Hydrol. Earth Syst. Sci. Discuss., 5, 3169–3211, 2008
www.hydrol-earth-syst-sci-discuss.net/5/3169/2008/

Hydrology and Earth System Sciences Discussions

Papers published in *Hydrology and Earth System Sciences Discussions* are under open-access review for the journal *Hydrology and Earth System Sciences*

Analysing the temporal dynamics of model performance for hydrological models

D. E. Reusser[1], T. Blume[1], B. Schaefli[2], and E. Zehe[3]

[1]University of Potsdam, Institute for Geoecology, Potsdam, Germany
[2]Delft University of Technology, Faculty of Civil Engineering and Geosciences, Water Resources Section, Delft, The Netherlands
[3]TU München, Institute of Water and Environment, München, Germany

Received: 16 September 2008 – Accepted: 18 September 2008 – Published: 19 November 2008

Correspondence to: D. E. Reusser (dreusser@uni-potsdam.de)

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

Abstract

The temporal dynamics of hydrological model performance gives insights into errors that cannot be obtained from global performance measures assigning a single number to the fit of a simulated time series to an observed reference series. These errors can include errors in data, model parameters, or model structure. Dealing with a set of performance measures evaluated at a high temporal resolution implies analyzing and interpreting a high dimensional data set. This paper presents a method for such a hydrological model performance assessment with a high temporal resolution and illustrates its application for two very different rainfall-runoff modeling case studies. The first is the Wilde Weisseritz case study, a headwater catchment in the eastern Ore Mountains, simulated with the conceptual model WaSiM-ETH. The second is the Malalcahuello case study, a headwater catchment in the Chilean Andes, simulated with the physics-based model Catflow. The proposed time-resolved performance assessment starts with the computation of a large set of classically used performance measures for a moving window. The key of the developed approach is a data-reduction method based on self-organizing maps (SOMs) and cluster analysis to classify the high-dimensional performance matrix. Synthetic peak errors are used to interpret the resulting error classes. The final outcome of the proposed method is a time series of the occurrence of dominant error types. For the two case studies analyzed here, 6 such error types have been identified. They show clear temporal patterns which can lead to the identification of model structural errors.

1 Introduction

Hydrological modelling essentially includes – implicitly or explicitly – five steps: 1) Deciding on the dominating processes and on appropriate concepts for their description. This first step is ideally based on data and process observations as it requires a thorough understanding of how the catchment functions. 2) Turning these concept into

HESSD

5, 3169–3211, 2008

equations. For the more common concepts in hydrology, equations are readily available. 3) Coding and numerically solving these equations. Again, we think that it is of great advantage to use existing work if it is available (Buytaert et al., 2008). 4) Once the model structure is defined, usually a number of model parameters have to be estimated (Gupta et al., 2005). 5) Finally the model has to be tested usually based on an independent data set and we have to decide whether the model is acceptable or not. In the latter case we have to revise the initially chosen concepts and repeat steps 2–5 (see Fenicia et al., 2008, for an example of how to step wise improve a model). However, a revision of our model concept requires a clear understanding of the model's structural deficits: what is going wrong, which part of the model is the origin and when does it go wrong?

Model evaluation is usually carried out by determining certain performance measures, thus quantitatively comparing simulation output and measured data. Various methods of model evaluation have been developed over time: Starting with visual inspection (usually used implicitly or explicitly during manual calibration) more objectivity was achieved with the calculation of performance measures, of which the most widely used in hydrology is certainly the Nash-Sutcliffe-Efficiency (Nash and Sutcliffe, 1970). Automatic calibration methods were developed based on these performance measures and lead to the realisation, that a single measure is not able to catch all the features that should be reproduced by the hydrological model (Gupta et al., 1998). As a result, multi objective calibration methods based on a range of performance measures have been and are still being developed (Gupta et al., 1998; Yapo et al., 1998; Vrugt et al., 2003).

Probably because of the development of automatic calibration procedures and their focus on the entire calibration period, the study of the *temporal dynamics* of model performance – which is implicitly used during visual inspection – did not undergo the same process of formalization.

However, we suggest that identification of temporal dynamics of performance measures can be very useful for detecting model structural errors as a first step of model im-

provement. This is of particular importance for operational flood forecasting, because detailed knowledge about the dominant processes is necessary for credible predictions. Global performance measures are only of little use in this context, because lead times for operational forecasts are typically very short 2 to 36 h. While to our knowledge there are no studies on high resolution temporal dynamics of model performance, it has been shown before that it might be useful to split time series (for example in seasons) to obtain some minimum temporal resolution of performance measures. Choi and Beven (2007) showed with their model conditioning procedure, that performance measures calculated on a seasonal scale give some additional indication of model structure deficiencies when compared to global performance measures. Similarly, Shamir et al. (2005) were able to improve identifiability of model parameters when looking at model performance on different time scales.

The rational behind this study is that we get a much clearer picture of structural model deficiencies if we know

- during which periods the model is or is not reproducing observed quantities and dynamics;
- what the nature of the error in times of bad model performance is;
- which parts/components of the model are causing this error.

A methodology to answer the first two questions is suggested here, while the third topic will be the subject of a subsequent publication (The idea is to combine this method with an approach to identify the model components that are active during times of bad performance by analysing the temporal dynamics of the sensitivity of model parameters). The main objective of this paper is thus to present a new method to analyse the temporal dynamics of the performance of hydrological models and to be more specific about the type of error. We propose to use a combination of a) vectors of performance measures to characterize different error types, b) synthetic peak errors to support error type characterization and c) the time series of the obtained error types to analyse their occurrence with respect to observed and modelled flow dynamics.

We use multiple performance measures to capture different types of model structural deficiencies, similar to multi objective calibration (e.g. Gupta et al., 1998; Yapo et al., 1998; Boyle et al., 2000; Vrugt et al., 2003). Dawson et al. (2007) assembled a list of about 20 performance measures commonly used in hydrology. In addition, we use several performance measures introduced by Jachner et al. (2007) to test the agreement between time series in the field of ecology and which, as we will discuss, are promising for the use in the field of hydrological model calibration

Synthetic peak errors with known characteristics will be used to better understand the model performance measures. Interpreting the values of performance measures based on reference time series has for example been proposed by Krause et al. (2005); Dawson et al. (2007) who used modified natural time series. We use an artificially generated peak as it is easier to control its properties.

As mentioned before hydrological modelling studies do generally not analyse the temporal dynamics of model performance. However, a similar approach to the one suggested here, but referring to parameter uncertainties has been used for the dynamic identifiability analysis (Wagener et al., 2003) and the multi-period model conditioning approach (Choi and Beven, 2007) where the temporal dynamics of parameter uncertainty is analysed. The temporal dynamics of model structure uncertainties have been analysed by Clark et al. (2008), who used more than 100 models from a model family for this study (model structures need to be fairly simple in this case).

While different aspects have been used before, their combination as well as the use of high resolution performance measure time series is a new and promising approach for model evaluation.

The large amount of data produced in such an analysis quickly becomes overwhelming and even confusing. Therefore an appropriate data reduction technique is essential to reduce the dimension of the data, while at the same time loosing as little information as possible. The number of simulated time steps (N) is usually large and multiple performance measures (M) are used at each time step, therefore a set of N^*M values has to be interpreted. Classical methods exist to reduce M, e.g. principle component

analysis, use of scatter plots (Cloke and Pappenberger, 2008), or removal of highly correlated measures (e.g. Gupta et al., 1998). In this study the third method was chosen as it is easy to apply and (contrary to principal component analysis) the variables are interpretable.

In a second step of data reduction, we propose self-organizing maps (SOM) (e.g. Kohonen, 1995; Haykin, 1999), which have already been used in several hydrological studies (see Herbst and Casper, 2008, for a short overview). The use of SOMs allows you to reduce the dimension of a data set while preserving the topology of the data in a two dimensional space (i.e. similar data sets are close to each other). During this step some of the variability is lost as the number of sets is drastically reduced (to be further explained in Sect. 2.3). From the SOM we will identify typical combinations of model performance measures, i.e. error types/error classes. This then leads to the assessment of the temporal dynamics of these typical combinations.

In this manuscript, we first present a detailed description of the methodology (Sect. 2) and then show its application for two case studies. These two case studies differ a) in catchment characteristics (size, topography, land use, soils etc.; Sect. 3) and b) in the hydrological model selected for simulation (process-oriented vs. physically based; Sect. 4). The results for the case studies are presented in Sects. 5 and 6. Main findings and suggested future tasks are summarized in Sect. 7.

2 Methods

The proposed methodology can be summarized as follows:

1. determination of a large set of different performance measures,
2. evaluation of the set of performance measures for a moving time window; this yields a vector of performance measures for each time step;
3. removal of highly correlated performance measures, i.e. of performance measures that have time series showing a high correlation with other time series;

4. use of synthetic peak errors to interpret the values of the remaining performance measures, i.e. to assess their error response;

5. use of SOMs and cluster analysis for further data reduction and classification of error types;

6. analysis and characterization of error types using box plots and synthetic peak errors;

7. analysis of temporal dynamics of error types with respect to measured and modelled time series.

A detailed description of these steps is given below.

