Hydrol. Earth Syst. Sci. Discuss., 7, 8387–8425, 2010
www.hydrol-earth-syst-sci-discuss.net/7/8387/2010/
doi:10.5194/hessd-7-8387-2010

Hydrology and Earth System Sciences Discussions

Series distance – an intuitive metric for hydrograph comparison

U. Ehret[1] **and E. Zehe**[1]

[1]Insitute of Water and Environment, Department of Hydrology and River Basin Management, Technische Universität München, Munich, Germany

Received: 18 October 2010 – Accepted: 18 October 2010 – Published: 25 October 2010

Correspondence to: U. Ehret (u.ehret@bv.tum.de)

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

Abstract

Applying metrics for hydrograph comparison 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 or the Mean Peak Time Error 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 is an event-based method and 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, namely by the same relative position in matching segments (rise or recession) of matching events. 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 (which is for example the case with metrics related to Peak Time Errors).

After presenting the Series Distance theory, we discuss its properties and compare it to those of standard metrics and visual inspection, both at the example of simple, artificial hydrographs and an ensemble of realistic forecasts. The results suggest that the Series Distance compares and evaluates hydrographs in a way comparable to visual inspection, but in an objective, reproducible way.

1 Introduction

Which is the best among a set of hydrological forecasts or simulations? All modellers in Hydrology are sooner or later confronted with this question, having to rank or choose among a set of forecasts or simulations, using some sort of metric. Applying metrics,

HESSD

7, 8387–8425, 2010

measures or objective functions (including subjective visual inspection) is therefore at the heart of hydrological modelling in its widest sense. They are used to analyze and classify hydrological systems, calibrate and validate hydrological models through comparison of observations and model output, identify scales at which to separate explicit and implicit representations of structures and processes, and also to quantify information about hydrological processes or models.

In its origins, hydrological modelling was mainly focused on analysis and reproduction of observed discharge timeseries at the catchment scale. Hence the repertoire of metrics in Hydrology was, and to a declining degree still is, mainly related to hydrographs. As hydrographs constitute a very particular subset in the large family of possible datasets, it is worth revisiting some of their properties before discussing the metrics most commonly used in hydrological modelling.

1.1 Hydrograph characteristics

A hydrograph is basically a two-dimensional dataset representative for one point in space (the river cross-section). The units of the two dimensions, namely discharge and time, are fundamentally different. This impedes any straightforward 2-dimensional distance calculations as it is for instance possible with spatial rainfall observations. This means that when evaluating a simulation or forecast, at some point a relation between errors in timing and amplitude has to be established. Often this is done only implicitly by choosing a certain metric (and ignoring others).

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?

HESSD
7, 8387–8425, 2010

Looking at the shape of a hydrograph reveals some obvious 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. Secondly, a hydrograph is not time-symmetrical: the rising and falling limbs of an event, with their shaping dominated by different parts of the hydrometeorological causal chain look different, the first usually being shorter and steeper than the latter. 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. 1.2).

Finally, any comparison of observed and simulated hydrographs requires a decision or rule on what to compare. For most metrics evaluating amplitude errors, this rule is simply equality in time, which allows one-to-one mapping of observations and simulations. The rules for metrics related to timing errors such as Peak Time Errors are less straightforward. They usually require the identification of characteristic points of a hydrograph, such as the peak of an event and subsequent matching of those points on the observed and simulated hydrograph.

Probably because they were simple, intuitive and straightforward to compute, the first metrics for 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. Peak Time Error. As both are still widely used in hydrological modelling, their characteristics will be briefly discussed in the following section.

1.2 Standard metrics for hydrographs

1.2.1 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. The RMSE is calculated as the mean of the squared distances between observations and simulations at the

HESSD
7, 8387–8425, 2010

same point in time, which is then backtransformed to discharge units by taking the root. Its range of values is [0, ∞], with zero being the optimum. The NSE is the RMSE normalized to [−∞, 1] by division with the deviation of the observations from their mean. 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 RMSE can be decomposed into three parts, evaluating the relative variability, the bias and the correlation coefficient. This means that the RMSE is essentially a weighted three-criteria objective function. However, using only the 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 (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

HESSD
7, 8387–8425, 2010

in shape, as shown in Fig. 1. The "observed event" (bold line) is of arbitrary length 17 hours and has a peak of 100 m^3/s. 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.
- 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.

