Hydrol. Earth Syst. Sci., 15, 877–896, 2011
www.hydrol-earth-syst-sci.net/15/877/2011/
doi:10.5194/hess-15-877-2011

Hydrology and
Earth System
Sciences

Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events

U. Ehret and E. Zehe

Institute of Hydrology, KIT Karlsruhe Institute of Technology, Karlsruhe, Germany

Received: 18 October 2010 – Published in Hydrol. Earth Syst. Sci. Discuss.: 25 October 2010
Revised: 25 February 2011 – Accepted: 2 March 2011 – Published: 11 March 2011

Abstract. Applying metrics to quantify the similarity or dissimilarity of hydrographs is a central task in hydrological modelling, used both in model calibration and the evaluation of simulations or forecasts. Motivated by the shortcomings of standard objective metrics such as the Root Mean Square Error (RMSE) or the Mean Absolute Peak Time Error (MAPTE) and the advantages of visual inspection as a powerful tool for simultaneous, case-specific and multi-criteria (yet subjective) evaluation, we propose a new objective metric termed Series Distance, which is in close accordance with visual evaluation. The Series Distance quantifies the similarity of two hydrographs neither in a time-aggregated nor in a point-by-point manner, but on the scale of hydrological events. It consists of three parts, namely a Threat Score which evaluates overall agreement of event occurrence, and the overall distance of matching observed and simulated events with respect to amplitude and timing. The novelty of the latter two is the way in which matching point pairs on the observed and simulated hydrographs are identified: not by equality in time (as is the case with the RMSE), but by the same relative position in matching segments (rise or recession) of the event, indicating the same underlying hydrological process. Thus, amplitude and timing errors are calculated simultaneously but separately, from point pairs that also match visually, considering complete events rather than only individual points (as is the case with MAPTE). Relative weights can freely be assigned to each component of the Series Distance, which allows (subjective) customization of the metric to various fields of application, but in a traceable way. Each of the three components of the Series Distance can be used in an aggregated or non-aggregated way, which makes the Series Distance a suitable tool for differentiated, process-based model diagnostics.

After discussing the applicability of established time series metrics for hydrographs, we present the Series Distance theory, discuss its properties and compare it to those of standard metrics used in Hydrology, both at the example of simple, artificial hydrographs and an ensemble of realistic forecasts. The results suggest that the Series Distance quantifies the degree of similarity of two hydrographs in a way comparable to visual inspection, but in an objective, reproducible way.

1 Introduction

Imagine the following situation: after a flood, the hydrologists responsible for the forecasts and the flood management personnel meet for post-event analysis. The head of the dike defence team was not satisfied with the forecasts: the peak water level was falsely predicted above dike height, so many people were unnecessarily activated for sandbag piling. The operator of a large retention basin claims that the event was not indicated in the long-term forecasts, which would have been necessary for pre-event waterlevel drawdown. Further he reports that during the event, the forecast of the flood rise was correct with respect to timing, so reservoir operation was started just in time. But, he continues, the recession was predicted much too long, resulting in valuable reservoir volume kept free in vain.

This conversation is fictitious, but nonetheless realistic according to the author's experience in operational hydrology. If we analyze it, several aspects stand out: first, the meeting took place after and was focused on an event. Second,

Correspondence to: U. Ehret
(uwe.ehret@kit.edu)

in the discussion the event was subdivided into several segments and points of interest (rising limb, peak, recession), that were deemed important enough for separate evaluation. Third, the discussion was mainly based on the comparison of observed and forecasted hydrographs, not e.g. observed groundwater levels. Fourth, the different users focused on completely different aspects of performance such as long-term event prediction, peak water level, timing etc. and used different metrics for evaluation (event occurrence Yes/No, visual comparison of hydrograph shape, water level exceedence Yes/No etc.).

These points, based on an example from hydrological forecasting also apply to hydrological modeling and the evaluation of hydrological model performance in a more general sense: be it for parameter estimation during model calibration, model validation, classification of hydrological systems or identification of scales at which to separate explicit and implicit representations of structures and processes: metrics, measures and objective functions (including subjective visual inspection) are applied in all disciplines of Hydrology. The data used for evaluation may vary with the purpose of the model, however in practice hydrographs from gauge observations are the most widely used: they are relatively easy to obtain and still the most meaningful and relevant expression of integral hydrological behaviour on catchment scale. Also, historically hydrological modelling was mainly focused on analysis and reproduction of observed discharge time series at the catchment scale. Hence the repertoire of metrics in Hydrology was, and to a declining degree still is, mainly related to hydrographs.

Hydrographs possess properties that make them (from a hydrological point of view) a particular subset of time series in general. These properties are worth being considered when evaluating the appropriateness of metrics to quantify the similarity or dissimilarity of hydrographs and we will therefore briefly discuss them in the following.

1.1 Hydrograph characteristics

A hydrograph basically is a time series, i.e. a two-dimensional, time-ordered dataset. This impedes any straightforward 2-dimensional Euclidean distance calculations as it is for instance possible with spatial rainfall observations. Hence metrics to quantify the similarity/dissimilarity of hydrographs can either evaluate the similarity in timing or amplitude, unless a relation between errors in timing and amplitude is established.

Further, the range of possible values differs among the dimensions: while time, loosely spoken, is quasi unbounded (and, with it, timing errors when comparing hydrographs), discharge has a lower limit of zero, which also limits the range of errors: a simulation (please note that henceforth, we will use the term "simulation" as representative of any hydrograph produced by a model, be it a simulation or a forecast), may therefore underestimate the observation by 100% at most (related to the observation), while the range of possible overestimations is basically unlimited. This may be an issue in hydrograph evaluation when considering relative rather than absolute values: to which underestimation does an overestimation of, say, 150% compare?

Looking at hydrographs from a more process-based point of view, it can be regarded as result and expression of a hydrometeorological process chain. As such it possesses characteristics that strongly influence both objective and subjective evaluation: firstly, a hydrograph is intermittent, with distinct rainfall-runoff events separated by periods of low flow. As indicated by the conversation sketched above, in Hydrology often the event is the time scale relevant for evaluation. Secondly, a hydrograph is not time-symmetrical: the shape of the rising and falling limbs of an event look different as they are dominated by different parts of the hydrometeorological causal chain. The first is mainly shaped by the rainfall event, the latter is mainly influenced by catchment properties such as shape, soil and inclination. As a consequence, when comparing hydrographs with a time offset, any metric evaluating amplitude errors at the same points in time possibly compares "apples with pears", i.e. rising with falling limbs (see also Sect. 2.2).

Keeping in mind the key points of the forecaster's discussion and the hydrograph characteristics outlined above, we suggest that a metric suitable to quantify the similarity of two hydrographs should have the right degrees of freedom to adapt to the user's subjective and case-specific perspective on the hydrographs, but in an objective and reproducible way, and it should take into account the special properties of hydrographs based on the knowledge of the underlying physical processes. As such, hydrological time series should neither be regarded as one single time series entity nor as individual records. In our eyes, the best scale of evaluation is the event scale, which lies in between. Or, as Spate et al. (2003) put it, "It seems natural to change the granularity of our (hydrological) time series from days into peaks or events."

It is the aim of this study to propose a new metric to quantify the distance of hydrographs which obeys these specifications. It is termed "Series Distance" and it closely follows subjective reasoning in visual inspection.

The remainder of the paper is structured as follows: in Sect. 2, we discuss established distance metrics for time series from various fields and their applicability to hydrographs. In the same section, we also present standard metrics for hydrograph comparison including visual inspection. In Sect. 3, we introduce the Series Distance method, its underlying assumptions and output. This is followed by an application to both simple synthetic and real-world hydrographs in Sect. 4, along with a discussion of results. Finally, conclusions are drawn and ways forward are discussed in Sect. 5.

2 Distance metrics for time series and their applicability for hydrographs

2.1 Distance metrics for time series – an overview

Time-series analysis has applications in many fields such as stock market, medicine, ecology, signal processing, etc. and a multitude of related metrics has been developed. In the following, we will present some well established methods and discuss their applicability to quantify the similarity of two hydrographs.

2.1.1 Frechet distance

The Frechet distance was introduced by Frechet (1906) and measures the closeness of two time series if stretching and compression in time is allowed, but temporal succession is to be preserved. An intuitive explanation of the Frechet distance is the minimum required length of a leash between a man and a dog, if both may walk along their predefined paths at varying speed including standstill, while walking back is prohibited. A variant of the Frechet distance for discrete time series was presented by Eiter and Mannila (1994). The Frechet distance is very useful when only the occurring events, not their occurring times, are determinant for the proximity evaluation. This explains the great success of Frechet distance in the domain of voice processing. However, as Chouakria-Douzal and Nagabhushan (2006) point out, the Frechet distance may lead to irrelevant results if the temporal interdependence of values is of importance, which is true in the case of hydrographs. As an alternative, they propose a dissimilarity index, which is a weighted combination of the Frechet distance and local temporal trend correlation (mutual rising or falling). While this is an improvement to the original Frechet distance, it is essentially a combination of two independent steps, where global similarity of shape is evaluated by the Frechet distance at the cost of giving up temporal interdependence, and temporal similarity is based on a value by value basis. However, events in a hydrograph are essentially trends of intermediate length, which are not explicitly captured by both components. This makes the use of the dissimilarity index a suitable, but not perfect metric for hydrograph comparison.

2.1.2 Dynamic Time Warping (DTW)

The Dynamic Time Warping (DTW) algorithm (Sakoe and Chiba, 1978) has been used very successfully in speech recognition. The basic assumption is that the shape of the test and the reference series are the same, but one may be stretched or compressed in time (e.g. a word slowly or quickly spoken by a test person and a reference word spoken at normal speed). By non-linear warping (stretching and compression) in time, the amplitude error of the two signals is minimized. The minimized amplitude error is the metric. Comparable to the Frechet distance, the DTW is a good metric to evaluate agreement of shape if temporal considerations play no role. Ouyang et al. (2010) have successfully used it in hydrological data mining to find years with similar discharge patterns from long discharge time series. Here, similarity is mainly defined by similarity in shape, not the timing, which is valid for long-term studies. However, the authors also state that "... the elastic shifting of the time axis loses the information regarding the exact time of the flood peak, which is absolutely critical in flood prediction." DTW is therefore not optimal for direct event-by-event based hydrograph comparison.

2.1.3 Dominant Mode Analysis

Dominant Mode Analysis approximates a series by decomposition with basis functions and evaluates agreement of two time series via agreement of their power spectrum. The two best known approaches are Fourier and Wavelet analysis.