2.1 Performance measures

Dawson et al. (2007) assembled around 20 performance measures used in hydrology into a test suite, including the Nash-Sutcliffe coefficient of efficiency CE, several measures based on the absolute or squared error e.g. the mean absolute error MAE, the root mean squared error RMSE and many more. The measures are listed in Table 1. Detailed descriptions are available from (Dawson et al., 2007) or https://co-public.lboro.ac.uk/cocwd/HydroTest/Details.html. Because CE in the positive range is of more interest, we used the following transformation of the standard Nash-Sutcliffe coefficient of efficiency CE^* in order to avoid large negative values, which can occur during periods with little dynamics in the time series:

$$CE = \begin{cases} CE^* & \text{if} \quad CE^* > -1 \\ -ln(-CE^*) - 1 & \text{if} \quad CE^* \leq -1 \end{cases} \tag{1}$$

Most of these measures are designed to capture the degree of exact agreement between modelled and observed values. However, we are also interested to measure the degree of qualitative agreement. Jachner et al. (2007) proposed a number of performance measures determining such a qualitative agreement (see also

http://cran.stat.ucla.edu/web/packages/qualV/qualV.pdf). Their measures are mainly based on MAE, MSE and RMSE defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum |x_{\mathrm{obs}} - x_{\mathrm{sim}}| \tag{2}$$

$$\mathrm{MSE} = \frac{1}{n}\sum (x_{\mathrm{obs}} - x_{\mathrm{sim}})^2 \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (x_{\mathrm{obs}} - x_{\mathrm{sim}})^2} \tag{4}$$

Where x_{obs} is the observed time series and x_{sim} the corresponding simulated time series. Depending on the desired qualitative comparison, they used data transformation to allow for shifts and/or changes in scaling. To obtain measures which are insensitive to shifts, data are centred (denoted by a "*C*"). In order to ignore scaling, data are standardized with a linear transformation, minimizing the deviance measure ("*S*").

In addition, Jachner et al. (2007) provide performance measures for different scales of interest. The absolute scale is most often used and applies to the measures defined above. If the difference calculated as a ratio is of more interest (e.g. simulating twice the observed discharge, regardless of the absolute value), a relative scale ("*P*" from percentage), log transformed data ("*L*") or geometric transformed data ("*G*") are more appropriate (see Jachner et al., 2007, for more details). Finally they define performance measures using an ordinal scale ("*O*" – after transformation of the data to ranks). They also define the longest common sequence (LCS) measure: The discharge time series is reduced to a sequence of letters indicating increases ("*I*"), constant values ("*C*"), or decreases ("*D*"). This sequence for the observed discharge (e.g. IIIIIICCDDDDDDCCCIII) is then compared to the sequence of the simulated discharge. LCS then is defined as the longest accumulation of characters with the same order in both sequences. Thereby the method allows deletions in one of the two series, i.e. characters can be ignored or missed (Jachner et al., 2007, for more details).

For this study, we complemented the above performance measures with the following set of five measures to obtain additional information: 1) The lag time t_L defined as the lag of the maximum in cross correlation, 2) the slope error r_d and 3) the recession error r_k defined as:

$$r_d = \frac{\partial x_{\text{obs}}}{\partial x_{\text{sim}}} \tag{5}$$

$$r_k = \frac{k(x_{\text{obs}})}{k(x_{\text{sim}})} \text{ with } k(x) = -\frac{\partial x}{x} \tag{6}$$

The two measures were calculated at the end point of the time window used to calculate the other measures (see below). 4) The direction error DE, which is obtained by counting the number of times the sign of the slope differs for the observed and the modelled time series. Measures 2–4) only work for "smoothed" time series where noise from the measurement on short time scales has been removed. 5) The error quantile was also calculated at the end point of the time window:

$$Q_e = \text{quantile}(x_{\text{sim}} - x_{\text{obs}}) \tag{7}$$

One way to use these measures would be to translate the modelling goal into some criteria (e.g. "reproduce timing and amplitude of extreme events well") and to select the most suitable performance measures to assess them. However, we prefer a different approach. All 47 measures are calculated for a moving time windows of a certain length and the vector of performance measure values for a window at a given time step t is then used as a finger print of the model performance during this time step. Periods with comparable finger prints can then be identified and characterized.

The selection of window size depends on the process of interest and the data quality (Wagener et al., 2003). For example slow recession processes require wider windows. If data quality is suboptimal, large windows will help to reduce the influence of data errors. After some preliminary tests we selected the window size large enough to capture

large events (Fig. 1). A time window of 10 days and 5 days was chosen as a compromise between looking for the local properties in the time series and having enough data to actually compute the values for the first and second case study, respectively. We repeated the first case study also with window sizes of 5 days and 15 days in order to test the sensitivity of the method with respect to the selected window length.

To reduce the number of performance measures M used for the subsequent steps, only one measure was used from each set of highly correlated performance measures (R>0.85). The vector $\boldsymbol{P}^{(t)}$ of the remaining M^* performance measures was then used as the finger print of the model performance for a given time step t. Of course the initial selection of the performance measures is likely to influence the result of the analysis. We regard our set of 47 measures as sufficiently large to cover the important aspects of deviations between two time series. Therefore we do not expect the results to change substantially if additional measures were added.

2.2 Synthetic errors

There is a need to better understand performance measures and their relationship. Two approaches exist in the literature to get familiarized with unknown measures: the first option is to calculate benchmark values for reference simple models (Schaefli and Gupta, 2007). The second option is to create artificial errors (Cloke and Pappenberger, 2008; Krause et al., 2005; Dawson et al., 2007). We used the second approach by generating synthetic errors for a single peak event as test cases (Fig. 2). The peak was modelled as

$$Q(t)=\begin{cases} Q_b & t<t_0 \\ Q_b * e^{(t-t_0)*k_c} & t_0<=t<t_{\max} \\ Q_b+\left(Q_b * e^{t_{\max}*k_c}-Q_b\right) * e^{(t-t_{\max})*k_r} & t_{\max}<=t \end{cases} \quad (8)$$

Where k_r is the recession constant (negative), k_c is the constant for the rise phase and Q_b is the base flow. t, t_0 and $t_{\max}$ are the time step, the time step when the event starts

and the time step of the maximum peak, respectively. We varied the timing, baseflow, the size of the event and the recession constant to obtain the combinations shown in Fig. 2. Each synthetic error was generated in both possible directions of deviation (e.g. under- and overestimation) and with three different levels (small, medium and large deviation).

2.3 Data reduction with SOM

The dimensionality of the simulated time steps N is reduced with self-organizing maps (SOMs). A SOM (for an example see Fig. 5) is a method to produce a (typically) two dimensional, discretized representation of a higher-dimensional input space (Kohonen, 1995). The topological properties of the input space are preserved in the representation of the SOM. Here, the SOM helps to generate and visualize a typology of the model performance finger prints. The matrix $\mathbf{P}=(\boldsymbol{P}^{(t)})_{t=1,\ldots,N}$ of all performance measures is used as an input to the SOM. The SOM is an artificial neural network with a number $x_{\max} * y_{\max}$ of cells (or neurons) corresponding to the dimension of the map $x_{\max}, y_{\max}$. Each cell has a position on the map x, y and a weight vector $\boldsymbol{v}=(v_j)_{j=1,..,M*}$ with the same dimension as the input vector $\boldsymbol{P}^{(t)}$. The weight vectors are initialized with random values. Then the training phase takes place with the following two steps cycling multiple times through all $\boldsymbol{P}^{(t)}$:

1. The cell most similar (best match, short BM) to the input vector $\boldsymbol{P}^{(t)}$ is determined using a Euclidean distance to the weight vector $\boldsymbol{v}$.

2. The weight for BM and its neighbours on the map are updated:

$$\boldsymbol{v}^{(i+1)}=\boldsymbol{v}^{i}+\sigma(x, y, \mathrm{BM}, i) * \alpha(i) * (\boldsymbol{P}^{(t)}-\boldsymbol{v}^{i}) \quad (9)$$

Where x, y are the cell coordinates, $\alpha(i)$ is the learning coefficient which monotonically decreases with iteration i and $\sigma(x, y, \mathrm{BM}, i)$ is the neighbourhood function – often a Gaussian function.

The resulting map arranges similar vectors of performance measures $\boldsymbol{P}^{(t)}$ close together while dissimilar are arranged apart. After the training phase, new input vectors can be placed on the map by finding the corresponding BM. The synthetic peak errors are placed on the map in this way in order to get a better understanding of the map.

We trained a SOM with a hexagonal and Gaussian neighbourhood with the matrix $\mathbf{P}$. Some of the performance measures were log-transformed (MARE, RAE, IRMSE, r_k, RSMSGE) or transformed with the fifth root (where log-transformation is not possible due to negative values: r_d and t_{test}) and all measures where normalized to the range [0, 1] in order to reduce effects from the differing distribution shapes and scales of the performance measures.