HESSD
7, 8387–8425, 2010

– 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. This is comparable to an RMSE of 12 m^3/s 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.

1.2.2 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 Peak Time Error (MPTE) then is the average of all peak time errors in a hydrograph. 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

HESSD
7, 8387–8425, 2010

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 MPTE to the synthetic triangular hydrographs and all possible pairs of time and multiplicative value offsets as described in Sect. 1.2.1. 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 MPTE 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.

When comparing the error surfaces for RMSE and MPTE, 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 MPTE 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.

1.2.3 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

HESSD
7, 8387–8425, 2010

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 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 timeseries of amplitude errors. This concept was further developed by Reusser et al. (2008), who analyzed the temporal dynamics of many metrics applied on hydrographs, clustering them into typical error classes and from this, drawing specific conclusions on structural deficits of the underlying models. This approach not only represents the trend of looking at hydrographs in a different 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 Bauwens (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

HESSD
7, 8387–8425, 2010

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).

In the light of these developments and the drawbacks of merely amplitude-based metrics as illustrated above, it is the aim of this study to propose a new objective metric for hydrographs termed "Series Distance", which closely follows subjective reasoning in visual inspection. The method, underlying assumptions and the output are presented in Sect. 2. This is followed by an application to both simple synthetic as well as real-world hydrographs in Sect. 3, along with a discussion of results. Finally, conclusions are drawn and ways forward are discussed in Sect. 4.

2 The metric "Series Distance"

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 (not just individual characteristic points such as a peak), simultaneously but separately for errors in amplitude and timing. A typical linguistic

HESSD
7, 8387–8425, 2010

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.

– 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 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.

2.1 Procedure

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

HESSD
7, 8387–8425, 2010

– Identify events: From the hydrograph, individual events are identified by applying a user-defined parameter termed "no-event threshold" [m^3/s]. 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 m^3/s.

– 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. 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 timeseries:

HESSD
7, 8387–8425, 2010

– 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.

In Fig. 5, each point of the observed and simulated hydrograph is marked with its hydrological case. An event invariably consists of the following sequence of components:

Start, *a**rise, *b**(peak, *c**recession, trough, *d**rise), peak, *e**recession, end with *a*, *b*, *c*, *d*, *e* [0, ∞].

This means that in the simplest case, an event consists of a start, a peak and an end (*a*, *b*, *c*, *d*, *e* = 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

HESSD
7, 8387–8425, 2010

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.

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

HESSD
7, 8387–8425, 2010

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. 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. 2.2). 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.

HESSD
7, 8387–8425, 2010

– 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], the match limit [h] and the manner of the smoothing.

2.2 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. 2.1, a number of metrics can be calculated:

– Contingency table: the frequency of matching, missing and false events can be listed in a contingency table. 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), which is the number of matching events divided by the sum of matching, missing and false events. Ranging from zero to one, a Threat Score of one indicates optimal reproduction of events.

– Overall amplitude and timing error: from the set of amplitude and timing errors, 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 and amplitude for the following reasons: firstly, taking the absolute value

HESSD
7, 8387–8425, 2010

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. In the following, we will use the abbreviation SDv and SDt (*S*eries *D*istance with respect to *v*alue and *t*iming) for the Mean Absolute Error of amplitude and timing, respectively.

– Many other metrics can be derived from the Series Distance procedure, e.g. scatterplots 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 SDv and the SDt. 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.

2.3 Alternatives

Development of the SD procedure as described in Sects. 2.1 and 2.2 was a matter of trial and error and frequently ended in dead ends. As we think that much can be learned

HESSD
7, 8387–8425, 2010

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.

HESSD
7, 8387–8425, 2010

hydrographs from alpine catchments with short, intensive events or hydrographs from large lowland catchments with drawn-out, smooth events).

3 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 Peak Time Error) and visual inspection.

3.1 Application on a synthetic hydrograph

Similar to the discussion of the RMSE and MPTE characteristics in Sects. 1.2.1 and 1.2.2, 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 MPTE, we calculated the SDv and SDt 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, "no-event" threshold = 1.9 m^3/s, 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 SDv and SDt 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 MPTE are:

- Both surfaces resemble a turned ridge roof, but in contrast to RMSE and MPTE, the (turned) ridges point in different directions: SDv is sensitive to amplitude offsets only, while SDt is sensitive to time offsets only. Both error surfaces are

HESSD
7, 8387–8425, 2010

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 SDv and SDt drop to zero, as for non-matching events, no distances are calculated (see Sect. 2.1). The disagreement of the observed and simulated hydrograph is in this case captured in the contingency table.