As Schaefli and Zehe (2009) summarize, "... the idea to use Fourier analysis in Hydrology is not new; Whittle (1953) proposed a method for parameter estimation in the Fourier-domain matching the theoretical power-density spectrum of the model to the estimated powerdensity spectrum of the process observations. The Whittle estimator has recently been applied to rainfall-runoff models by Montanari and Toth (2007)." ... "However, as shown by Contreras-Cristán et al. (2006), it can produce unreliable estimates for non-Gaussian processes or show an important loss of efficiency if the autocorrelation of the process is high." Hence the most important drawback of Fourier analysis is that timing aspects of the original series are not retained. This is not the case with Wavelet analysis, which makes it a suitable tool for rainfall-runoff model calibration and performance analysis (Schaefli and Zehe, 2009). The challenge of this method lies mainly in the choice of the similarity measure between the wavelet power spectra and to a lesser degree in the choice of the base wavelet (Schaefli and Zehe, 2009).

2.1.4 Wasserstein Distance (WD)

The Wasserstein Distance (WD) is a robust and intuitive metric to quantify the distance between two probability density distributions. Also known as the Earth Mover's Distance, the WD is the numerical cost of moving one distribution onto the other (with the probability being the mass and the transportation distance in the units of the data). The optimal way for this can be found with a transhipment plan solved by a network simplex algorithm. Among many others, it has found applications in the distance-based analysis of the long-term behaviour of non-linear dynamical systems on the basis of probability distributions derived from time series (Moeckel and Murray, 1997; Muskulus and Verduyn-Lunel, 2011). While this approach is suitable to evaluate if a model has captured the essential behaviour of a dynamical system, it retains temporal aspects which are important in an event-based comparison. However, if one replaces the pdf's with an event, the distance between an observed and simulated event could, after normalization, be calculated in the same way. To our knowledge, this has not been tried for hydrographs yet. The drawback of calculating the distance between two events with the WD is that "apples could be compared with pears", when mass (in this case discharge) would e.g. be moved from a rising to a falling limb.

2.2 Standard metrics for hydrographs

Probably because they were simple, intuitive and straightforward to compute, the first metrics used to evaluate the similarity of hydrographs were either time-aggregated average measures of amplitude error, e.g. the Root Mean Square Error or metrics for timing errors of characteristic points, e.g. the Peak Time Error. A notable recent exception is the Multicomponent Mapping approach proposed by Pappenberger and Beven (2004), where the distance of two hydrographs is measured by the fuzzy degree of membership in boxes placed around one hydrograph which are intersected by the other. This allows simultaneous but not separate consideration of timing and amplitude errors.

As both Root Mean Square Error and Peak Time Error are, despite their known deficits, still widely used in hydrological modelling, their characteristics will be briefly discussed in the following section.

2.2.1 Metrics for errors in amplitude

Arguably the most widely used metrics in hydrograph analysis are amplitude errors and their derivatives, e.g. the Mean Square Error, Root Mean Square Error (RMSE), Nash-Sutcliffe efficiency NSE (Nash and Sutcliffe, 1970) etc.

$$\mathrm{RMSE} = \sqrt{\frac{1}{T} \cdot \sum_{t=1}^{T} (o_t - s_t)^2} \tag{1}$$

A formulation of RMSE for discharge [$\mathrm{m^3\,s^{-1}}$] is given in Eq. (1), where T is the number of steps in a time series [$-$], o_t and s_t are the observation at time step t, respectively [$\mathrm{m^3\,s^{-1}}$]. Its range of values is [0, ∞], with zero being the optimum. The NSE is the RMSE normalized to [$-\infty$, 1] by division with the deviation of the observations from their mean – see Eq. (2), with $\overline{o}$ denoting the mean of (o_1, ..., o_T).

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T} (o_t - S_t)^2}{\sum_{t=1}^{t} (o_t - \overline{o}_t)^2} \tag{2}$$

Here, the optimum value is one. As these metrics are in essence the same, we will discuss their properties only with the example of the RMSE.

Intuitively, amplitude errors and their derivatives are thought to be sensitive mainly to errors in amplitude. However, applied on hydrographs, they show interesting and sometimes non-intuitive characteristics which have been the subject of many studies. As Murphy (1988) and later Gupta et al. (2009) discussed, the NSE (or RMSE) can be regarded as a combination of three criteria, which relate to "... the correlation, the bias, and a measure of relative variability in the simulated and observed values". Consequently, "... optimizing NSE is essentially a search for a balanced solution among the three components..." (quoted from Gupta et al., 2009). The point we want to stress here is that the relative weight of each component is implicitly fixed by the definition of the NSE (or RMSE) and cannot be adjusted according to user needs, which would be possible in a true multiple criteria optimization. Further, using only the NSE (or RMSE) for evaluation or optimization introduces systematic problems such as volume balance errors, undersized variability and a tendency to underestimate large peaks (Gupta et al., 2009). Further, Weglarczyk (1998) reported on interdependencies of the RMSE with other metrics, Krause et al. (2005) compared several, mainly amplitude-based metrics, Legates (1999) described the limits of correlation-based measures such as the RMSE. Along the same lines, Schaefli and Gupta (2007) as well as Jain and Sudheer (2008) found that NSE is a poor metric if the test series show strong seasonality. In this case, even very simple periodical models can produce high values of NSE. McCuen et al. (2006) investigated the influence of sample size, outliers, magnitude bias and time offsets on the NSE, identifying the adverse effect of time offsets and magnitude bias. Summarizing the findings of the above studies, the RMSE and related metrics should not be used by themselves, but only in combination with additional, preferably orthogonal measures and their results should be put in a proper context, e.g. by comparison of the evaluated simulations to benchmarks.

In addition to the findings reported in the literature, we found more characteristics of RMSE related to the interplay of errors in timing and amplitude. We will discuss them with the example of synthetic triangular hydrographs, simple but roughly realistic in shape, as shown in Fig. 1. The "observed event" (bold line) is of arbitrary length 17 h and has a peak of 100 $\mathrm{m^3\,s^{-1}}$. From it, artificial simulations were derived by applying all possible combinations of time offsets in the range [-20, 20] hours and 1-h increments and multiplicative value offsets in the range [0, 2] in increments of 0.1. In Fig. 1, three example simulations are shown. For each combination of time and amplitude offset, we calculated the RMSE and, for reasons of display and comparison, normalized it by the maximum RMSE to [0, 1]. The resulting 2-D surface of errors is shown in Fig. 2. Its main characteristics are:

- Starting from the centre (time and value offset zero), the error increases both with increasing time and value offset. This is in accordance with intuition.

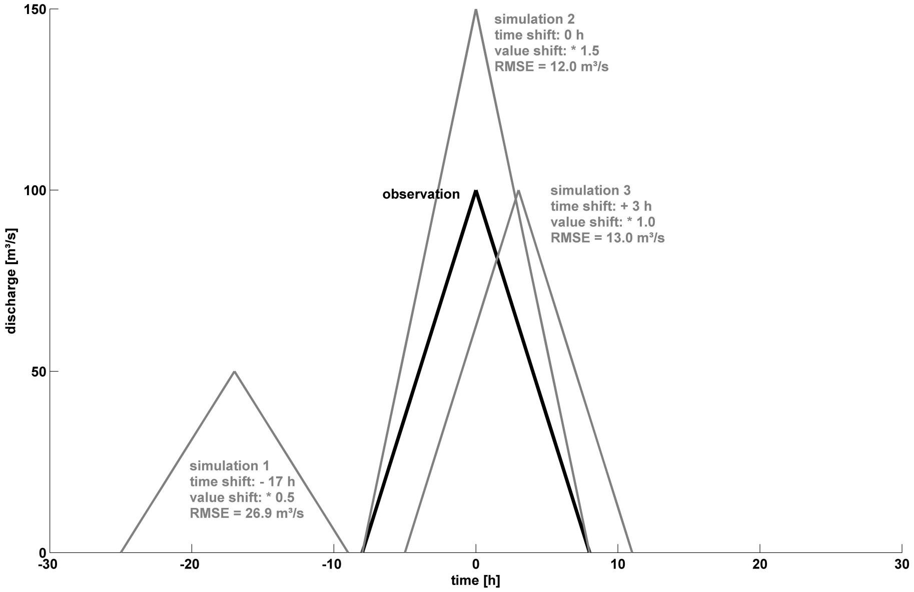

Fig. 1. Synthetic, triangular events. "Observation" (bold line) and three example "simulations" (normal lines) derived from the "observation" by time offsets and multiplicative value offsets.

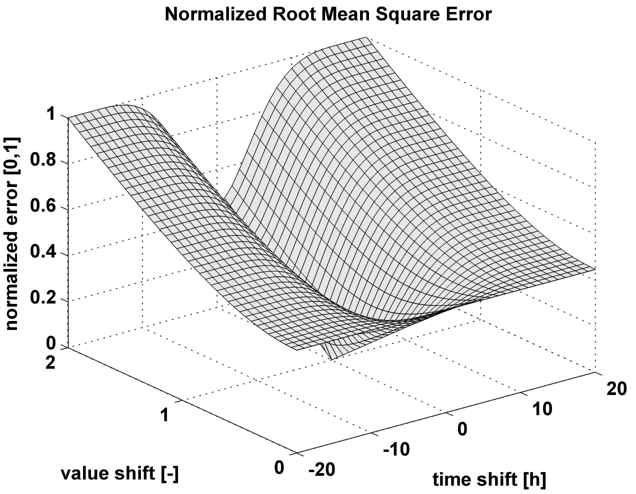

Fig. 2. Error surface of the Root Mean Square Error (RMSE) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [−20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum error.

- Considering time offsets, the error surface is symmetrical to time offset zero, rising steeply at first until, beyond a time offset of around ±10 h, the gradient of the error surface becomes very small and completely levels out at time offsets ≥ ±18 h. Note that symmetry occurs only if either at least one of the two hydrographs (observed and simulated) is time-symmetrical or if they are identical in shape. As can be seen in Fig. 1, simulation 1, a time offset larger than ±18 h completely separates the observed and simulated hydrograph. This means that the RMSE, especially for short, steep hydrographs is strongly sensitive to small time offsets, hardly sensitive to larger offsets and completely insensitive to time offsets larger than the event duration. Note also that for all time offsets, the RMSE compares "apples with pears": first rising with falling limbs, with increasing offset each "event" is more and more compared to zero, i.e. "no event".
- Considering value offsets, the error surface is only symmetrical for time offset zero. With increasing time offset, the error surface becomes more and more asymmetric. This means that a simulation with a time offset, which overestimates the observation by 50%, leads to a much larger RMSE than a simulation with the same time offset but 50% underestimation.
- As for the relation between RMSE values for time and value offsets, the triangular hydrograph as used here, shifted by 3 h (and no value offset), leads to an RMSE value of 13 $m^3\,s^{-1}$. This is comparable to an RMSE of 12 $m^3\,s^{-1}$ for a simulation with a value offset of factor 1.5 and time offset zero (see simulation 2 and 3 in Fig. 1). This relation may or may not be in accordance with the user's subjective weighting, but the point is that it is fixed by the nature of the RMSE calculation and the shape of the hydrograph. And in the author's subjective view, especially in cases of short events with fast rise and recession, RMSE puts too much weight on timing errors compared to errors in amplitude.