The representation of the SOM (for an example see Fig. 5 top left plot) is based on work by Cottrell and de Bodt (1996). Each cell of the neural network is represented as a polygon. The intensity of the colouring represents the number of $\boldsymbol{P}^{(t)}$ associated with the cell (i.e. the cell weight vector $\boldsymbol{v}$ was the best match BM to the input vector $\boldsymbol{P}^{(t)}$). The shape of the polygon represents the distance (Euclidean distance) to the eight neighbouring cells. Large polygons indicate a small distance to the neighbour while if the polygon shrinks in one direction, the distance to the cell in this direction is large. Colouring of the cells can also be used to show the distribution of a specific performance measure on the map.

2.4 Identification of regions of the SOM

To further summarize the results, characteristic regions of the SOM with similar weight vectors $\boldsymbol{v}$ were determined using fuzzy c-means clustering (Bezdek, 1981). As in all clustering algorithms, the $\boldsymbol{v}$ are divided into clusters, such that they are as similar as possible within the same cluster and as different as possible between clusters. In fuzzy clustering, the $\boldsymbol{v}$ can belong to multiple clusters with all the fuzzy membership values μ_i summing up to 1. In c-means clustering the cluster memberships μ_{ki} are found by

minimizing the function

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ki})^m \|\boldsymbol{v}_k - \boldsymbol{w}_i\|^2 \quad (10)$$

where the $\boldsymbol{w}_i$ are the cluster centres, $\boldsymbol{v}_k$ are the weight vectors of the SOM, and m is a parameter modifying the weight of each fuzzy membership, and $\| \|^2$ is the Euclidean distance.

As suggested by Choi and Beven (2007), the validity index V_{XB} from Xie and Beni (1991) can be used to determine the optimal number of clusters:

$$V_{XB} = \frac{\sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ki})^m \|\boldsymbol{v}_k - \boldsymbol{w}_i\|^2}{c \left(\min_{i \neq k} \|\boldsymbol{w}_i - \boldsymbol{w}_k\|^2 \right)} \quad (11)$$

The number of clusters is thereby optimized in correspondence with the goal of the cluster analysis to have the $\boldsymbol{v}$ as similar a possible within a cluster (compactness – numerator in Eq. 11) and as dissimilar as possible between classes (separation – denominator in Eq. 11). The optimal number of clusters is where V_{XB} is at its minimum.

For the interpretation of the SOM, box plots of the performance measures for each cluster, the occurrence of the clusters in the time series and a visual inspection of the SOM are used.

3 Study areas

3.1 The Weisseritz catchment

For the first case study, the catchment of the Wilde Weisseritz, situated in the eastern Ore Mountains at the Czech-German border was used (Fig. 3a). The lowest gauging station used in the study was Ammelsdorf (49.3 km^2). The study area has an elevation of 530 to about 900 m a.s.l. and slopes are gentle with an average of 7°, 99% are

<20°; calculated from a 90 m digital elevation model (SRTM, 2002). Soils are mostly campisoils. The climate is moderate. Land use is dominated by forests (≈30%) and agriculture (≈50%). Annual precipitation for this catchment is 1120 mm/year for the two years of the simulation period from 1 June 2000 until 1 June 2002. During winter, the catchment usually has a snow cover of up to about 1 m for 1 to 4 months with high flows during the snow melt period (Fig. 8a shows the pronounced peaks during spring). High flows can also be induced by convective events during summer. WASY (2006) conclude from their analysis based on topography, soil types and land use, that subsurface stormflow is likely to be the dominant process. Meteorological data for 11 surrounding climate stations was obtained from the German Weather Service (DWD, 2007). Discharge data, as well as data about land use and soil was obtained from (LfUG, 2007).

3.2 The Malalcahuello catchment

As a second case study the Malalcahuello catchment (Chile) was used. This research area is located in the Reserva Forestal Malalcahuello, on the southern slope of Volcán Lonquimay. The catchment covers an area of 6.26 km^2. Elevations range from 1120 m to 1856 m a.s.l., with average slopes of 51%. 80% of the catchment is covered with native forest. There is no anthropogenic intervention.

The soils are young, little developed and strongly layered volcanic ash soils (Andosols, in Chile known as Trumaos) (Iroumé, 2003; Blume et al., 2008). High permeabilities (saturated and unsaturated), high porosities and low bulk densities are typical for volcanic ash soils. Soil hydraulic conductivities for the soils in the Malalcahuello catchment range from $1.22*10^{-5}$ to $5.53*10^{-3}$ m/s for the top 45 cm. Porosities for all horizons sampled range from 56.8% to 82.1%. Layer thickness is also highly heterogeneous, and can range from 2–4 cm to several meters. For a more detailed description of the Malalcahuello catchment see (Blume et al., 2008).

The climate of this area is humid-temperate with altitudinal effects. There is snow at higher elevations during winter and little precipitation during the summer months

January and February. Annual rainfall amounts range from 2000 to over 3000 mm, depending on elevation. An overview of catchment topography and basic instrumentation is given in Fig. 3b.

4 Hydrological models

4.1 WaSiM-ETH

As subsurface storm flow is deemed to be a dominant process in the Weisseritz catchment, the topmodel approach (Beven and Kirby, 1979) appears suitable to conceptualise runoff generation. We therefore selected WaSiM-ETH, which is a modular, deterministic and distributed water balance model based on the topmodel approach (Schulla and Jasper, 2001). It was used for the Weisseritz catchment with a regularly spaced grid of 100 m resolution and an hourly time step. Interception, evapotranspiration (Penman-Monteith), and infiltration (Green and Ampt approach) as well as snow dynamics are also included as modules. The unsaturated zone is described based on the topmodel approach with the topographic index (Beven and Kirby, 1979), which determines flow based on the saturation deficit and its spatial distribution, instead of modelling the soil water movement explicitly. For the exact formulations of WaSiM-ETH see (Schulla and Jasper, 2001). We used an extension by Niehoff et al. (2002) which includes macropore flow, siltation and water retention in the landscape. Direct flow and interflow are calculated as linear storage per grid cell, while baseflow is calculated as linear storage for the entire subcatchment. The snow cover dynamics are simulated with a temperature index approach (Rango and Martinec, 1995). The routing of streamflow is computed with the kinematic wave approach (Niehoff et al., 2002). The model was run with hourly time steps.

4.2 Catflow

The hillslope module of the physically based model Catflow (Zehe and Fluhler, 2001; Zehe and Bloschl, 2004; Zehe et al., 2005) was used to model runoff generation in the Malalcahuello catchment. It relies on detailed process representation such as soil water dynamics with the Richards equation (mixed form), evapotranspiration with the Penman-Monteith equation, surface runoff with the convection diffusion approximation to the 1D Saint Venant equation. The processes saturation and infiltration excess runoff, reinfiltration of surface runoff, lateral subsurface flow and return flow can be simulated. Macropores can be included with a simplified effective approach (Zehe et al., 2001). The simulation time step is dynamically adjusted to achieve a fast convergence of the Picard iteration. The hillslope is discretized as a 2-D vertical grid along the main slope line. This grid is defined by curvilinear coordinates (Zehe et al., 2001). As the hillslope is defined along its main slope line each element extends over the whole width of the hillslope, making the representation quasi-3-D. Catflow has proved to be successful for a number of applications (Zehe et al., 2005, 2001; Lindenmaier et al., 2005; Lee et al., 2007; Graeff et al., 2008).

For this investigation the hillslope module was used to simulate a single hillslope. As the outflow at the lower end of the slope is compared with stream hydrographs measured at the main stream gauging station, this carries the inherent assumption that the structure and physical characteristics of this single slope are representative of all slopes in the catchment. While this is a strong assumption it is not completely unrealistic for the Malalcahuello catchment.

For soil parametrization values of saturated hydraulic conductivities, porosities, pF curves and fitted Van Genuchten parameters were used. Details on set-up and parametrization can be found in (Blume, 2008). 2004 data from a climate station just outside the catchment was used as climatic input data with a temporal resolution of 30 min. Rainfall time series stem from a rain gauge close to the catchment outlet.

5 Weisseritz case study – results

5.1 Performance measures

The performance measures introduced in Sect. 2.1 were calculated for the entire simulation period with a moving 10 day window (N=14821). After removing highly correlated performance measures (see Table 2), a set of 23 measures remained. The summary of the measures shows how the measures vary greatly in their range of values (Table 3). The error measures based on the differences (PDIFF, ME) are not bound at the upper or lower end and have a value of 0 for perfectly matching time series. The error measures based on absolute or squared differences (AME, MAE, RAE, MARE, MSLE, MSDE, CMSE, MAOE) start from 0 for no error and are not limited at the upper end. The geometric error (RSMSGE) is bound at the lower end at 1 and unbound at the upper end. LCS varies between 1 (for no error) and 0. The relative measures (r_k, and r_d) have a value of 1 for no error and are above or below 1 (always positive) depending on the direction of the error. The t_{test} statistics has a value of 0 for perfect agreement and has positive or negative values for underestimated and overestimated time series, respectively. The other measures have the following ranges: $-1 \leq$Rsqr ≤ 1, $-\infty \leq$CE≤ 1, $0 <$IoAd≤ 1 (1 indicates no error for these three measures), t_L and NSC are limited by the window width with a "best" value of 0, $0 \leq Q_e \leq 1$ (the value for no error is defined by the shape of the error distribution – for normally distributed errors Q_e=0.5 for no error).