3.2 Application on realistic hydrographs

Finally, we applied the SD procedure to eight realistic pairs of observed and simulated 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 km^2. The discharge was observed during a small 5-day flood event from 21–27 April 2008. The related simulations 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, 2005). We chose an ensemble forecast as with this, a number of different simulations is available which are all related to the same observed hydrograph. This facilitates performance comparisons among the simulations and allows ranking. 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 simulations 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.

HESSD
7, 8387–8425, 2010

In order to apply the Series Distance, its free parameters were set to the following values: match limit = 0 h, "no-event" threshold = 88 m^3/s (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 MPTE 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 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 Fig. 6 and Fig. 9. In addition, connection lines between related points (i.e. the point pairs used for distance calculations) on the two timeseries 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 MPTE is calculated only on a single pair of points.

HESSD
7, 8387–8425, 2010

All metrics (RMSE, MPTE, Threat Score, SDv and SDt) for each of the eight simulations are shown in Table 1. 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 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 2.

In addition to ranking the individual metrics (columns I, II, IV, V, and VII), we also calculated the ranks of combined metrics. First, we combined RMSE and MPTE, giving equal weights to each of them. To this end, the ranks of RMSE and MPTE 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 MPTE 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 SDv and SDt were combined in the same manner as RMSE and MPTE. In contrast to RMSE and MPTE, however, the rankings of the two SD distances are dissimilar. For example, hydrograph eight was ranked best by the SDv and worst by the SDt. 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

HESSD
7, 8387–8425, 2010

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 SDv 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, SDv and SDt. 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%. SDv and SDt ranks were equally weighted with 25%, respectively.

Finally, the author's subjective ranking of the eight test hydrographs is also shown in Table 2, column IX. During the underlying visual hydrograph inspection, we followed the general guidelines discussed in Sect. 2. 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 2 ("Rank Diff"). Comparing the Rank Errors for the different metrics reveals several interesting points:

- Combining RMSE and MPTE 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-

HESSD
7, 8387–8425, 2010

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. 2.1).

– In contrast to RMSE and MPTE, combination of the SD metrics continually improves the agreement with subjective classification: while SDv and SDt taken separately still show relatively weak agreement (although better than for RMSE or MPTE), a combination of the two leads to a Rank Error of only 10 (column VII).

– Finally, combining the Threat Score, SDv and SDt (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 SDv and SDt are essentially non-redundant information, as their combination decreased the Rank Error substantially.

4 Summary and conclusions

In this paper, we proposed a new metric to compare simulated and observed 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 evaluates three hydrograph characteristics: agreement of event occurrence, expressed by a contingency table, and the distance of matching events with respect to amplitude and timing. The latter two are based on distance calculations between matching points on the observed and simulated hydrographs: matching points are located at the same relative position of comparable segments of matching events (e.g. in the middle of the first rise). This procedure is closer to the subjective way of relating points than e.g. the way it is done for most amplitude-related metrics such as the Root Mean Square Error, where matching points are simply the ones at the same point in time. Based on the point pairs (which cover the complete event), amplitude and timing errors are calculated simultaneously but separately. For the example of simple, triangular hydrographs we demonstrated that

HESSD
7, 8387–8425, 2010

the resulting Mean Absolute Error in Timing and Amplitude is less redundant than the Root Mean Square Error and the Mean 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 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 an observed and a simulated event 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. This remains to be done in the future.

Which is the best among a set of hydrological forecasts or simulations? This question we asked at the beginning of this article has no unique answer as it will always be asked with a case-specific background and intention. The challenge is therefore not to find a unique all-purpose metric but rather to find a standardized and traceable way to apply many non-redundant objective metrics, but with enough degrees of freedom to subjectively attune them (or rather attune the way they are combined) to the user's needs. Although the Series Distance only evaluates one aspect of hydrological modelling, namely the hydrograph, it combines several non-redundant aspects of it. Thus, we hope to contribute to this task.

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

HESSD
7, 8387–8425, 2010

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 and Markus Casper for helpful comments.

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, Journal of Hydrology, 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.