2.2.2 Metrics for errors in timing

When comparing two hydrographs, time offsets are easily detected by the examiners eye and strongly influence the process of opinion making. Hence, metrics to quantify timing errors are, after metrics of amplitude errors, also well-known, especially the Peak Time Error. This is the time offset between an observed and the related simulated peak (e.g. Yilmaz et al., 2005). The Mean Absolute Peak Time Error (MAPTE) in unit [h] then is the average of all absolute peak time errors in a hydrograph (see Eq. 3, where N is the number of matching peaks – observed, simulated – in the time series and P_o and P_s are timing of the observed and simulated peaks, respectively).

$$\mathrm{MAPTE} = \frac{1}{N} \cdot \sum_{n=1}^{N} \left| P_{\mathrm{o},n} - P_{\mathrm{s},n} \right| \tag{3}$$

However, peak time metrics are much easier verbalized and applied in visual inspection than formulated and coded, as it requires automated identification of individual events and within the events unique peaks, which may be difficult in case of multi-peak events. Further, once the peaks are found, matching pairs in the observed and simulated hydrograph have to be found. This is usually done by temporal proximity, but this may not always be correct. Hence, metrics for time offsets are less frequently applied than amplitude-based metrics. An elegant solution to this problem is to find the average time offset of the complete hydrograph by maximizing correlation of the observed and the shifted simulated series (e.g. Fenicia et al., 2008). However, this does not consider the event-based nature of hydrographs, where individual events may occur too early and others too late.

Some interesting new approaches were proposed by Lerat et al. (2010), who calculate time offsets not only from event peaks or centroids, but also from comparison of the cumulative volume of two hydrographs and by the phase difference in a cross wavelet approach. Liu et al. (2010) also proposed to estimate timing errors in scale-time space using cross-wavelet transformations, which provides information on scale-dependent time offsets.

For reasons of comparison to the RMSE, we also applied the MAPTE to the synthetic triangular hydrographs and all possible pairs of time and multiplicative value offsets as described in Sect. 2.2.1. The resulting 2-D error surface, again normalized by division with the maximum error to [0, 1], is shown in Fig. 3. Its main characteristics are:

- Its shape is rather simple and resembles a turned ridge roof. As the MAPTE is insensitive to any differences in peak magnitude, the error along the transect at time offset zero is always zero.
- Similar to RMSE, the error surface is symmetrical to time offset zero. But, in contrast, it continuously rises as a linear function of time offset.

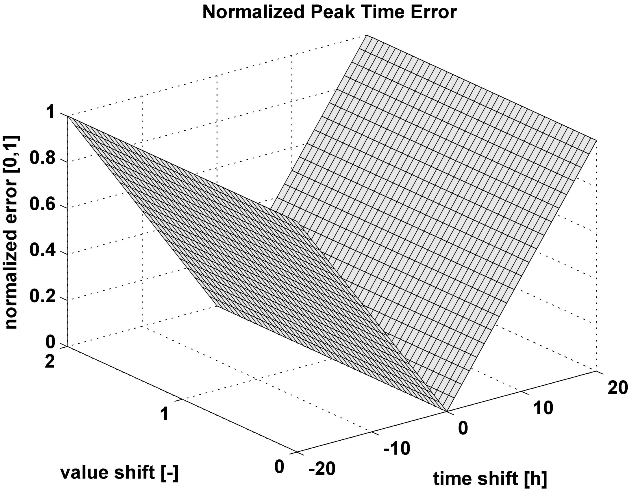

Fig. 3. Error surface of the Mean Absolute Peak Time Error (MAPTE) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [−20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum error.

When comparing the error surfaces for RMSE and MAPTE, it becomes apparent that basically, the directions of largest and smallest gradients are identical. This indicates that when comparing observed and simulated hydrographs with short and steep events and small but present time offsets (which is frequently the case with real-world hydrographs), RMSE and MAPTE are essentially redundant metrics. We tried this also for rectangle-shaped synthetic hydrographs (not shown): the results were less pronounced but essentially the same. This is on one hand unfavourable as errors in amplitude should be distinguishable from errors in timing in order to provide useful feedback for model calibration. On the other hand it supports the findings of Murphy (1988) and Gupta et al. (2009), stating that NSE evaluates not only amplitude errors, but several aspects of a hydrograph.

2.2.3 Visual inspection

Apart from objective metrics, perhaps even more important, is visual inspection and comparison of hydrographs. Eye and brain are a powerful expert system for simultaneous, case-specific multi-criteria evaluation which provides results in close accordance with the user's needs. Due to these obvious advantages, visual inspection is still standard procedure for calibration and validation in engineering practice.

At this point the reader is, before reading on, encouraged to rank the set of example simulations displayed in Fig. 4 by her or his own subjective judgement. The ranking can later be compared to the author's subjective ranking and the result of objective ranking schemes.

However, visual inspection has two major drawbacks: it is subjective and hence irreproducible and it is not applicable

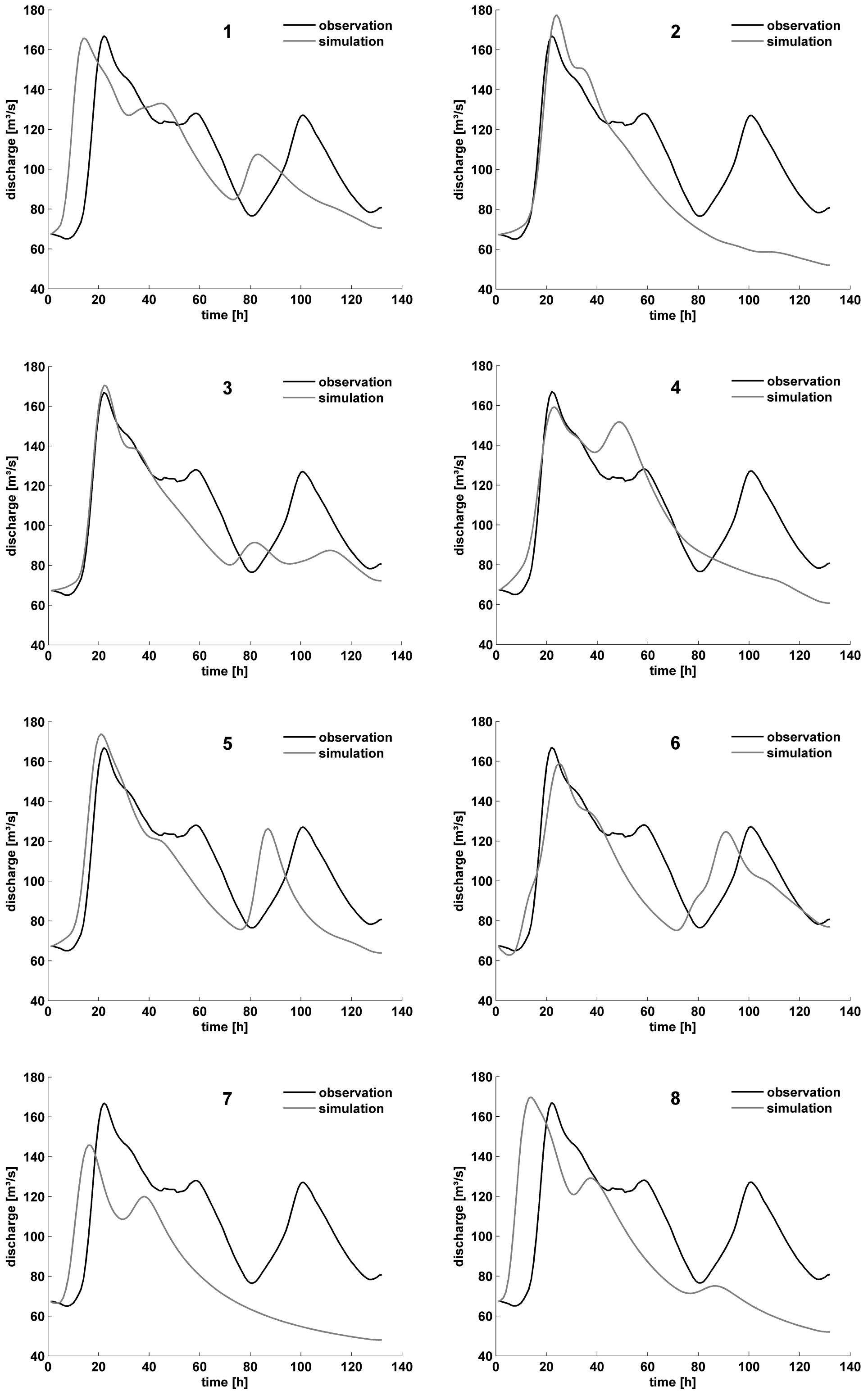

Fig. 4. Observed discharge at gauge Kempten/Iller (954 km^2) for period 21 April 2008 14:00–27 April 2008 00:00 (132 h) and 8 simulations with hydrological model "Fgmod" (Ludwig, 1982) based on Cosmo-Leps ensemble weather forecasts (Marsigli et al., 2005).

on large data sets. In order to overcome this, in recent years several objective metrics were proposed which more closely resemble subjective reasoning in visual inspection (Bastidas et al., 1999; Boyle et al., 2000, 2001). One major step towards this goal was to change the way of looking at a hydrograph, away from considering it merely as a sequence of values towards seeing it as the result of a hydrometeorological process chain, producing distinguishable features such as low flow, events, rising and falling limbs etc. which contain valuable information on both the processes and the models to be evaluated. For instance, Pebesma et al. (2005) evaluated the temporal characteristics of time series of amplitude errors. This concept was further developed by Reusser et al. (2009), who analyzed the temporal dynamics of many metrics applied on hydrographs, clustering them into typical error classes and from this, drew specific conclusions on structural deficits of the underlying models. The same trend away from merely amplitude-based scores towards more intuitive, feature-based comparison can be noticed in the atmospheric sciences: Ebert (2008) proposed Fuzzy and neighbourhood-based approaches to account for approximate agreement; Casati et al. (2004) used scale-decomposition techniques to isolate physical features such as large-scale frontal systems of small-scale convective showers. Davis et al. (2006) used object-based techniques to compare identifiable objects such as rain cells, Keil and Craig (2009) used field verification techniques for the same purpose.