5.2 Synthetic errors

The synthetic peak errors are used to improve our understanding of the performance measures. In Fig. 4, five plots show the response of some exemplary measures (y-axis) to the synthetic peak errors, each of which is shown with a different symbol. On the x-axis, no error would be in the centre and the severity of the error increases to each side. Some performance measures are very specific to a certain type of error.

9 out of 23 measures react (similar to the Nash-Sutcliffe efficiency CE in Fig. 4) to all peak errors (AME, MAE, RAE, MARE, IoAd, MSLE, MSDE, CMSE). PDIFF and PEP are insensitive to the error in recession (error 3), lag (error 4) and width (error 6), which do not change the maximum of the peak. The ME and the t_{test} show no or only little sensitivity to the lag time error (error 4) and the error in peak size with correct total volume (error 5). Rsqr, LCS and RSMSGE are insensitive to errors related to shifts (errors 2 and 8) and the former two are also insensitive to peak size errors (1). t_L, DE and MAOE are only sensitive to the lag time (error 4) and the missed/false peak (error 7). r_d and r_k (not shown) are sensitive to errors that result in a different slope at the end of the time window compared to the reference peak (errors 3, 5 and 6). Similarly, Q_e (not shown) is only sensitive to errors that result in a different level at the end of the time window (errors 2, 3, 6, 7 and 8).

5.3 Data reduction with SOM

Based on the normalized $\boldsymbol{p}^{(t)}$ of the model performance a SOM was created. The representation according to Cottrell and de Bodt (1996) is shown in the upper left corner of Fig. 5. Remember that the shape of the polygons indicates the distance between the cells and the intensity of the colour is proportional to the number of $\boldsymbol{p}^{(t)}$ represented by a cell.

The next 23 representations of the SOM (one for each of the performance measures found in Sect. 5.1) in Fig. 5 help to identify a typology of the model performance finger prints. The value associated with each cell is colour coded in grey tones starting with white for no error and ending in black at the highest deviation from the optimal value. For performance measures with a central optimal value, no error is – again – shown in white while errors are displayed in red in one direction and blue in the other direction. A careful inspection of the SOMs (Fig. 5) allows identification of patterns that are related to certain errors. For instance on the right hand side and the centre of the SOM positive lag times can be found. In the bottom right corner the model strongly overestimates observed peaks as indicated by negative values for t_{test} and ME, PEP,

and PDIFF. However, a clear interpretation appears still to be difficult. Hence, a further condensation of the SOMs is necessary to identify how different criteria cluster into different error classes and how we can interpret these error classes with respect to model failure.

5.4 Identification of regions of the SOM

In order to cluster the SOM for further identification of error classes, fuzzy c-means clustering was applied to the weight vectors $\boldsymbol{v}$ of the SOM. The validity index V_{XB} for the identification of the optimal cluster number is shown in Fig. 6. Based on the V_{XB} we chose the solution with 6 clusters for further analysis. We also checked if the clustering algorithm could be applied to the $\boldsymbol{p}^{(t)}$ directly. However, we did not find satisfactory results: The validity index was lowest for two clusters with a value of about 100 – indicating no successful separation.

The 6 clusters are represented with colour coding in the SOM in the bottom right corner of Fig. 5. No $\boldsymbol{p}^{(t)}$ vectors are associated with uncoloured cells; i.e. these cells were never identified as best match to any input vector.

Looking at Fig. 5 allows us to make some first statements about the model performance found in each cluster, e.g. the overestimation by the model (negative PDIFF, ME, PEP and t_{test}) is found in cluster B. To support the interpretation of the clusters, box plots for each cluster (Fig. 7a) were created for each of the performance measures from the normalized weight vectors $\boldsymbol{v}$ of the cells in the SOM. Note that the y-axis in the box plot shows normalized values as described in Sect. 2.3, whereas non-normalized values were used for the labels in Fig. 5. The normalized weight vectors $\boldsymbol{v}$ do not span the entire range from 0 to 1 because each cell in the SOM only represents the centre of the associated $\boldsymbol{p}^{(t)}$.

The findings from the box plots are summarized in Table 4. If the median of the $\boldsymbol{v}_j$ belonging to a cluster was closest or furthest to the value for no error, this cluster was entered into the table as “best” or “worst”, respectively. “Worst” was replaced by “high”

and "low" if the deviation occurred to both sides of the optimal value. If the median of the second highest/lowest cluster was within the inner quartiles, it was also entered into the table.

Summarizing the SOMs (Fig. 5), the box plot (Fig. 7a) and Table 4 we find that cluster A shows the best fit according to over half (13) of the performance measures. In this cluster there is thus a good agreement in dynamics and amounts of simulated and observed stream flows. Furthermore, low values for r_k (the recession constant is overestimated by the model) are observed. The bad values for NSC indicate that the modelled time series changes often between under- and overestimation, which is most likely caused by small deviations. Cluster B has good values for 9 error measures indicating that the observed and modelled time series match well in terms of correlation (Rsqr, DE, LCS, t_L) and size of the peaks (PDIFF). Data agree well after centring (CMSE), ordering (MAOE) or rescaling (RSMSGE, MSLE). Low values for PEP, ME and for the t-test indicate that the model overestimates the observed data. Low values for r_k indicate, that also the recession constant is overestimated by the model. Cluster C performs well for AME, MAE, ME, PEP t_{test} and CMSE which shows that the error is always relatively small and that the maxima agree. Bad performance in terms of NSC, Rsqr, LCS as well as high lag times show that the timing is poor. This type of error might occur for small peaks where the timing is not well represented in the model. Cluster D shows good values for the t-test which indicates that the modelled and observed time series can not be distinguished based on this test, i.e. the distribution of the differences can not be distinguished from zero. However, bad values indicate that the overall magnitude (MAE), the magnitude of the peaks (AME, PDIFF), and the dynamics (MSDE, DE, CMSE, MAOE) are not reproduced well. The negative lag times indicate problems with the timing. Cluster E has good values for CMSE which indicates that the time series agree well after centring. Also the recession (r_k) is represented well. High values for ME, PEP, t_{test} and Q_e show that the model strongly underestimates the observed values. In Cluster F the dynamics are not well represented as indicated by bad values for CE, RAE, Rsqr, IoAd, r_d, MAOE and LCS. Good values for AME

and PDIFF show that the maximum errors are small. Good values for MSDE, DE and CMSE show that the derivatives and the direction of the flow is represented well.

In order to associate the synthetic peak errors (Sect. 5.2) with the error clusters, the synthetic peak errors were placed on the SOM by finding the best matching cell (BM). Table 5 shows to which clusters the synthetic peak errors belong. Level 1 to 3 corresponds to overestimated values by the model compared to the observed data (the darker grey peaks in Fig. 2) while levels 4 to 6 correspond underestimated values (to the lighter grey peaks). The short cluster descriptions in parentheses in the following paragraph are a very condensed summary of the boxplot analysis. None of the errors were placed within Cluster A (good fit between model and observation). Cluster B (model performs well but overestimates data) includes mostly the small and intermediate overestimations. Cluster C (relatively small errors but positive lag times) includes lag time errors and a number of errors which indicate underestimation as well as missing peaks. Cluster D (badly represented peaks and the negative lag times) includes a number of strong overestimating errors, strong negative lag times and modelled peaks, where no peak was present in the observed data. Cluster E (strong underestimation due to shift) includes underestimating peak errors, mainly due to shifts (in presence or absence of peaks). Cluster F (bad representation of the dynamics, small maximum error, underestimation of peaks) includes all peaks that are too narrow, and peaks overestimating the recession constant, both indicating an underestimation of the observed data.

The occurrence of the error classes in time is shown in Fig. 8a as colour bars in the discharge time series. The colour coding is equivalent to Figs. 5 and 7a. The plot shows clear patterns in the occurrence of the error classes. The following patterns were identified by visual inspection: Cluster A (model performs well) and B (model performs well, but overestimates observed data) occur during snow melt events and the fall season. Cluster C (underestimation with positive lag times) occurs mainly during fall and spring season. Cluster D (strong overestimation of peaks and negative lag times) occurs during periods with intermediate dynamics in winter. The timing and

representation is unsatisfactory for data belonging to this cluster. Further investigations (not presented here) showed that the simple temperature index approach used for the snow model is not able to capture both, temperature and radiation induced snow melt events. Therefore, while the model is able to capture some larger snow melt events (Cluster A) in a satisfactory way, other snow melt events are not represented well. This is in agreement with Rango and Martinec (1995) who report that the degree-day method can lead to errors because of the missing radiation component. Cluster E (strong underestimation due to shift) occurs during the low flow period in summer, where flows are strongly underestimated with the current model. Finally, cluster F (bad representation of dynamics, too small peaks) occurs during times where model shows dynamics which do not occur in the observed data.

The entire case study was repeated two more times, once with a time window size of 5 days and once with a 15 day window, in order to test the sensitivity of the method for this choice. Detailed results from this comparison are available from the corresponding author. In short, the alternative window sizes resulted in 3 and 4 clusters for the 5 and 15 day window, respectively. Clusters A and E were found in both cases with an equivalent descriptions of the error types and temporal occurrence of the error clusters.