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. Model. 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. Model. 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, 1975, 321–326, 1975.

Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44(13), W01402, doi:10.1029/2006wr005563, 2008.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic mod-

HESSD

7, 8387–8425, 2010

els: 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.
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: Geophysical Research Abstracts, Vol. 12, EGU2010-3832, EGU General Assembly 2010, Vienna, 2010,
Liu, Y., Brown, J., Demargne, J., and Seo, D.-J.: Using wavelet analysis to assess timing errors in streamflow predictions, in: Geophysical Research Abstracts, Vol. 12, EGU2010-5456, EGU General Assembly 2010, Vienna, 2010,
Ludwig, K.: The Program System FGMOD for Calculation of Runoff Processes in River Basins, Zeitschrift für Kulturtechnik und Flurbereinigung, 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., 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.
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.
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.

HESSD
7, 8387–8425, 2010

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.
Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth Syst. Sci. Discuss., 5, 3169–3211, doi:10.5194/hessd-5-3169-2008, 2008.
Schaefli, B., and Gupta, H. V.: Do nash values have value?, Hydrol. Process., 21, 2075–2080, 2007.
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, 1348, doi:10.1029/2003wr002284, 2003.
Weglarczyk, S.: The interdependence and applicability of some statistical quality measures for hydrological models, J. Hydrol., 206, 98–103, 1998.
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), W12422, doi:10.1029/2009wr007706, 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.

HESSD
7, 8387–8425, 2010

Table 1. Metrics for 8 pairs of simulated and observed hydrographs as shown in Fig. 4. RMSE = Root Mean Square Error, MPTE = Mean Peak Time Error, SDv = Amplitude Error of Series Distance, SDt = Timing Error of Series Distance.

Sim #	RMSE [m^3/s]	MPTE [h]	Threat Score [–]	SDv [m^3/s]	SDt [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

HESSD
7, 8387–8425, 2010

Table 2. Ranked metrics from Table 1 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, MPTE = Mean Peak Time Error, I&II = ranks of columns I and II added and ranked, SDv = Amplitude Error of Series Distance, SDt = Timing Error of Series Distance, V&VI = ranks of columns V and VI added and ranked, IV&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 I	MPTE II	I&II III	Threat Score IV	SDv V	SDt VI	V&VI VII	IV&VII VIII	Subjective 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

HESSD
7, 8387–8425, 2010

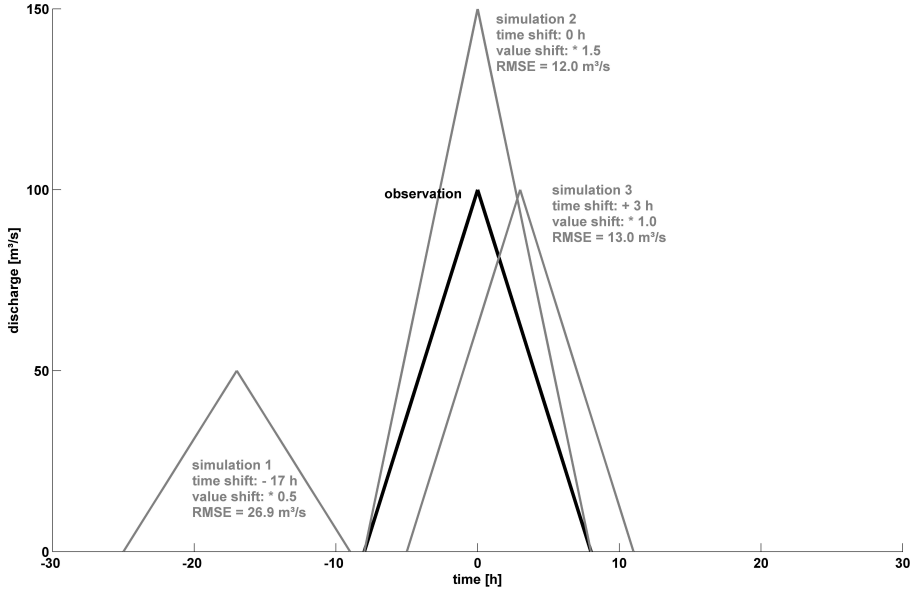

Fig. 1. Synthetic, triangular events. "Observation" (bold line) and three example "simulations" (normal lines) derived from the "observation" by time offsets and multiplicative value offsets.