These approaches not only represent the trend of looking at data (in our case hydrographs) in a more process-based way, but also the move from single- towards multi-objective evaluation. Much work has been done in this field in recent years, and both new metrics (e.g. Dawson et al., 2007, 2010) as well as ways to jointly evaluate them have been proposed, e.g. Taylor (2001), Yapo et al. (1998), Gupta et al. (1998), van Griensven and Bauwen (2003). Applications of multi-objective calibration are manifold (e.g. Beldring, 2002); however the metrics applied are still mainly of the amplitude-error type. Recently, Gupta et al. (2008) proposed a step beyond multi-objective evaluation towards diagnostic, behavioural evaluation of catchment/process signature indices. The concept has been applied by Yilmaz et al. (2008), using three behavioural functions: water balance, vertical and temporal water redistribution. Other steps towards multi-objective evaluation with hard and soft information have been proposed by Winsemius et al. (2009).

3 The metric "Series Distance (SD)"

The Series Distance (SD) was developed with the aim to closely reflect subjective reasoning in visual hydrograph inspection. In our view, this is mainly characterised by the following points:

- A hydrograph is the result and expression of a hydrometeorological process chain and as such, individual events, separated by periods of low flow are distinguished and considered individually.
- Each event is composed of characteristic features, namely peaks, troughs, and segments of rise or recession.
- When comparing observed and simulated hydrographs, only matching events and matching segments within them are compared. There may be events, simulated or observed, that have no match.
- Subjective evaluation of an event is typically done by complete comparison of matching segments, simultaneously but separately for errors in amplitude and timing. A typical linguistic evaluation could be: "The simulated flood rise is too early and too steep and the peak too high, the falling limb drops too slowly and lasts too long". The resulting synoptic evaluation compares the overall shape of the hydrographs. This is in our eyes superior to the approach proposed by Perng et al. (2000), who uses patterns of single characteristic landmarks such as peaks or troughs for time series comparison.
- Each user weighs errors in amplitude and timing differently, depending on the intended use of the simulation. For example in flood forecasting, a person operating a small flood-retention basin is dependent on accurate peak timing, while a person responsible for dike defence is more interested in maximum water levels.
- The overall comparison of an observed and simulated hydrograph includes the following components: did the simulation produce matches of all observed events, or were there missing or false events? Did the overall shape of the matching events agree with respect to timing and amplitude? These individual components may point towards different sources of error (poor input data, deficits in different parts of the underlying model structure, etc.). It is therefore useful to also allow their separate, non-aggregate evaluation.

As the SD aims to consider all these points, a precondition for its use is that the investigated hydrograph pairs (i) contain events and (ii) have at least something to do with each other in the sense that they are to a certain degree correlated and that observed and simulated events can be related. If this is not the case, e.g. for long spells of low flow, an event-based comparison is not useful and other measures such as simple amplitude metrics can and should be applied.

3.1 Procedure

The SD is not a single metric based on a single formula; it is rather a procedure which allows a combined determination of

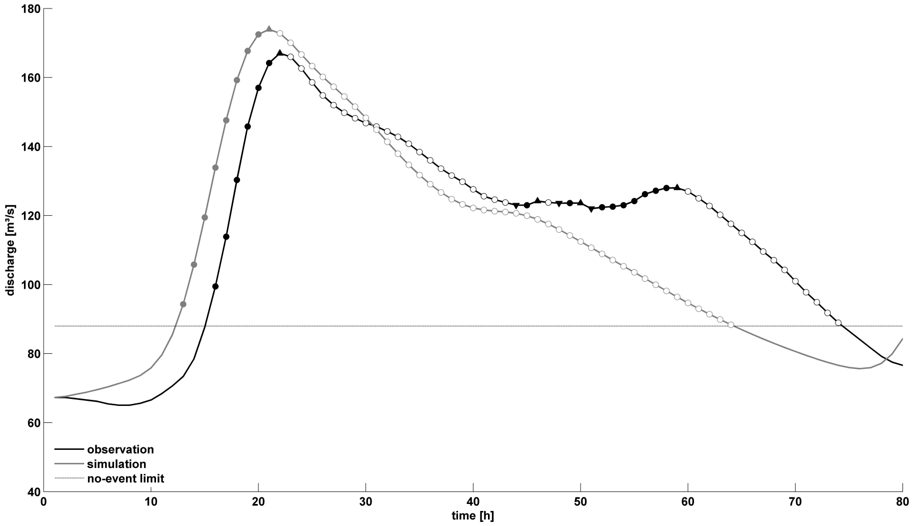

Fig. 5. Example of a matching observed (black) and simulated (grey) event (detail of event 5 in Fig. 4). The hydrological case is shown for each point: "rise" (filled circle), "peak" (upward triangle), "recession" (empty circle), "trough" (downward triangle), "no event" (no marker). The "no-event" threshold (thin grey line) separating events from low flow conditions is set to 88 $\mathrm{m}^3\,\mathrm{s}^{-1}$.

how many of the observed and simulated events match and how the matching events differ with respect to timing and amplitude. It consists of the following steps:

- Identify events: from the hydrograph, individual events are identified by applying a user-defined parameter termed "no-event threshold" [$\mathrm{m}^3\,\mathrm{s}^{-1}$]. In its simplest form, this is a constant discharge threshold separating baseflow conditions from an event. More elaborate baseflow separation techniques are of course possible. Each event starts with an upward and ends with a downward crossing of the "no-event" threshold. In the example hydrograph shown in Fig. 5, the threshold was set to 88 $\mathrm{m}^3\,\mathrm{s}^{-1}$.

- Match events: in order to relate events in the observed and simulated hydrograph, a parameter termed "match limit" [h] is applied. This is a time offset separating matching from non-matching events. Two events are considered matching, if the end of the earlier and the start of the later are no longer apart than the match limit. Hence, in an observed and simulated hydrograph, there can, following the nomenclature used for contingency tables, be matching events ("hits"), observed events with no match ("misses") and simulated events with no match ("false events"). Only 1:1 relations are allowed, i.e. in the case of two simulated events matching one observed (or vice versa), the relation is only established between the pair with larger overlap. "Match limit" can assume negative or positive values, usually it is set to zero. For more detailed information on the matching algorithm, see the pseudo code in Appendix A1. In Fig. 5, with match limit set to zero, the two events were considered matching. In simulations based on observed forcing, events usually match. Simulations based on weather forecasts however, especially long-term forecasts in small catchments, may contain misses or false events.

- Assign hydrological cases: each point of the observed and simulated hydrograph is assigned one of the following hydrological cases, defined by the sequence of gradients from the previous to the current and from the current to the next point: "rise" (positive–positive), "peak" (positive–negative), "recession" (negative–negative), "trough" (negative-positive). In addition, all points below the no-event threshold are labelled "no event". Ensuring meaningful assignments usually requires pre-processing of the time series:

  - Smoothing: peaks and troughs mark important turning points in the hydrograph. In order to capture only the relevant peaks and troughs by the gradient-based approach, and not just small fluctuations (possibly caused by the manner of observation), the latter should be removed, e.g. by a moving average filter.

  - Avoid equal values: sequences of equal values sometimes occur under low-flow conditions, corrupt data or human impact (e.g. weir operation). As this obviates unique determination of hydrological cases, we modify them in a very simple manner: each value in the sequence is raised by 1/1000 of its precursor. The impact of this modification on the overall result is in most cases negligible.

For more detailed information on the algorithm, see the pseudo code in Appendix A2. In Fig. 5, each point of the observed and simulated hydrograph is marked with its hydrological case. An event invariably consists of the sequence of components shown in Eq. (4), where $x_i \epsilon$ [0, ∞].

$$\text{start}, \; x_1 \cdot \text{rise}, \; x_2 \cdot (\text{peak}, \; x_3 \cdot \text{decline}, \; \text{through}, \; x_4 \cdot \text{rise}), \; \text{peak}, \; x_5 \cdot \text{decline}, \; \text{end} \tag{4}$$

This means that in the simplest case, an event consists of a start, a peak and an end (x_1, x_2, x_3, x_4, $x_5 =$ zero). Note that the sequence of peaks and troughs alternates and that it always starts and ends with a peak. Hence, there is always one more peak than the number of troughs.

- Attune matching events: Although the principal order and relative frequency of peaks and troughs is predetermined, the absolute number can differ between matching observed and simulated events. For example in Fig. 5, there are 4 peaks and 3 troughs in the observed event, and only 1 peak and no trough in the simulated. However, in order to calculate the distance between the observed and simulated event (explained below), the number of peaks and troughs in the observed and simulated event must be equal. This is achieved by eliminating the less relevant peaks and troughs in the event with the higher number of turning points:
  - In the event, find the sequence of peak$_n$/trough$_n$/ peak$_{n+1}$ where the amplitude difference calculated as (peak$_n$ − trough$_n$) + (peak$_{n+1}$ − trough$_n$) is minimal. In other words, this is the least pronounced "dent" in the event.
  - From this sequence, erase the trough and the smaller (less important) of the two peaks. "Erase" here does not mean that the point are removed, but their hydrological case is changed to "rise" or "recession", depending on the neighbouring points.
  - This is repeated until the number of turning points in the observed and simulated event is equalized.
  - Having thus ensured that each segment of the observed event finds its counterpart in the simulated event, the distance calculation is done in a loop over all segments.
  - Note that for misses and false events, this procedure is not required.

For more detailed information on the attuning algorithm, see the pseudo code in Appendix A3. In the example shown in Fig. 5, this procedure removes the last three peaks and troughs from the observed hydrograph. This is in accordance with visual inspection, as the dominant peak at the beginning of the event is maintained.