6 Malalcahuello case study – results

6.1 Performance measures and synthetic errors

For the Malalcahuello case study a time window of 120 h (5 days) was chosen as streamflow here is faster in response and dynamics than in the Weisseritz catchment. A set of 17 performance measures (N=3240) remained after excluding correlated measures as well as measures that are sensitive to noise in the measured data (i.e. all measures based on derivatives). 14 of these measures were also used in the Weisseritz case study, the new ones being MRE, IRMSE and RSMSE. The 8 synthetic errors proposed in Sect. 3.2 were adapted for the time window as well as the range in flows.

6.2 SOM and fuzzy clustering

As in the Weisseritz case study, data reduction was achieved by producing a self-organizing map. The cluster analysis of the SOM regions resulted in the identification of 6 error clusters. The box plots of for each performance measure and cluster are shown in Fig. 7b. A summary of the specific characteristics of each cluster is given in Table 4. While it is difficult to identify a single "best" cluster, cluster E can easily be identified as having the worst performance measures (scores worst on 6 of the performance measures and best only on 1. Peaks as well as overall time series are underestimated (values of PDIFF, PEP and ME slightly above target value). The correlation between modelled and measured time series is low as it has the worst scores on Rsqr, t_L, and DE. Furthermore cluster D stands out as having best performances for the measures focusing on the peaks (PDIFF and PEP), while it also scores good to medium on the other performance measures. It thus describes times where the model has only slight over and underestimation in peaks, quite good correlation and quite low mean errors. Cluster A shows the best performance for those measures looking at the correlation of the time series (CE, Rsqr, DE, LCS), but also has the characteristic values for overestimating peaks (PDIFF and PEP below aim) as well as overestimating the time series in general (ME below aim). Cluster B also has good/best values for CE and Rsqr (good correlation) but strongly overestimates the time series (ME), also if measured on a relative scale (MARE) and after rescaling (RSMSE). Also, the peaks are strongly overestimated in this cluster(AME, PDIFF and PEP). Cluster C also shows good correlation (Rsqr) and little time lag (t_L) but scores low on CE (which means peaks are badly reproduced), and the relative errors MARE and IRMSE. Peaks as well as the overall time series are generally underestimated (PDIFF, PEP and ME above aim). Cluster F scores well on mean and mean relative errors (ME, MARE, MRE) as well as on the measures describing the peaks (PDIFF and PEP). In this case a slight over- or underestimation of the peaks is possible. Bad scores were achieved for the measures NSC and t_L.

Placing the error vectors produced by the synthetic peak errors (corresponding to Fig. 4) on the SOM further improves the characterization of the error clusters (see Table 5): Cluster A (good correlation, but overestimation) contains a number of overestimating synthetic errors and the earliest peak (error 4). Again, the short description in parentheses is the condensed summary of the boxplot analysis. Cluster B (good correlation, strong overestimation, also for relative and rescaled data, bad reproduction of the peak) contains all but one errors in peak size with and without correct total volume (errors 1 and 5). In addition, all but the extreme lag time errors (4) and a number of other, intermediate errors are found in this cluster. Cluster C (underestimation) contains all errors shifting the modelled below the measured time series (error 2, error 7a-level 1 and error 8). Cluster D (slight over and underestimation in peaks, good correlation, low mean errors) does not contain any of the synthetic peak errors. Cluster E (worst performance, underestimation) also does not contain any of the synthetic peak errors, however, introducing a small random noise to the reference recession in error 8 (not shown) causes all shifts below this reference to be placed in this cluster. Cluster F (small mean error, peaks well represented but with lag times) contains a number of underestimating synthetic errors (3, 6, 7b) and the latest peak (4).

Looking at the distribution of the error clusters over the time series (Fig. 8b) we find a distinct pattern of errors, which mainly occur in larger blocks. Cluster A (good correlation, but overestimation) was attributed to a longer period in April and May, while cluster B (good correlation, strong overestimation) is allocated to a series of peaks in June (high errors for PDIFF and PEP as well as AME, see Fig. 7b). Times attributed to cluster C (quite good correlation, underestimation) are the late recessions in May and August. These periods have very little dynamics and the model does indeed show a general underestimation of flow. Cluster D was characterized as quite balanced around the measured data with little deviation and did not contain any of these synthetic errors. This error occurs in shorter time blocks throughout the time series. Cluster E is attributed to the late recessions in June and August where flow as well as dynamics are underestimated. Some of these discrepancies in dynamics, especially in August, are

the result of snow melt. As Catflow does not contain a snow model, these dynamics cannot be reproduced in the simulation. This confirms the findings from the synthetic error analysis where only the introduction of variability or noise in the reference time series which was not reproduced by the model, resulted in an error that would be classified as belonging to cluster E. For the portion of the time series attributed to cluster F, the long term behaviour seems to be reproduced quite well. However, for the small short term variability during this period there seems to be little correlation between the time series (late July early August) and we find an over- as well as an underestimation of peaks. This is in line with the results in Fig. 7b: low Rsqr, but best performance on ME, MARE and MRE. Combining the findings from the three parts of our analysis: a) cluster description with the help of box plots, b) placement of synthetic errors within the SOM and thus allocation to a specific cluster as well as c) the analysis of cluster allocation over time improves our understanding of model performance and model shortcomings.

7 Summary and conclusions

We presented a new method to analyse the temporal dynamics of the performance of hydrological models and to characterize the types of errors in more detail. The methodology was applied successfully in two case studies, differing strongly in both, model type and streamflow dynamics and thus seems to be applicable for a wide range of research areas and modelling approaches. In the two case studies, we used a set of uncorrelated performance measures calculated for a moving 5 or 10 day window to characterize the temporal dynamics of the model performance (model performance finger print). A set of synthetic peak errors was used to test the sensitivity of the performance measures. Some performance measures were very specific for a certain type of error, while others reacted to all types of error. As our results showed, the combination of multiple measures provides a better characterization of the performance compared to any single measure, which agrees with the basic idea of multi-objective

calibrations.

Self organizing maps (SOM) were used to reduce the amount of data and in a subsequent step, different clusters of performance finger prints were identified. Using the raw data (before data reduction with the SOM) did not result in an acceptable separation of error clusters. The synthetic peaks were very helpful to characterize the different clusters in addition to the pattern observed from the different performance measures.

In both case studies we found 6 classes or clusters differing in various performance measures (Fig. 7). A temporal pattern of the occurrence could be identified in both cases, indicating that the model has different deviations during the different phases. For the Weisseritz case study, errors in timing (indicated by lag times) are observed more often during snow accumulation and melt periods. Acceptable agreement between modelled and observed data occurs in the winter and fall season. During low flow periods in summer, the discharge is strongly underestimated. In the Malalcahuello case study flow was found to be underestimated during the longer recession periods. In some recession periods the model completely fails to reproduce stream flow dynamics, causing attribution to a different error class. The three major events in June form a distinct group as they are strongly overestimated by the model. Both the missed dynamics as well as this strong overestimation are likely to be the result of the lacking representation of snow dynamics in the model. While some of these errors are already apparent in a first visual inspection of the model output, others are less obvious and might be overlooked. The here proposed methodology allows for a simple classification of these errors and at the same time gives a clear indication of what type of errors are occurring at what time. This way even less obvious errors can be found to appear repeatedly over time and especially these patterns of error repetition are likely to contain valuable information if they can be connected with parameter sensitivities.

The next step will thus be to combine the analysis of the temporal dynamics of model performance with the analysis of the temporal dynamics of parameter sensitivity in order to enhance our understanding of the model. The model performance will tell us during which periods the model is failing while the parameter sensitivity will show which

model component is the most important during this periods. Overall the methodology presented here proves to be viable and valuable for the analysis of the temporal dynamics of model performance.

Acknowledgements. This study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Färderaktivität "Risikomanagement extremer Hochwasserereignisse" (RIMAX). We would like to thank Jenny Eckart for her support with the data preprocessing for WaSiM-ETH. A major part of the analysis was carried out with the free statistical software R and contributed packages, we would like to thank its community.

