U., Walser, A., and Fundel, F.: Superposition of three sources of uncertainties in operational flood forecasting chains, Atmos. Res., 100, 246–262, doi:10.1016/j.atmosres.2010.12.005, 2011.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The Economic Value Of Ensemble-Based Weather Forecasts, B. Am. Meteorol. Soc., 83, 73–83, doi:10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2, 2002.