- Distance calculation for matching events: Having ensured that the number of peaks and troughs (and with it, the number of rising and falling segments) is attuned, the distance between matching segments can be calculated. *This is the core of the Series Distance procedure*. The idea is that the shape of each observed segment, expressed by the number of points and their respective time and amplitude values, is the reference, against which the matching simulated segment is compared. As the simulated segment may be longer or shorter than the observed, 1:1 mapping of observed and simulated points is usually not possible. To overcome this, the simulated segment is considered as a polygon line. From this, applying linear interpolation, points are sampled with equal temporal spacing, the number being equal to the number of points in the observed segment. With this, each point in the observed segment can be assigned a point in the simulated segment. Now for each pair of points the offset in time and amplitude can be calculated. For more detailed information on the distance algorithm, see the pseudo code in Appendix A5. The advantage is thus that (i) only matching segments are compared, (ii) not single points (e.g. peaks) are used to calculate the distance, but complete segments are scanned, (iii) the relative contribution/importance of each segment to the overall event is determined by the length of the observed segment, (iv) matching points are found in a way comparable to visual inspection and (v) timing and amplitude errors are calculated between the same pairs of points, simultaneously but separately. To illustrate this, connecting lines between matching points are shown in Fig. 6. The small inserted figure reveals that the observed points in a segment do not necessarily match with a simulated point, but with a point on the polygon line representing the simulation, located at the same fraction of overall segment length.

- Distance calculation for non-matching events: in the case of misses and false events, there is no matching event available for comparison. Consequently, there is neither a timing error nor an amplitude error that can be calculated from them. This may seem non-intuitive at first, as misses and false events are most unfavourable and should therefore strongly affect any metric. In fact, their influence is accounted for by the third component of the Series Distance, a contingency table (see also Sect. 3.2). The advantage of this procedure is that three basically independent characteristics of agreement between two hydrographs (do the features match? is the timing of the matching features correct? is the magnitude of the matching features comparable?) are treated separately. With a suitable weight of the contingency table in a final combined evaluation of the three metrics, misses and false events can be considered appropriately.

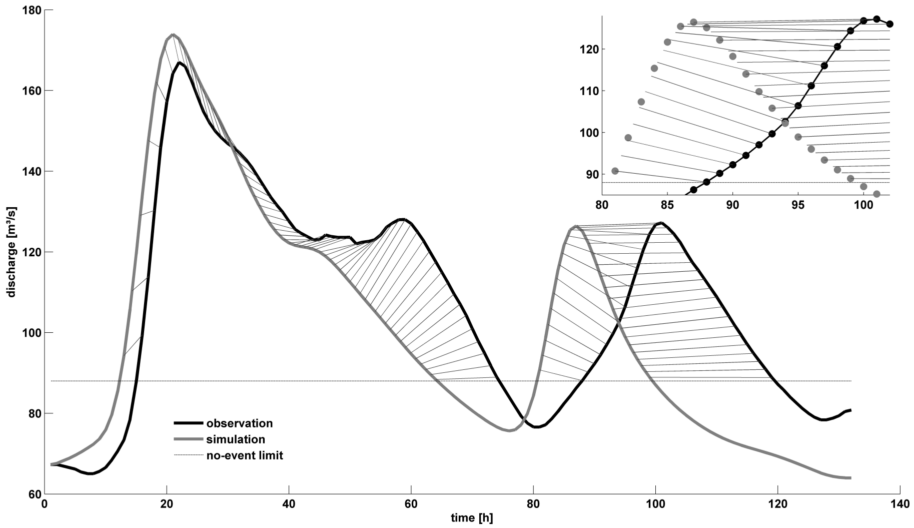

Fig. 6. Example of a matching observed (black) and simulated (grey) event (event 5 in Fig. 4). Connections (thin grey lines) between matching points of observation and simulation according to the Series Distance procedure are shown. The small inserted figure reveals that the observed points in a segment (rise or recession) do not necessarily match with a simulated point, but with a point on a polygon line representing the simulation at the same fraction of overall segment duration.

- Distance calculation for low flow periods: As the Series Distance focuses on comparison of events, neither time nor value errors are calculated for values below the no-event limit.

- Altogether, the SD procedure has three free parameters, namely the "no-event" threshold [$m^3\,s^{-1}$], the match limit [h] and the manner of the smoothing.

3.2 Output

Based on the identification of events in the observed and simulated hydrograph and the distances in magnitude and timing, calculated for all matching point pairs as described in Sect. 3.1, a number of metrics can be calculated:

- Contingency table: the frequency of matching, missing and false events can be listed in a contingency table as shown in Table 1. This provides useful information on the overall agreement of simulated and observed events. Note that here the number of correct negatives, i.e. occasions where both the observation and simulation show no event, cannot be calculated as this would require the definition of a typical period of time for evaluation (in weather forecasting, this is typically the aggregation time of interest, e.g. 12 h). However, as the SD is intended to evaluate the agreement of events, this is in our eyes no substantial drawback.

- Threat Score: the information in the contingency table can be further condensed to the well known Threat Score or Critical Success Index (Donaldson et al., 1975) as shown in Eq. (5). Ranging from zero to one, a Threat Score of one indicates optimal reproduction of events. For the definition of hits, misses and false alarms see Table 1.

$$\text{Threat Score} = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}} \tag{5}$$

- Mean Absolute Amplitude and Timing error: from the set of amplitude and timing errors (all point pairs in all segments in all matching events), standard aggregate metrics such as the mean, mean absolute or mean squared error can be calculated. In this work, we applied the Mean Absolute Error both for timing (SD_t) and value (SD_v) for the following reasons: firstly, taking the absolute value avoids cancellation of positive and negative errors. Secondly, we used the simple (i.e. non-squared) distance, as the goal of the Series Distance is to evaluate overall agreement rather than amplifying individual gross errors. SD_t and SD_v are also displayed in Eqs. (6) and (7), respectively, with M being the number of time steps within all observed events that have a matching simulated event. Dist_t and Dist_v are the differences between matching observations and simulations, respectively, as explained in Appendix A4.

$$\text{SD}_t = \frac{1}{M} \cdot \sum_{m=1}^{M} |\text{Dist\_t(m)}| \tag{6}$$

$$\text{SD}_\text{v} = \frac{1}{M} \cdot \sum_{m=1}^{M} |\text{Dist\_v(m)}| \tag{7}$$

Table 1. Contingency table.

		observation	
		> threshold	≤ threshold
simulation	> threshold	hits	false alarms
	≤ threshold	misses	correct negatives

- Many other metrics can be derived from the Series Distance procedure, e.g. scatter plots of timing error vs. amplitude error, which potentially allows insight into typical error combinations useful for deficit analysis of the underlying models. This could be further refined by doing the analysis separately for each hydrological case.

Applied in the manner as proposed above, the Series Distance procedure yields three metrics, namely the Threat Score, the $\mathrm{SD_v}$ and the SD_t. They are essentially non-redundant, as the first evaluates agreement in overall event occurrence, the second agreement in amplitude and the last agreement in timing and as such, they can be evaluated separately. For tasks such as automated model optimization however, a single metric may be desirable. In this case the three metrics can be combined to one, using some kind of weighted combination function. The choice of this function and the relative metric weights of course introduces a subjective element in the evaluation procedure. However, as discussed above, each user weighs errors in event occurrence, amplitude and timing differently, depending on the intended use of the simulation. In contrast to visual inspection, where the weighted combination is carried out in an irreproducible way, the application of a combination function is objective and reproducible while still giving the user the freedom of customizing it according to her or his subjective needs.

3.3 Alternatives

Development of the SD procedure as described in Sects. 3.1 and 3.2 was a matter of trial and error and frequently ended in dead ends. As we think that much can be learned from going astray, we will now present a line of thought we tested and abandoned.

Seeking a way to compare hydrographs in a more holistic manner, it was tempting to establish a relation between errors in amplitude and timing at the very beginning of the SD procedure. This can be done either in a subjective, user-specific manner by formulating a direct relation (e.g. "an error in timing of one hour is equivalent to an error in magnitude of ±10%"), or it can be done in the form of an objective relation based on hydrograph characteristics (e.g. for each event, the difference of peak and lower threshold is considered as 100% error in amplitude, while a time offset equal to the event length is considered 100% error in timing). Thus transforming both errors to dimensionless units allows 2-D distance calculations in the transformed time-amplitude space. With this, matching points on the observed and simulated hydrograph are simply those that are closest to each other, given that they are of the same hydrological case. The 2-D point distances can then simply be added to the overall Series Distance. This approach, however, had two major disadvantages. Firstly, it may lead to non-intuitive sets of point pairs as complete scanning of each segment is not assured. For instance, if a simulated flood rise severely underestimates the observed rise, for most points on the simulated hydrograph the closest points will be found in the lower part of the observed hydrograph, leaving the upper part completely unconsidered. Secondly, while on one hand combining errors in time and amplitude from the beginning is attractive as it allows direct computation of a single metric, on the other hand it means a loss of information which can be drawn from the relative contributions and correlations of errors in timing and amplitude.

Although this line of thought is no longer pursued at the moment, it may at a later time be interesting to relate (i.e. normalize) the components of the Series Distance to characteristic features of the hydrograph under consideration, such as mean event duration, mean event distance, distribution of discharge values, etc. Thus transforming the errors to dimensionless numbers would facilitate combination to a single metric and make their relative weighting more objective. Also, it would facilitate comparison of metrics among hydrographs from different sites with different characteristics (e.g. hydrographs from alpine catchments with short, intensive events or hydrographs from large lowland catchments with drawn-out, smooth events).

4 Application, results and discussion

In this section, we apply the Series Distance both to artificial and realistic hydrographs in order to evaluate its behaviour under different conditions and to compare its results both to standard metrics (RMSE and Mean Absolute Peak Time Error) and visual inspection.

4.1 Application on a synthetic hydrograph

Similar to the discussion of the RMSE and MAPTE characteristics in Sects. 2.2.1 and 2.2.2, respectively, we first applied the SD procedure to the synthetic triangular hydrographs shown in Fig. 1. Each "simulated" event is simply derived from the "observed" event by an offset in time and a multiplicative offset in amplitude. As with RMSE and MAPTE, we calculated the $\mathrm{SD_v}$ and SD_t for all offset combinations in the range of [−20, 20] hours and multiplicative value offsets in the range [0, 2]. The free SD parameters were set to the following values: match limit = 0 h,

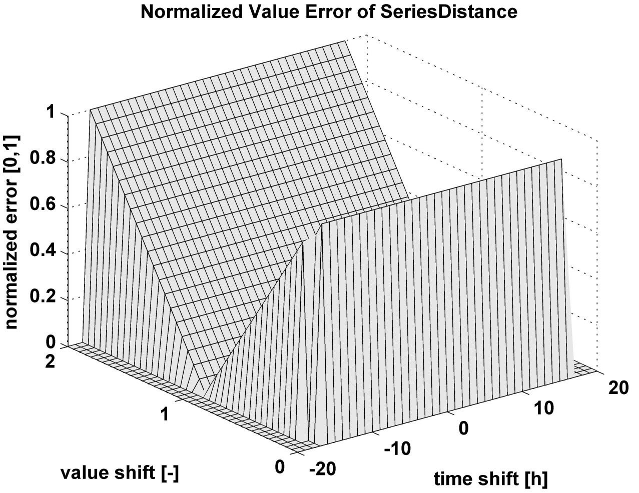

Fig. 7. Error surface of the value/amplitude error of the Series Distance (SD) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [-20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum error.