References

Beven, K. and Kirby, M.: A physically based variable contributing area model of basin hydrology, Hydrol. Sci. B., 24, 43–69, 1979. 3183

Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 272 pp., 1981. 3180

Blume, T.: Hydrological processes in volcanic ash soils – Measuring, modelling and understanding runoff generation in an undisturbed catchment, Ph.D. thesis, University of Potsdam, 2008. 3184

Blume, T., Zehe, E., Reusser, D., Bauer, A., Iroumé, A., and Bronstert, A.: Investigation of runoff generation in a pristine, poorly gauged catchment in the Chilean Andes. I: A multi-method experimental study, Hydrol. Process., 22, 3661–3675, 2008. 3182

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. 3173

Buytaert, W., Reusser, D., Krause, S., and Renaud, J.-P.: Why can't we do better than Topmodel?, Hydrol. Process., 22, 4175–4179, 2008. 3171

Choi, H. T. and Beven, K.: Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, J. Hydrol., 332, 316–336, 2007. 3172, 3173, 3181

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework

to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008. 3173
Cloke, H. and Pappenberger, F.: Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures, Meteorol. Appl., 15, 181–197, 2008. 3174, 3178
Cottrell, M. and de Bodt, E.: A Kohonen map representation to avoid misleading interpretations, in: 4th European Symposium on Artificial Neural Networks, available at: http://www.dice.ucl.ac.be/esann/proceedings/papers.php?ann=19%96, 1996. 3180, 3186
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. 3173, 3175, 3178, 3199
DWD: Deutscher Wetter Dienst (German Weather Service) Climatological data for 11 cllimate stations around the Weisseritz catchment, data, 2007. 3182
Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi:10.1029/2006WR005563, 2008. 3171
Graeff, T., Zehe, E., Reusser, D., Lück, E., Schröder, B., Wenk, G., John, H., and Bronstert, A.: Process identification through rejection of model structures in a mid-mountainous rural catchment: observations of rainfall-runoff response, geophysical conditions and model intercomparison, Hydrol. Process., accepted, 2008. 3184
Gupta, H., Beven, K., and Wagener, T.: Encyclopedia of Hydrological Sciences, Model Calibration and Uncertainty Estimation, John Wiley & Sons, chap. 131, 1–17, 2005. 3171
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. 3171, 3173, 3174
Haykin, S.: Neural networks – A comprehensive foundation, Self-organizing maps, Prentice-Hall, 2nd edn., chap. 9, 425–474, 1999. 3174
Herbst, M. and Casper, M. C.: Towards model evaluation and identification using Self-Organizing Maps, Hydrol. Earth Syst. Sci., 12, 657–667, 2008, http://www.hydrol-earth-syst-sci.net/12/657/2008/. 3174
Iroumé, A.: Transporte de sedimentos en una cuenca de montaña en la Cordillera de los Andes de la Novena Regíon de Chile, Bosque, 24, 125–135, 2003. 3182
Jachner, S., van den Boogaart, K. G., and Petzoldt, T.: Statistical Methods for the Qualitative

Assessment of Dynamic Models with Time Delay (R Package qualV), J. Stat. Softw., 22, 1–30, 2007. 3173, 3175, 3176, 3199
Kohonen, T.: Self-Organizing Maps, in: Series in Information Sciences, Springer, Heidelberg, 2nd edn., Vol. 30, 521 pp., 1995. 3174, 3179
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/. 3173, 3178
Lee, H., Zehe, E., and Sivapalan, M.: Predictions of rainfall-runoff response and soil moisture dynamics in a microscale catchment using the CREW model, Hydrol. Earth Syst. Sci., 11, 819–849, 2007, http://www.hydrol-earth-syst-sci.net/11/819/2007/. 3184
LfUG: Landesamt für Umwelt und Geologie Sachsen (State office for environment and geology), Data about land use, soils, discharge, and the digital elevation model, data, 2007. 3182
Lindenmaier, F., Zehe, E., Dittfurth, A., and Ihringer, J.: Process identification at a slow-moving landslide in the Vorarlberg Alps, Hydrol. Process., 19, 1635–1651, 2005. 3184
Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, available at: http://www.sciencedirect.com/science/article/B6V6C-487FF7C-1XH/1/75ac51a8910cad95dda46f4756e7a800, 1970. 3171
Niehoff, D., Fritsch, U., and Bronstert, A.: Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany, J. Hydrol., 267, 80–93, available at: http://www.sciencedirect.com/science/article/B6V6C-46HBKF8-2/2/e7d43db548caa8d7c0ee195052aa4e98, 2002. 3183
Rango, A. and Martinec, J.: Revisiting The Degree-Day Method For Snowmelt Computations, Water Resour. Bull., 31, 657–669, 1995. 3183, 3190
Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, 2007. 3178
Schulla, J. and Jasper, K.: Model Description WaSiM-ETH, ETH-Zrich, Zrich, Switzerland, 186 pp., 2001. 3183
Shamir, E., Imam, B., Gupta, H. V., and Sorooshian, S.: Application of temporal streamflow descriptors in hydrologic model parameter estimation, Water Resour. Res., 41, W06021, doi:10.1029/2004WR003409, 2005. 3172
SRTM: Shuttle Radar Topography Mission (SRTM) Elevation Data Set, dataset, 2002. 3182

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39(8), 1214, doi:10.1029/2002WR001746, 2003. 3171, 3173

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrol. Process., 17, 455–476, available at: http://dx.doi.org/10.1002/hyp.1135, 2003. 3173, 3177

WASY: Schätzung dominanter Abflussprozesse mit WBS FLAB (Assessment of dominant runoff processes with WBS FLAB), Tech. rep., WASY Gesellschaft für wasserwirtschaftliche Planung und Systemforschung mbH and Internationales Hochschulinstitut Zittau, 10 pp., 2006. 3182

Xie, X. and Beni, G.: A validity measure for fuzzy clustering, IEEE T. Pattern Anal., 13, 841–847, 1991. 3181

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, J. Hydrol., 204, 83–97, 1998. 3171, 3173

Zehe, E. and Bloschl, G. N.: Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions, Water Resour. Res., 40, W10202, doi:10.1029/2003WR002869, 2004. 3184

Zehe, E. and Fluhler, H.: Preferential transport of isoproturon at a plot scale and a field scale tile-drained site, J. Hydrol., 247, 100–115, 2001. 3184

Zehe, E., Maurer, T., Ihringer, J., and Plate, E.: Modeling water flow and mass transport in a loess catchment, Phys. Chem. Earth Pt. B, 26, 487–507, 2001. 3184

Zehe, E., Becker, R., Bardossy, A., and Plate, E.: Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation, J. Hydrol., 315, 183–202, 2005. 3184

Table 1. List of performance measures and their abbreviations.

Abr.	Full Name	Abr.	Full Name
from Dawson et al. (2007)		from Jachner et al. (2007)	
MSE	mean squared error	CMAE	centred mean absolute error
RMSE	root mean squared error	CMSE	centred mean squared error
IRMSE	inertia root mean squared error	RCMSE	root centred mean squared error
R4MS4E	fourth root mean quadrupled error	SMAE	scaled mean absolute error
CE	Nash-Sutcliffe efficiency	SMSE	scaled mean squared error
PI	coefficient of persistence	RSMSE	root scaled mean squared error
AME	absolute maximum error	MAPE	mean absolute percentage error
PDIFF	peak difference	MSPE	mean squared percentage error
MAE	mean absolute error	RMSPE	root mean squared percentage error
ME	mean error	MALE	mean absolute log error
NSC	number of sign changes	MSLE	mean squared log error
RAE	relative absolute error	RMSLE	root mean squared log error
PEP	percent error in peak	MAGE	mean absolute geometric error
MARE	mean absolute relative error	MSGE	mean squared geometric error
MdAPE	median absolute percentage error	RMSGE	root mean squared geometric error
MRE	mean relative error	RMSOE	root mean squared ordinal error
MSRE	mean squared relative error	MAOE	mean absolute ordinal error
RVE	relative volume error	MSOE	mean squared ordinal error
Rsqr	the square of the Pearson correlation coefficient	RSMSGE	root scaled mean squared geometric error
IoAd	index of agreement	LCS	longest common sequence
MSDE	mean squared derivative error	additional measures	
t_{test}	value of the t-test statistics	t_L	lag time
		r_k	recession error
		r_d	slope error
		DE	direction error
		Q_e	error quantile

HESSD

5, 3169–3211, 2008

Table 2. Performance measures to remove based on high correlation for the Weisseritz study.

Measure to keep	Correlated measure (R>0.85) to be removed
CE	PI
AME	RMSE, R4MS4E, CMAE, MSE, RCMSE, RSMSE, SMAE
PEP	MRE, MSRE, RVE
MARE	MdAPE, MSRE, IRMSE, MAPE
MSLE	MAGE, MALE, RMSGE, RMSLE
CMSE	MSE, RCMSE, RSMSE, SMAE, SMSE
MAOE	MSOE, RMSOE
RSMSGE	RSMSLE, SMAGE, SMALE, SMSLE

Table 3. Summary of performance measures.

Measure	Min	1st Q.	Median	Mean	3rd Q.	Max
CE	−8.98	−3.38	−2.03	−2.14	−0.52	0.84
AME	0.0072	0.0328	0.0651	0.0880	0.1042	0.4367
PDIFF	−0.1966	−0.0074	0.0277	0.0383	0.0552	0.4137
MAE	0.0036	0.0149	0.0239	0.0299	0.0404	0.1139
ME	−0.0955	−0.0054	0.0113	0.0082	0.0248	0.0689
NSC	0.0	0.0	1.0	1.8	3.0	9.0
RAE	0.32	1.19	2.21	4.38	4.05	254.21
PEP	−718.2	−8.4	31.7	6.3	63.9	98.6
MARE	0.10	0.32	0.60	0.77	0.88	6.08
Rsqr	1.4e-10	1.3e−01	3.6e−01	3.8e−01	5.9e−01	9.8e−01
IoAd	3.2e-07	2.7e−01	4.0e−01	4.4e−01	6.1e−01	9.5e−01
MSLE	0.017	0.164	0.654	2.210	2.672	22.404
MSDE	3.4e-09	8.3e−07	3.1e−06	1.3e−05	1.1e−05	2.1e−04
t_{test}	−66.8	−5.7	8.7	17.4	35.1	183.4
t_l	−20.0	0.0	3.0	3.5	15.0	20.0
r_d	−1.7e+03	0.0e+00	8.2e−02	1.0e+00	1.0e+00	4.7e+03
DE	0	28	43	48	62	197
r_k	0.000	0.021	0.037	6.029	0.080	109.351
CMSE	1.3e−06	6.1e−05	2.0e−04	1.2e−03	9.4e−04	1.7e−02
MAOE	0.0017	0.1363	0.2053	0.2277	0.3069	0.5021
LCS	0.01	0.45	0.58	0.57	0.70	0.98
RSMSGE	1.1	1.3	1.5	1.7	1.9	3.7
Q_e	6.2e−05	2.4e−01	5.2e−01	5.1e−01	7.6e−01	1.0e+00

Table 4. Characterization of performance measures clusters derived from visual inspection of the SOMs in Fig. 5 and from the box plots in Fig. 7a and b.