HESSD
7, 8387–8425, 2010

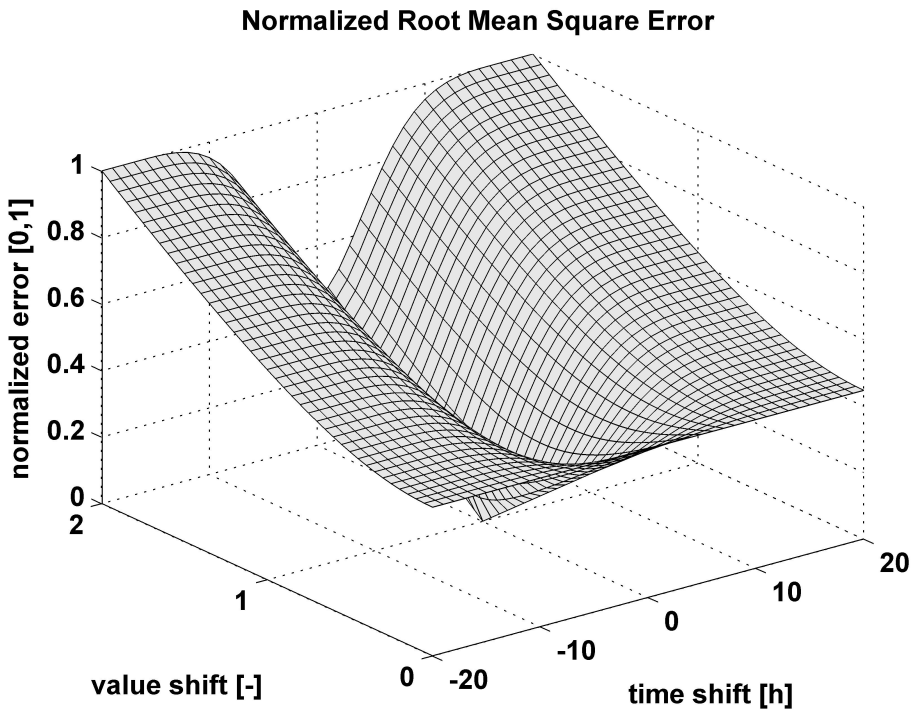

Fig. 2. Error surface of the Root Mean Square Error (RMSE) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [−20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum error.

HESSD
7, 8387–8425, 2010

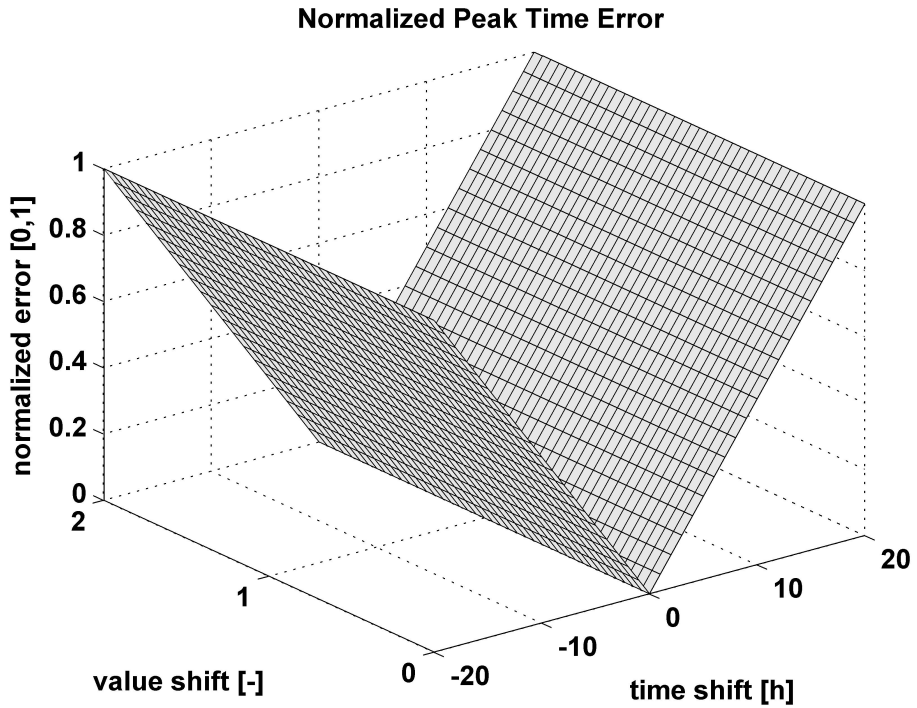

Fig. 3. Error surface of the Mean Peak Time Error (MPTE) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [−20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum error.