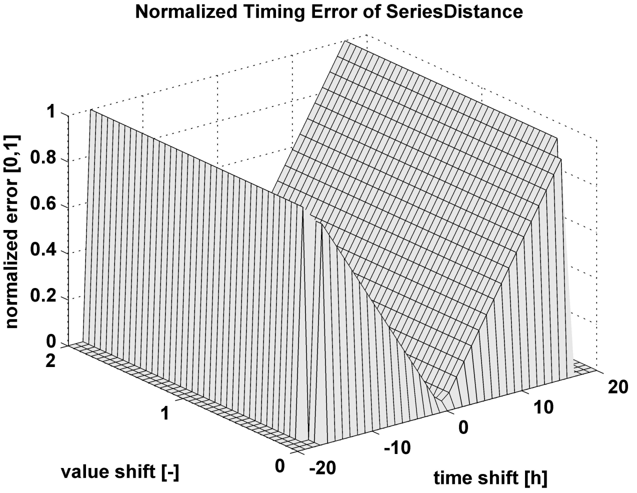

Fig. 8. Error surface of the timing error of the Series Distance (SD) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [-20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum error.

"no-event" threshold = 1.9 $\mathrm{m^3\,s^{-1}}$, smoothing = none. With the "observed" values ranging from 0 to 100 and an event length of 17 h, time shifts ≥18 h lead to non-matching events. The contingency table here simply contains one "hit" for time offsets smaller than 18 h and one "miss" and one "false alarm" beyond. With the event threshold set to a very low value, even strongly downsized simulations are still above the threshold and thus considered as events. The resulting 2-D surfaces of error for $\mathrm{SD_v}$ and SD_t are shown in Figs. 7 and 8, respectively, again normalized by division with the maximum error to [0, 1]. Their main characteristics, especially in comparison to those of RMSE and MAPTE are:

- Both surfaces resemble a turned ridge roof, but in contrast to RMSE and MAPTE, the (turned) ridges point in different directions: $\mathrm{SD_v}$ is sensitive to amplitude offsets only, while SD_t is sensitive to time offsets only. Both error surfaces are symmetrical to the respective ridge (amplitude offset one and time offset zero, respectively) and, unlike RMSE, rise linearly. This means that the two metrics are basically orthogonal, which makes them suitable for joint, non-redundant evaluation.
- For time offsets beyond the matching limit (≥18 h), both $\mathrm{SD_v}$ and SD_t drop to zero, as for non-matching events, no distances are calculated (see Sect. 3.1). The disagreement of the observed and simulated hydrograph is in this case captured in the contingency table.

4.2 Application on realistic hydrographs

Finally, we applied the SD procedure to eight realistic pairs of observed and modelled hydrographs as shown in Fig. 4. The observed hydrograph is from the Kempten gauge on the river Iller (Germany), which drains an alpine catchment of 954 $\mathrm{km^2}$. The discharge was observed during a small 5-day flood event from 21–27 April 2008. The related modelled hydrographs are based on forecasts from an operational, conceptional flood forecasting model based on Larsim (Ludwig, 1982; Ludwig and Bremicker, 2006), driven by Cosmo-Leps ensemble weather forecasts (Marsigli et al., 2005), which are widely used in operational hydrological forecasting . We chose an ensemble forecast as with this, a number of different modelled hydrographs are available which are all related to the same observed hydrograph. This facilitates performance comparisons among the simulations and allows ranking. However, this does not mean that the Series Distance is only applicable on hydrological forecasts; the hydrograph ensemble might just as well have been a set of simulations based on different model parameter sets in a calibration procedure.

As the model application is not of central interest here, for the sake of brevity we are not going into greater detail on the model setup. We also did not use the hydrographs as produced by the hydrological model directly, but modified them slightly. We did so because the aim of this study is to present and analyse the behaviour of SD for a variety of hydrograph pairs with different characteristics such as overestimation, timing errors, matching and missing events, etc. This is hard to find in a single forecast ensemble. The modifications we carried out were small changes in magnitude (of the order of ±10%) or timing (of the order of ±5 h). However, care was taken that the resulting hydrographs remained realistic.

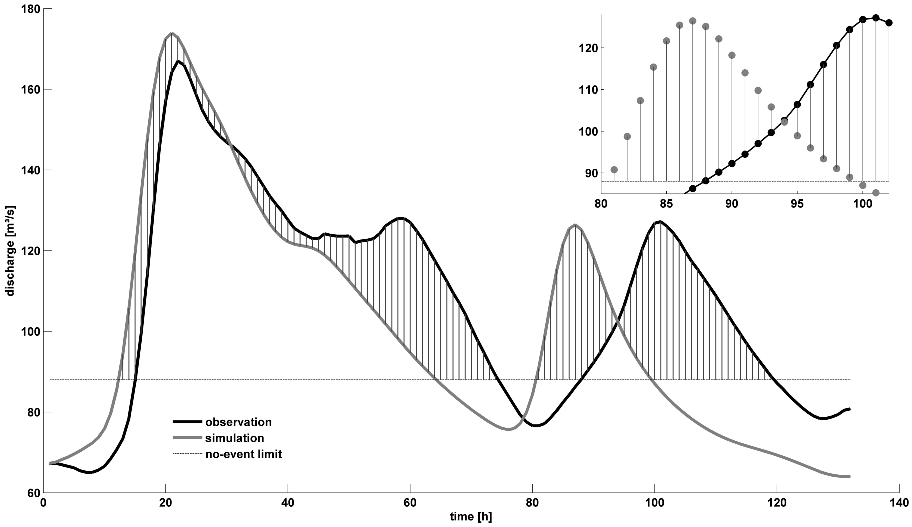

Fig. 9. Example of a matching observed (black) and simulated (grey) event (event 5 in Fig. 4). Connections (thin grey lines) between matching points of observation and simulation according to the RMSE are shown. Note that connections may exist between non-matching segments of the hydrographs (rise with recession or vice versa).

In order to apply the Series Distance, its free parameters were set to the following values: match limit = 0 h, "no-event" threshold = 88 $\mathrm{m^3\,s^{-1}}$ (see e.g. Fig. 5), smoothing = 5 h moving average. Note that we deliberately omitted the threshold from Fig. 4 to avoid biasing the reader's own subjective evaluation and ranking.

For comparison, we also calculated the RMSE and MAPTE for all eight events. In order to base them on the same dataset as the Series Distance metrics, RMSE was also only calculated for values above the "no-event" threshold (i.e. low flow was omitted) and the Mean Absolute Peak Time Error was only calculated between peaks of events that were considered matching by the SD procedure.

The observed and simulated hydrographs for event 5 are shown in Figs. 6 and 9. In addition, connection lines between related points (i.e. the point pairs used for distance calculations) on the two time series are shown in Fig. 6 according to the SD procedure and in Fig. 9 as used by the RMSE. While in both cases points below the "no-event" threshold are neglected, there are obvious differences for the points above: RMSE relates points with equal position in time, while SD relates points at equal relative position in matching segments of matching events. In our view, the latter is in closer accordance with intuition than the first. For example, the detailed subplot in Fig. 9 reveals that between time steps 88 and 99, RMSE is calculated between non-matching parts of the hydrographs: the simulation already recedes while the observation still rises. Another example is the first steep flood rise at time steps 15 to 20. Here, the simulated hydrograph closely resembles the observed one, but runs ahead for about two hours. The resulting point pairs for RMSE are far apart with respect to amplitude, which results in large values of RMSE, while a user might consider the simulation as relatively good, despite the time offset. In our opinion, the distance between the hydrographs is in this case better represented by the point pairs of SD as shown in Fig. 6. They also have the advantage that both the errors in amplitude and timing are calculated on the same point pairs, simultaneously but separately. In contrast, the MAPTE is calculated only on a single pair of points.

Table 2. Metrics for 8 pairs of simulated and observed hydrographs as shown in Fig. 4. RMSE = Root Mean Square Error, MAPTE = Mean Absolute Peak Time Error, $\mathrm{SD_v}$ = Amplitude Error of Series Distance, SD_t = Timing Error of Series Distance.

Sim #	RMSE [$\mathrm{m^3\,s^{-1}}$]	MAPTE [h]	Threat Score [−]	$\mathrm{SD_v}$ [$\mathrm{m^3\,s^{-1}}$]	SD_t [h]
1	22.2	13.0	1.0	6.7	13.8
2	15.5	2.0	0.5	18.1	12.1
3	15.2	0.0	0.3	7.5	4.6
4	14.0	1.0	0.5	10.3	5.5
5	17.9	7.5	1.0	5.8	8.4
6	15.8	6.5	1.0	6.8	6.5
7	24.1	6.0	0.5	10.6	15.5
8	25.8	8.0	0.5	5.0	15.6

All metrics (RMSE, MAPTE, Threat Score, $\mathrm{SD_v}$ and SD_t) for each of the eight simulations are shown in Table 2. Irrespective of whether the eight simulations stand for a set of ensemble forecasts or a set of simulations in a parameter optimization process, the task is the same: to evaluate them according to their performance and then select the best (or the

Table 3. Ranked metrics from Table 2 for 8 pairs of simulated and observed hydrographs as shown in Fig. 4. Ranks are determined separately for each column. Highest ranks are shaded grey. RMSE = Root Mean Square Error, MAPTE = Mean Absolute Peak Time Error, I and II = ranks of columns I and II added and ranked, SD$_v$ = Amplitude Error of Series Distance, SD$_t$ = Timing Error of Series Distance, V and VI = ranks of columns V and VI added and ranked, IV and VII = ranks of columns IV and VII added and ranked, Subjective = subjective classification by the authors, Rank Diff = Accumulated rank difference between subjective ranking (column IX) and the ranks in the respective column.