Cluster	Description	
Weisseritz Case Study		
A	best:	CE, ME, RAE, PEP, MARE, Rsqr, IoAd, MSLE, t_{test}, r_d, MAOE, LCS, Q_e;
	worst:	NSC, r_k
B	best:	PDIFF, Rsqr, MSLE, t_L, DE, CMSE, MAOE, LCS, RSMSGE;
	worst:	r_d, r_k;
	low:	ME, PEP, t_{test}, Q_e
C	best:	AME, MAE, ME, PEP, t_{test}, CMSE, Q_e;
	worst:	NSC, Rsqr, r_k, LCS;
	high:	t_L
D	best:	t_{test};
	worst:	AME, PDIFF, MAE, MSDE, r_d, DE, r_k, CMSE, MAOE;
	low:	t_L
E	best:	NSC, r_k, CMSE;
	worst:	MARE, MSLE, RSMSGE;
	high:	ME, PEP, t_{test}, t_L, Q_e
F	best:	AME, PDIFF, MSDE, DE, r_k, CMSE, Q_e;
	worst:	CE, RAE, Rsqr, IoAd, r_d, MAOE, LCS;
	low:	PEP
Malalcahuello Case Study		
A	best:	CE, MARE, Rsqr, DE, MAOE, LCS;
	low:	MRE, t_{test}, Q_e
B	best:	CE, Rsqr, IRMSE, MAOE, LCS;
	worst:	AME, NSC, MARE, RSMSE;
	low:	PDIFF, ME, PEP, MRE, t_{test}
C	best:	AME, NSC, Rsqr, t_L, MAOE, RSMSE, Q_e;
	worst:	CE, MARE, IRMSE, LCS;
	high:	PDIFF, ME, PEP, MRE, t_{test}
D	best:	PDIFF, Rsqr
E	best:	NSC;
	worst:	MARE, Rsqr, t_L, DE, MAOE;
	high:	PDIFF, ME, PEP, MRE, Q_e
F	best:	ME, PEP, MARE, MRE, t_{test};
	worst:	NSC, t_L

HESSD

5, 3169–3211, 2008

Table 5. Cluster allocation of synthetic peak errors. For details on peak characteristics see Figs. 2 and 4. Levels 1–3 generally overestimate flow while levels 4–6 underestimate it.

		Weisseritz						Malalcahuello					
nr.	error	1	2	3	4	5	6	1	2	3	4	5	6
1	peak size	D	B	B	C	C	C	A	B	B	B	B	B
2	shift	D	B	B	C	E	E	A	A	A	C	C	C
3	recession	B	B	B	F	E	F	A	A	B	F	F	F
4	lag	D	B	B	B	C	C	F	B	B	B	B	A
5	size./integr.	D	B	B	B	C	C	B	B	B	B	B	B
6	width	B	B	B	F	F	F	A	A	B	B	F	F
7	peak/no peak	B	D	D	C	C	E	C	A	A	B	F	F
8	late recession	B	B	B	E	E	E	A	A	A	C	C	C

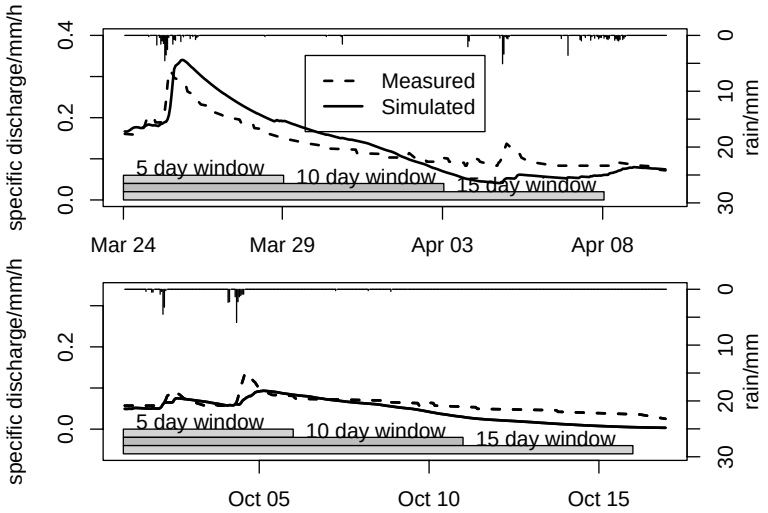


Fig. 1. Size of the selected time window with respect to observed events (Case study Weisseritz catchment).

HESSD
5, 3169–3211, 2008

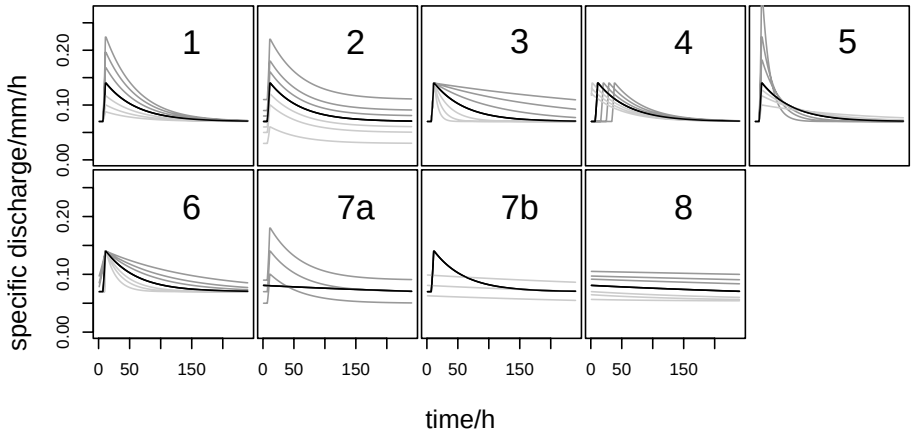
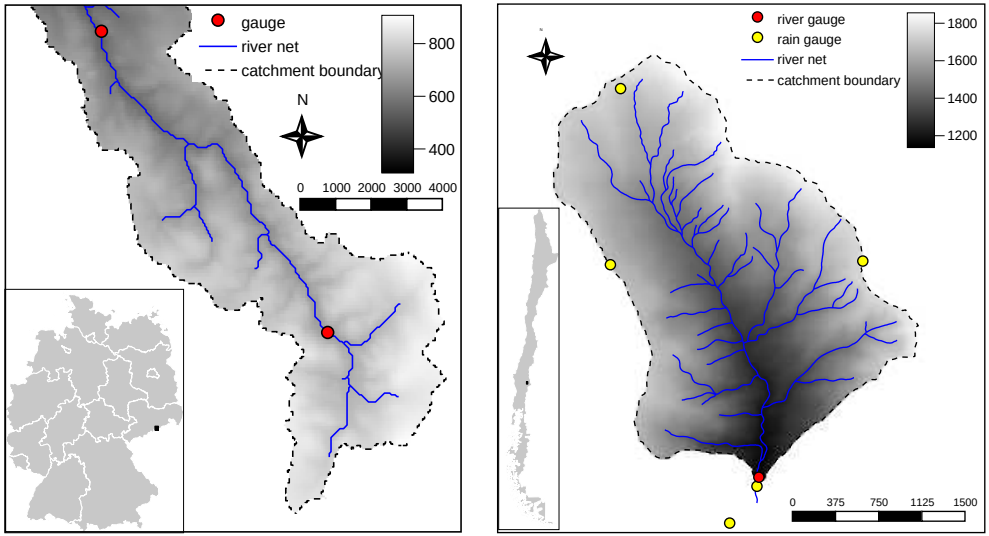


Fig. 2. Examples of synthetic errors for a single peak event: Peak over- or underestimation (1), baseflow over- or underestimation (2), recession too fast or too slow (3), timing: too late or too early (4), maximum peak flow over- or underestimation, but with correct total volume (5), peak too wide (start too early, recession too slow) or too narrow (6), erroneously simulated peak (7a) or missed peak (7b), and over- or underestimation during a late recession phase (8). The dark grey peaks will be labelled 1 to 3 with decreasing error in the remainder of this paper, while light grey peaks will be labelled 4 to 6 with increasing error.

HESSD
5, 3169–3211, 2008



(a) Wilde Weisseritz (b) Malalcahuello

Fig. 3. Maps of both research catchments (scales in m).