HESSD
7, 8387–8425, 2010

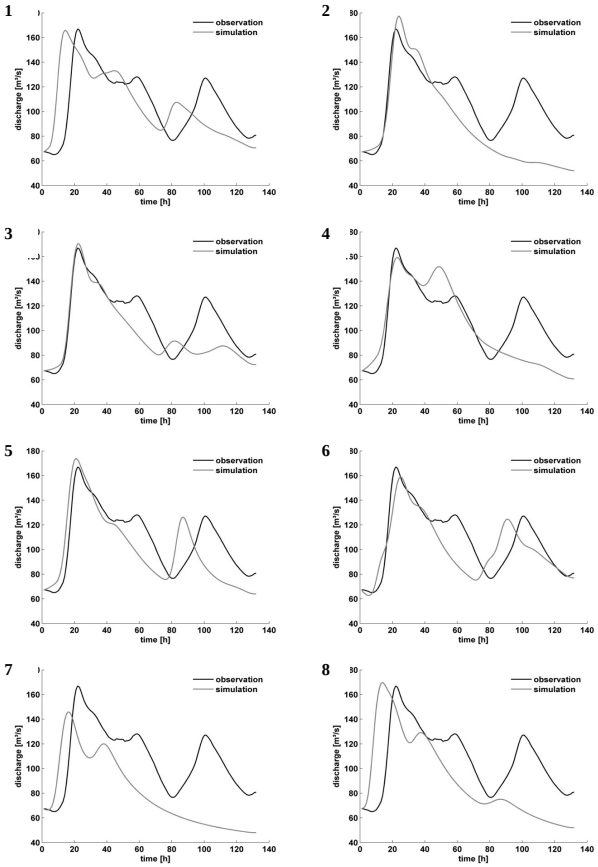

Fig. 4. Observed discharge at gauge Kempten/Iller (954 km^2) for period 21 April 2008 14:00–27 April 2008 00:00 (132 h) and 8 simulations with hydrological model "Fgmod" (Ludwig, 1982) based on Cosmo-Leps ensemble weather forecasts (Marsigli et al., 2005).

HESSD
7, 8387–8425, 2010

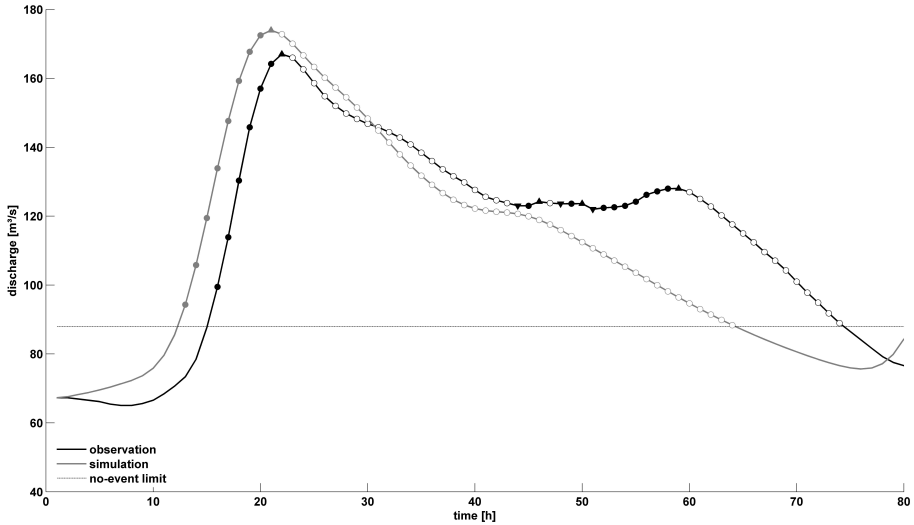

Fig. 5. Example of a matching observed (black) and simulated (grey) event (detail of event 5 in Fig. 4). The hydrological case is shown for each point: "rise" (filled circle), "peak" (upward triangle), "recession" (empty circle), "trough" (downward triangle), "no event" (no marker). The "no-event" threshold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