Sim #	RMSE	MAP-TE	I and II	Threat Score	SD$_{\mathrm{v}}$	SD$_t$	V and VI	IV and VII	Subjective
	I	II	III	IV	V	VI	VII	VIII	IX
1	6	8	7	2	3	6	5.5	3	3
2	3	3	3	5.5	8	5	7	7	6
3	2	1	1.5	8	5	1	1.5	4.5	4
4	1	2	1.5	5.5	6	2	4	4.5	5
5	5	6	5.5	2	2	4	1.5	1	1
6	4	5	4	2	4	3	3	2	2
7	7	4	5.5	5.5	7	7	8	8	8
8	8	7	8	5.5	1	8	5.5	6	7
Rank Diff	20	26	23	11	14	16	10	3	0

best few). This is no problem if single metrics are used, but if several metrics with different units are jointly considered the problem of unit mixing and of assigning relative weights to individual metrics occurs. The first can, for example, be overcome by transforming values to relative ranks within the set while the latter requires a (subjective) fixing of weights by the user. With respect to the first problem, in this study we used a simple ranking transformation: for each metric, the relative rank of each simulation is shown in Table 3.

In addition to ranking the individual metrics (columns I, II, IV, V, and VI), we also calculated the ranks of combined metrics. First, we combined RMSE and MAPTE, giving equal weights to each of them. To this end, the ranks of RMSE and MAPTE for each simulation were added and the resulting sums ranked again (see column III). It is noteworthy that for the set of simulations presented in this study, both RMSE and MAPTE lead to rather similar ranking orders: hydrographs three and four (both with small timing errors for the main event, but almost completely missing the secondary event) were placed at the top, hydrograph five (both events reproduced in the correct order of magnitude but with a timing error) was placed in the lower half. As a consequence of the similar ranks, the combined ranking is comparable to the ranking of the individual metrics.

Moreover, we merged the two SD distance metrics: in column VII, the ranks of SD$_{\mathrm{v}}$ and SD$_t$ were combined in the same manner as RMSE and MAPTE. In contrast to RMSE and MAPTE, however, the rankings of the two SD distances are dissimilar. For example, hydrograph eight was ranked best by the SD$_{\mathrm{v}}$ and worst by the SD$_t$. In that case, the matching simulated and observed hydrographs were similar in shape and amplitude, but offset by a large time shift. Note that for hydrograph eight, SD identified only one matching event: the secondary observed event found no match. Consequently, the Threat Score was low (rank 5.5 in column IV, row "8"). In contrast to this, in hydrograph one (where simulation and observation of the main event are also similar in amplitude and offset in time), the secondary observed event matches a simulated one. This results in a high rank for the Threat Score. Ranks for SD$_{\mathrm{v}}$ were lower, though, as the matching simulation underestimated the observed secondary event.

Also, all three SD metrics were combined in column VIII by adding the (weighted) ranks of Threat Score, SD$_{\mathrm{v}}$ and SD$_t$. We (subjectively) chose the following relative weights: as principal agreement of the hydrographs (expressed by the Threat Score) was considered to be most important, we gave it a weight of 50%. SD$_{\mathrm{v}}$ and SD$_t$ ranks were equally weighted with 25%, respectively.

Finally, the author's subjective ranking of the eight test hydrographs is also shown in Table 3, column IX. During the underlying visual hydrograph inspection, we followed the general guidelines discussed in Sect. 3. The resulting ranks are of course highly subjective and may or may not be in accordance with the reader's ranking, nevertheless we compared the agreement of the rankings based on the objective metrics (columns I–VIII) with the subjective ranking by calculating the Sum of Absolute Rank Errors. This is simply the sum of absolute deviations from the subjective ranks, accumulated for all eight hydrographs, separately for each objective metric. The magnitude of the Rank Error expresses the degree of agreement between the objective and the subjective ranking scheme: the smaller it is, the better the agreement. The results are shown in the last line of Table 3 ("Rank Diff"). Comparing the Rank Errors for the different metrics reveals several interesting points:

- Combining RMSE and MAPTE results in a Rank Error of 23. This is in between those of the two metrics evaluated separately. It seems that in the example presented here, combining the two did not improve much the overall closeness to subjective classification.

- The Threat Score seems to be a good metric to mimic visual inspection: without combination with other metrics it has a Rank Error of only 11, which is the third-best from the tested eight metrics. It should be noted, though, that it is only useful for simulations or forecasts, where substantial numbers of false alarms or misses really occur (see also Sect. 3.1).

- In contrast to RMSE and MAPTE, combination of the SD metrics continually improves the agreement with subjective classification: while $\mathrm{SD_v}$ and SD_t taken separately still show relatively weak agreement (although better than for RMSE or MAPTE), a combination of the two leads to a Rank Error of only 10 (column VII).

- Finally, combining the Threat Score, $\mathrm{SD_v}$ and SD_t (column VIII) leads to the smallest Rank Error of only 3. This suggests that this final combination constitutes a metric reflecting visual inspection relatively closely. Further, it seems that the Threat Score and the combined $\mathrm{SD_v}$ and SD_t are essentially non-redundant information, as their combination decreased the Rank Error substantially.

5 Summary and conclusions

In this paper, we proposed a new metric to quantify the similarity of hydrographs. Termed Series Distance, it is aimed to reproduce the advantages of visual inspection, namely simultaneous, case-specific multi-criteria evaluation, but in an objective manner. The Series Distance quantifies the similarity of two hydrographs on the scale of hydrological events. It consists of three parts, namely a Threat Score which evaluates overall agreement of event occurrence, and the overall distance of matching observed and simulated events with respect to amplitude and timing. Within matching events, point pairs on the observed and simulated hydrographs for distance calculation are identified by the same relative position in matching segments (rise or recession) of the event, indicating the same underlying hydrological condition. Thus, amplitude and timing errors are calculated simultaneously but separately, from point pairs that also match visually, considering complete events rather than only individual points (as is the case with Peak Time Errors). Relative weights can be freely assigned to each component of the Series Distance, which allows (subjective) customization of the metric to various fields of application in a traceable way. Each of the three components of the Series Distance components can be used in an aggregated or non-aggregated way, which makes the Series Distance a suitable tool for differentiated, process-based model diagnostics.

For the example of simple, triangular hydrographs we demonstrated that the resulting Mean Absolute Errors in Timing and Amplitude are less redundant than the Root Mean Square Error and the Mean Absolute Peak Time Error, two metrics commonly used in hydrograph evaluation. Applied on an ensemble of real hydrographs, the three Series Distance metrics lead to different rankings, but in combination came close to the author's subjective ranking, at least closer than single or combined rankings based on the Root Mean Square Error and the Mean Absolute Peak Time Error. Although this reasoning is partly based on strongly subjective components, namely the ranking by the authors and the way of combining the three metrics, the results seem to suggest that the Series Distance jointly evaluates several hydrograph characteristics in a way similar to visual inspection.

The Series Distance currently requires the selection of three parameters: a discharge threshold separating events from low flow conditions, a minimum time overlap to consider two events as matching, and the way of hydrograph smoothing to remove minor peaks and troughs. In order to facilitate and standardize selection of these parameters and also the weighting of the three components, it could be helpful to relate them to general hydrograph properties such as the mean event duration and distance or the distribution of discharge values. This could also facilitate the intercomparison of metrics based on hydrographs from different sites with different characteristics. Also, the Series Distance as presented makes two events comparable by equalizing their number of segments, but does not consider the degree of attunement necessary to achieve this. We propose to count the number and magnitude of peak/trough removals necessary to achieve attunement and to include this information of disagreement in the overall Series Distance metric. This remains to be done in the future.

Recalling the fictitious post-flood conversation of hydrologists and flood managers from the introduction, we hope to contribute with the Series Distance to a better (i.e. non-redundant and traceable) evaluation of hydrological models adaptable to a range of user-specific needs.

The Series Distance is available as Matlab code from the corresponding author.

Appendix A

Pseudocode

A1 Match events

Algorithm to find matching events in an observed and simulated hydrograph. A match occurs when two events are closer to each other than the distance defined by "limitformatch". Matches are unique, i.e. only one event can match another. If several simulated events match one observed (or vice versa), the pair with the largest overlap is used.

```
function MatchEvents

input
$O = (o_1, ..., o_n)$: all observed events, each represented by an object with properties start time ($o.t_s$), end time ($o.t_e$), and number of the matching simulated event ($o$.match)

$S = (s_1, ..., s_n)$: all simulated events, each represented by an object with properties start time ($s.t_s$), end time ($s.t_e$), and number of the matching observed event ($s$.match)

limitformatch: maximum time gap between end of the first event (obs or sim) and start of the second (sim or obs) to be still considered matching

returns: for each $o$ in $O$ and $s$ in $S$ the number of the matching event or $-999$ (if no match was found)

begin
dim overlap $(1, ..., n, 1, ..., m)$
for $i = 1$ to $n$
    for $j = 1$ to $m$
            overlap $(o_i, s_j) = \min(o_i.t_e, s_j.t_e) - \max(o_i.t_s, s_j.t_s) + 1$
            if overlap $(n, m) <$ limitformatch then overlap $(n, m) = -999$
    next $j$
next $i$

while max (overlap) $> -999$
    $i, j$ = index (max(overlap))
    $o_i$.match $= j$; $s_j$.match $= i$
    overlap $(i,*) = -999$; overlap $(*, j) = -999$;
end

end
```

A2 Hydrological cases

Algorithm to find the hydrological case for each time step of a hydrograph.