HESSD
5, 3169–3211, 2008

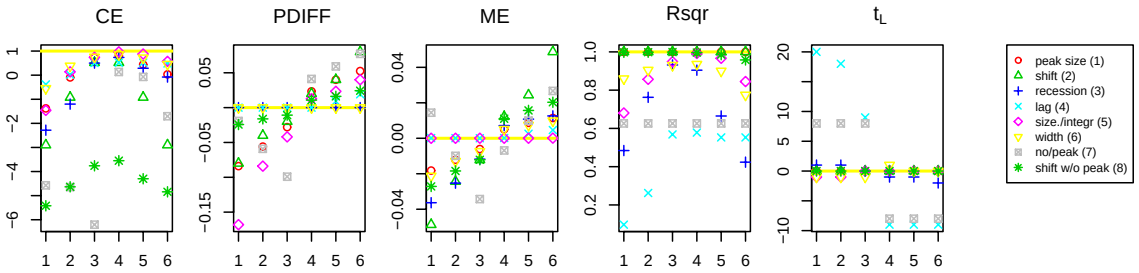


Fig. 4. Performance measures for synthetic peak errors. Along the x-axes, the degree of error varies, with index 1 to 3 indicating a peak that is much (some, little) too large (shift to too high discharges, too slow recession, too late, too wide) and 4 to 6 indicating too small peaks. The yellow line indicates the position of "perfect fit".

HESSD
5, 3169–3211, 2008

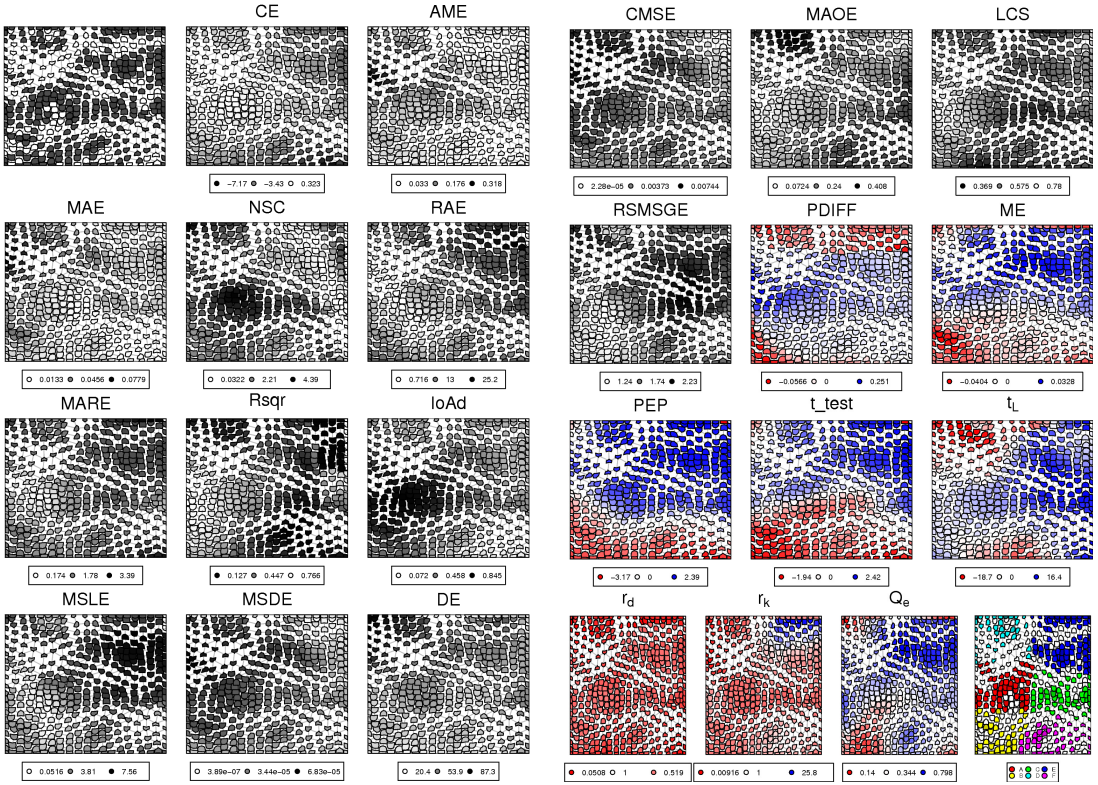


Fig. 5. Top-left: self-organizing map of the performance ,,finger prints" (containing 23 measures) for all $N = 14\,821$ 10-day time windows; bottom-right: locations of error clusters on this SOM (see Sect. 5.4); all other plots show the median of the performance measure values attributed to each cell of the SOM, white cells indicate no error, increasing saturation of grey (for single sided performance measures), and blue and red (for double sided performance measures) indicate increasing deviation from optimal performance (see Sect. 5.3 for more details).

HESSD
5, 3169–3211, 2008

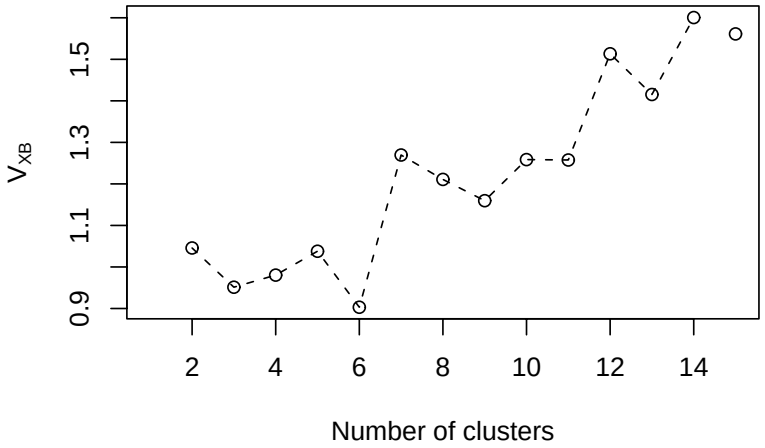


Fig. 6. Validity index for the identification of the optimal cluster number for c-means clustering (Weisseritz case study).

HESSD
5, 3169–3211, 2008

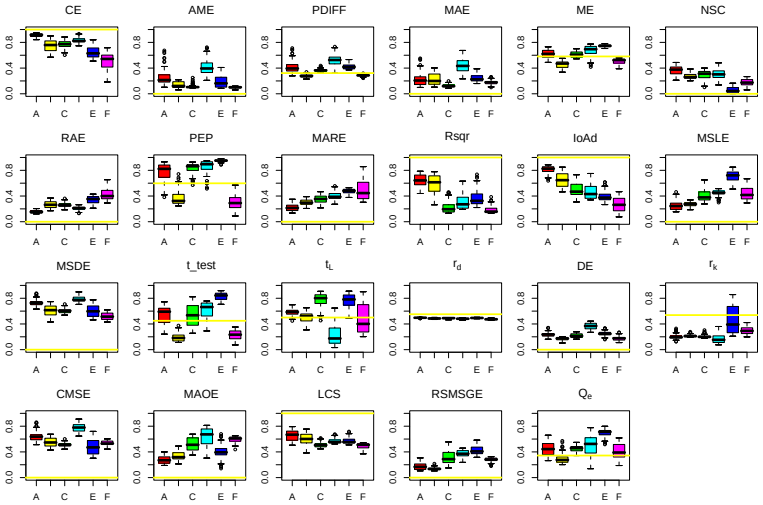


Fig. 7. Matrix of box plots comparing the normalized error measure values v_j attributed to the cells in each of the performance measure clusters (see Sect. 2.3 and 2.4). The yellow line indicates the "perfect fit" for each of the performance measures.

HESSD
5, 3169–3211, 2008

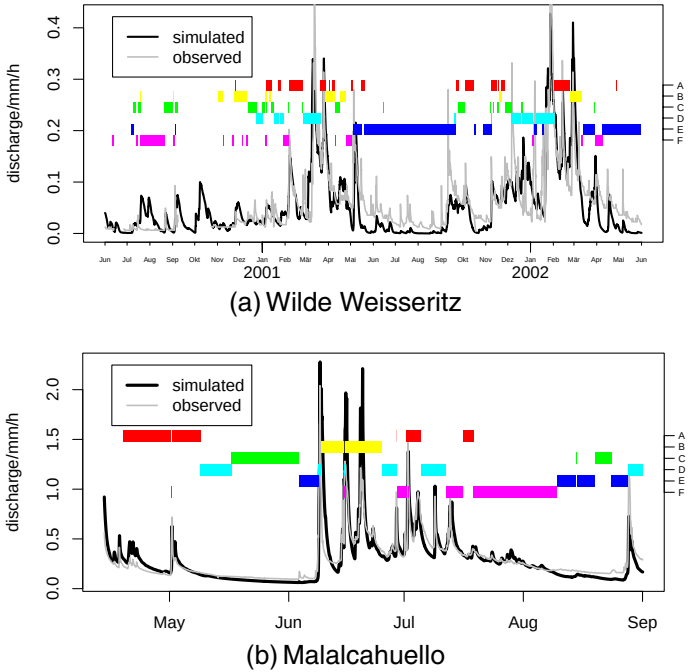


Fig. 8. Simulated and observed discharge series. The colour bars indicate the error class during this time period.

HESSD

5, 3169–3211, 2008