HESSD
7, 8387–8425, 2010

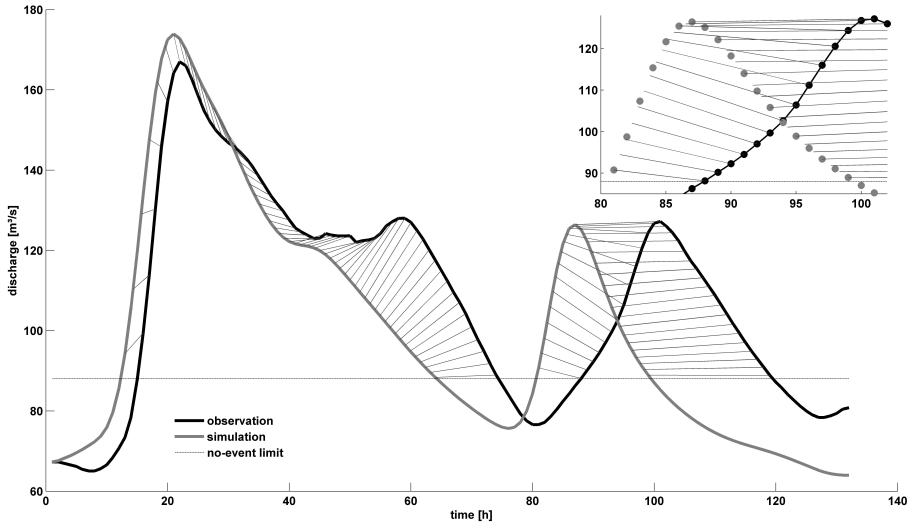

Fig. 6. Example of a matching observed (black) and simulated (grey) event (event 5 in Fig. 4). Connections (thin grey lines) between matching points of observation and simulation according to the Series Distance procedure are shown. The small inserted figure reveals that the observed points in a segment (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.

HESSD
7, 8387–8425, 2010

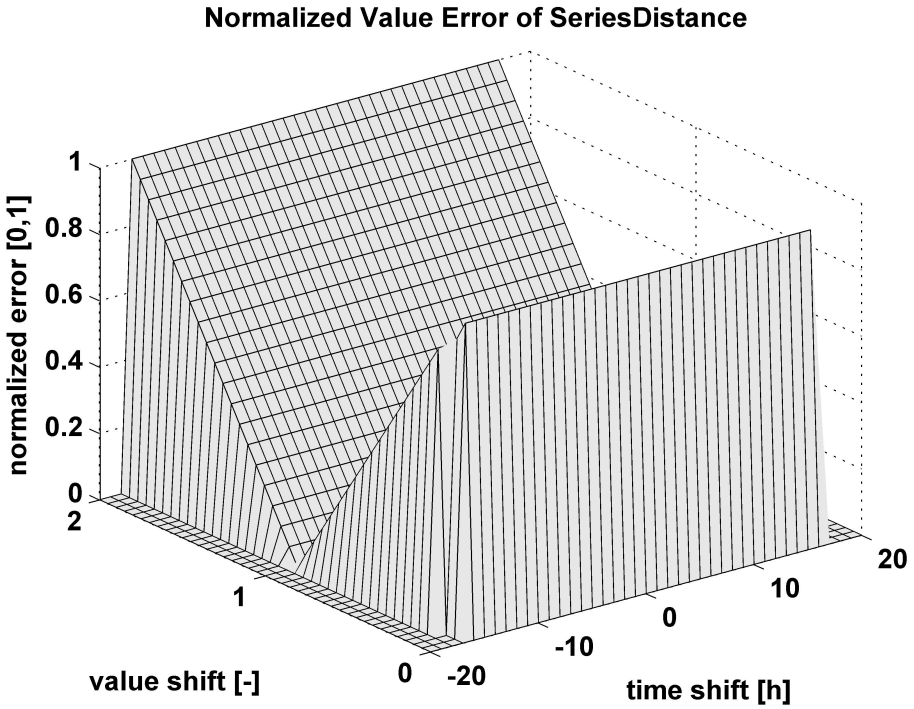

Fig. 7. Error surface of the value/amplitude error of the Series Distance (SD) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [−20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with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