Possible cases are: -2 = valley, -1 = drop, 0 = not within an event, 1 = rise, 2 = peak.

```
function HydCase

input
$Q = (q_1, ..., q_n)$: hydrograph with n observations or simulations, each represented by an object with property value ($q.v$) and hydrological case ($q$.hydcase)

lolim = discharge threshold; any values below are considered as not within an event

returns: for each $q$ $Q$ the hydrological case in property $q$.hydcase

begin
$q_1$.hydcase = 0; $q_n$.hydcase = 0;
for $i = 2$ to $n-1$
    if $q_i >$ lolim then
    if $(q_i.v - q_{i-1}.v) < 0$ and $(q_{i+1}.v - q_i.v) > 0$ then $q_i$.hydcase $= -2$; end
    if $(q_i.v - q_{i-1}.v) < 0$ and $(q_{i+1}.v - q_i.v) < 0$ then $q_i$.hydcase $= -1$; end
    if $(q_i.v - q_{i-1}.v) > 0$ and $(q_{i+1}.v - q_i.v) > 0$ then $q_i$.hydcase $= 1$; end
    if $(q_i.v - q_{i-1}.v) > 0$ and $(q_{i+1}.v - q_i.v) < 0$ then $q_i$.hydcase $= 2$; end
    else
            $q_i$.hydcase = 0
    end
next $i$
end
```

A3 Equalize events

Algorithm to equalize the number of peaks and valleys in matching observed and simulated events by erasing the least pronounced pairs of peaks and troughs.

```
function Equalize Events

input
$O$ = observed event, represented by an object with property $.P = (p_1, ..., p_n)$ containing all peak values and property $.T = (t_1, ..., t_{n-1})$ containing all trough values, both ordered by time.

$S$ = simulated event matching the observed, represented by an object with property $.P = (p_1, ..., p_\mathrm{m})$ containing all peak values and property $.T = (t_1, ..., t_{\mathrm{m}-1})$ containing all trough values, both ordered by time.

returns: $O$ and $S$, with the number of peaks and troughs equalized

begin

for $i = 1$ to abs $(n - m)$
        if $(n - m) > 0$ then
        dim diff $(1, ..., n-1)$
        for $j = 1$ to $n-1$
                diff $(j) = (O.P(j) - O.T(j)) + (O.P(j+1) - O.T(j))$
next $j$
```

```
        $j$ = index (min (diff))
        $O.T(j)$ = nothing; min $(O.T(j), O.T(j+1))$ = nothing
        elseif $(n-m)<0$ then
        dim diff $(1, ..., m-1)$
        for $j = 1$ to $m-1$
            diff $(j) = (S.P(j) - S.T(j)) + (S.P(j+1) - S.T(j))$
        next $j$
        $j$ = index (min (diff))
        $S.T(j)$ = nothing; min $(S.T(j), S.T(j+1))$ = nothing
        end
next $i$
end
```

A4 Segment distance

Algorithm to calculate the time and amplitude distance between matching segments in an observed and simulated event. The simulated segment is approximated by a polygon line. The polygon is sampled with n equally spaced points (n being the number of points in the observed segment). These are used to calculate the distance from the n observed points in the segment.

*function*Segment Distance

input

O = segment (rise or decline) of length n in an observed event, represented by an object with property $.V=(v_1, ..., v_n)$ containing the amplitude of all segment points and property $.T=(t_1, ..., t_n)$ containing the timing of all segment points. The elements are ordered in time.

S = segment (rise or decline) of length m in a simulated event matching the observed segment. It is represented by an object with property $.V=(v_1, ..., v_\mathrm{m})$ containing the amplitude of all segment points and property $.T=(t_1, ..., t_\mathrm{m})$ containing the timing of all segment points. The elements are ordered in time.

returns
Dist_v = (d_v$_1$, ..., d_v$_n$) containing n amplitude differences between O and S
Dist_t = (d_t$_1$, ..., d_t$_n$) containing n timing differences between O and S

```
begin dim poly_t = $(1, ..., n)$; dim poly_v = $(1, ..., n)$
poly_t = linspace $(S.T, n)$      rem: create n equally spaced points within $[S.T(1), S.T(m)]$
poly_v = linintpol ($S.T$, poly_t)      rem: find point values by linear interpolation

for $i = 1$ to $n$
        Dist_t($i$) = poly_t($i$) $- O.T(i)$
        Dist_v($i$) = poly_v($i$) $- O.V(i)$
next $i$
end
```

Acknowledgements. The authors wish to thank the Flood forecasting centre Iller/Lech for supplying both the observations at Kempten gauge and the Cosmo-Leps forecasts. Also, we thank Conrad Jackisch, Pedro Restrepo, Olga Semenova, Massimiliano Zappa, Markus Casper and two anonymous referees for helpful comments.

Edited by: N. Verhoest

References

Bastidas, L. A., Gupta, H. V., Sorooshian, S., Shuttleworth, W. J., and Yang, Z. L.: Sensitivity analysis of a land surface scheme using multicriteria methods, J. Geophys. Res.-Atmos., 104, 19481–19490, 1999.

Beldring, S.: Multi-criteria validation of a precipitation-runoff model, J. Hydrol., 257, 189–211, 2002.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., 36, 3663–3674, 2000.

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z. Y., and Smith, M.: Toward improved streamflow forecasts: Value of semidistributed modeling, Water Resour. Res., 37, 2749–2759, 2001.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S.: Forecast verification: Current status and future directions, Meteorol. Appl., 15, 3–18, doi:10.1002/met.52, 2008.

Chouakria-Douzal, A. and Nagabhushan, P.: Improved fréchet distance for time series, in: Data science and classification, Studies in classification, data analysis, and knowledge organization, Springer, Berlin, Heidelberg, Germany, 13–20, 2006.

Contreras-Cristán, A., Guti'errez-Peña, E., and Walker, S. G.: A Note on Whittle's Likelihood, Commun. Stat.-Simul. C., 35, 857–875, 2006.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts, Part i: Methodology and application to mesoscale rain areas, Mon. Weather Rev., 134, 1772–1784, 2006.

Dawson, C. W., Abrahart, R. J., and See, L. M.: Hydrotest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environ. Modell. Softw., 22, 1034–1052, 2007.

Dawson, C. W., Abrahart, R. J., and See, L. M.: Hydro test: Further development of a web resource for the standardised assessment of hydrological models, Environ. Modell. Softw., 25, 1481–1482, 2010.

Donaldson, R. J., Dyer, R. M., and Kraus, M. J.: An objective evaluator of techniques for predicting severe weather events, in: 9th Conf. on Severe Local Storms, Norman, OK, USA, 321–326, 1975.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework, Meteorol. Appl., 15, 51–64, doi:10.1002/met.25, 2008.

Eiter, T. and Mannila, H.: Computing discrete frechet distance, Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria, 1994.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, 13, 2008.

Frechet, M.: Sur quelques points du calcul fonctionnel, Rendiconto del Circolo Mathematico di Palermo, 22, 1–74, 1906.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, 1998.

Gupta, H. V., Wagener, T., and Liu, Y. Q.: Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, 2008.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, 2009.

Jain, S. K. and Sudheer, K. P.: Fitting of hydrologic models: A close look at the nash-sutcliffe index, J. Hydrol. Eng., 13, 981–986, 2008.

Keil, C. and Craig, G. C.: A displacement and amplitude score employing an optical flow technique, Weather Forecast., 24, 1297–1308, doi:10.1175/2009waf2222247.1, 2009.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, 2005,
http://www.adv-geosci.net/5/89/2005/.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, 1999.

Lerat, J., Anderssen, B., and Gouweleeuw, B.: How to estimate timing errors in flood forecasting systems?, in: Geophys. Res. Abstr., Vol. 12, EGU2010-3832, EGU General Assembly 2010, Vienna, Austria, 2010,

Liu, Y., Brown, J., Demargne, J., and Seo, D.-J.: Using wavelet analysis to assess timing errors in streamflow predictions, in: Geophys. Res. Abstr., Vol. 12, EGU2010-5456, EGU General Assembly 2010, Vienna, Austria, 2010,

Ludwig, K.: The Program System FGMOD for Calculation of Runoff Processes in River Basins, Z. Kulturtech. Flurber., 23, 25–37, 1982.

Ludwig, K. and Bremicker, M.: The water balance model larsim – design, content and applications, Freiburger schriften zur hydrologie, Institut für Hydrologie, Uni Freiburg i. Br., Germany, 2006.

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.

McCuen, R., Knight, Z., and Cutter, G.: Evaluation of the nash-sutcliffe efficiency index, J. Hydrol. Eng., 11, 597–602, 2006.

Moeckel, R. and Murray, B.: Measuring the distance between time series, Physica D, 102, 187–194, 1997.

Montanari, A. and Toth, E.: Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins?, Water Resour. Res., 43, W05434, doi:10.1029/2006wr005184, 2007.

Murphy, A. H.: Skill scores based on the mean-square error and their relationships to the correlation-coefficient, Mon. Weather Rev., 116, 2417–2425, 1988.

Muskulus, M. and Verduyn-Lunel, S.: Wasserstein distances in the analysis of time series and dynamical systems, Physica D, 240, 45–58, 2011.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part i – a discussion of principles, J. Hydrol., 10, 282–290, 1970.

Ouyang, R., Ren, L., Cheng, W., and Zhou, C.: Similarity search and pattern discovery in hydrological time series data mining, Hydrol. Process., 24, 1198–1210, doi:10.1002/hyp.7583, 2010.

Pappenberger, F. and Beven, K.: Functional classification and evaluation of hydrographs based on multicomponent mapping, Int. J. River Basin Manage., 2, 89–100, 2004.

Pebesma, E. J., Switzer, P., and Loague, K.: Error analysis for the evaluation of model performance: Rainfall-runoff event time series data, Hydrol. Process., 19, 1529–1548, 2005.

Perng, C.-S., Wang, H., Zhang, S., and Parker, D. S.: Landmarks: A new model for similarity-based pattern querying in time series databases, Proceedings of the 16th International Conference on Data Engineering, 33, 2000.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth Syst. Sci., 13, 999-1018, doi:10.5194/hess-13-999-2009, 2009.

Sakoe, H. and Chiba, S.: Dynamic-programming algorithm optimization for spoken word recognition, IEEE T. Acoust. Speech, 26, 43–49, 1978.

Schaefli, B. and Gupta, H. V.: Do nash values have value?, Hydrol. Process., 21, 2075–2080, 2007.

Schaefli, B. and Zehe, E.: Hydrological model performance and parameter estimation in the wavelet-domain, Hydrol. Earth Syst. Sci., 13, 1921–1936, doi:10.5194/hess-13-1921-2009, 2009.

Spate, J., Croke, B., and Jakeman, A.: Data mining in hydrology, MODSIM 2003, Proceedings of the 2003 International Congress on Modelling and Simulation, Townsville, Australia, 2003, 422–427, 2003.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, 2001.

van Griensven, A. and Bauwens, W.: Multiobjective autocalibration for semidistributed water quality models, Water Resour. Res., 39, 9, 2003.

Weglarczyk, S.: The interdependence and applicability of some statistical quality measures for hydrological models, J. Hydrol., 206, 98–103, 1998.

Whittle, P.: Estimation and information in stationary time series, Arkiv för Matematik, 2, 423–434, doi:10.1007/bf02590998, 1953.

Winsemius, H. C., Schaefli, B., Montanari, A., and Savenije, H. H. G.: On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, Water Resour. Res., 45, 15, 2009.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, J. Hydrol., 204, 83–97, 1998.

Yilmaz, K. K., Hogue, T. S., Hsu, K. L., Sorooshian, S., Gupta, H. V., and Wagener, T.: Intercomparison of rain gauge, radar, and satellite-based precipitation estimates with emphasis on hydrologic forecasting, J. Hydrometeorol., 6, 497–517, 2005.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi:10.1029/2007WR006716, 2008.