HESSD
7, 8387–8425, 2010

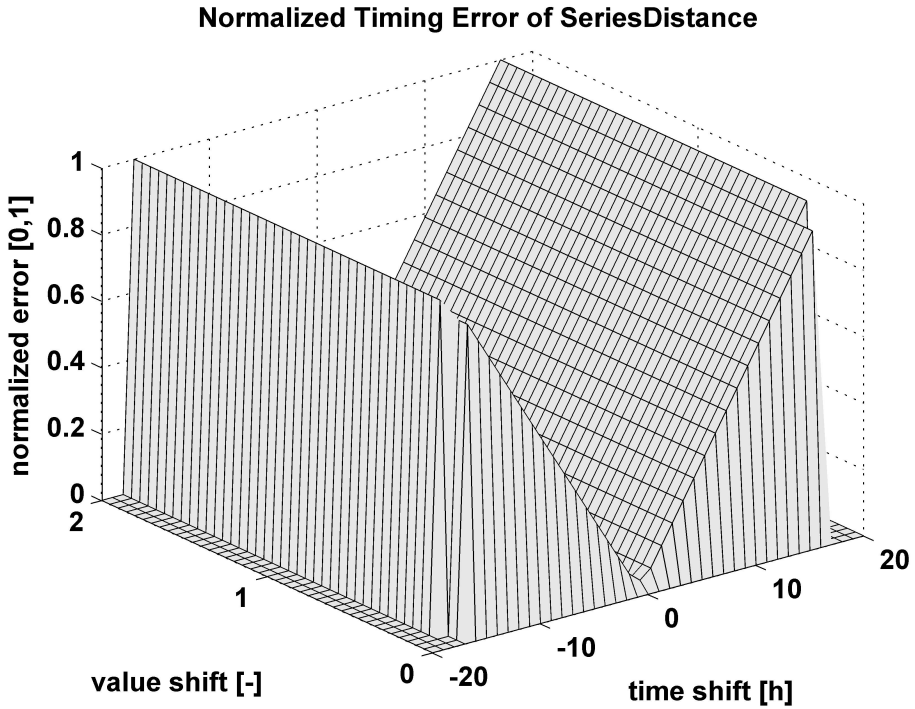

Fig. 8. Error surface of the timing error of the Series Distance (SD) for synthetic, triangular events as shown in Fig. 1. Simulations are shifted in time (offset range [−20 h, 20 h]) and amplitude (multiplier range [0, 2]). The error surface is normalized to [0, 1] by means of division with the maximum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

HESSD
7, 8387–8425, 2010

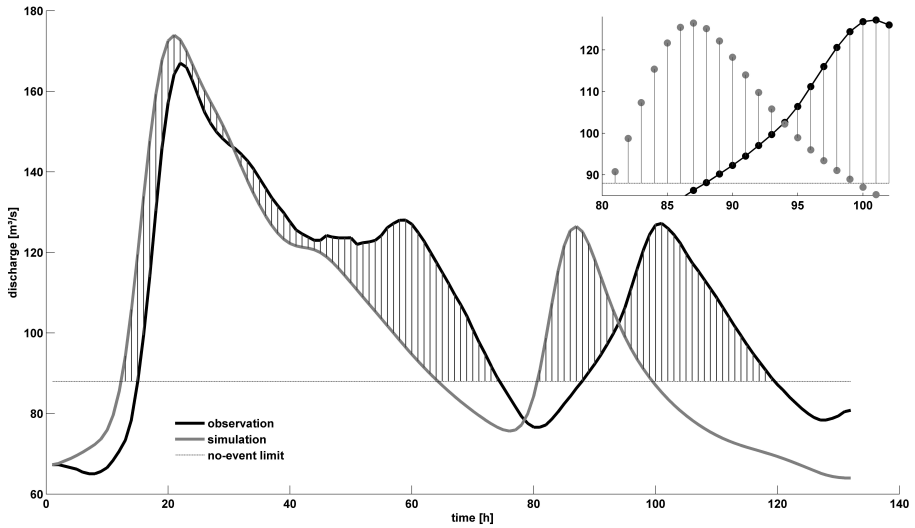

Fig. 9. Example of a matching observed (black) and simulated (grey) event (event 5 in Fig. 4). Connections (thin grey lines) between matching points of observation and simulation according to the RMSE are shown. Note that connections may exist between non-matching segments of the hydrographs (rise with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

HESSD
7, 8387–8425, 2010