Hydrol. Earth Syst. Sci., 13, 999–1018, 2009
www.hydrol-earth-syst-sci.net/13/999/2009/

Hydrology and Earth System Sciences

Analysing the temporal dynamics of model performance for hydrological models

D. E. Reusser[1], **T. Blume**[1,2], **B. Schaefli**[3], **and E. Zehe**[4]

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

Received: 16 September 2008 – Published in Hydrol. Earth Syst. Sci. Discuss.: 19 November 2008
Revised: 12 June 2009 – Accepted: 12 June 2009 – Published: 7 July 2009

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.

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

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 is ideally based on data and process observations as it requires a thorough understanding of how the catchment functions. 2) Turning these concept into 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 code 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 stepwise improve a model). However, a revision of our model concept requires a clear understanding of the model's structural deficits: what is going wrong, when does it go wrong and which part of the model is the origin?

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 improvement. 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 i.e. in the order of 2 to 36 h. To our knowledge, there are no studies on high resolution temporal dynamics of model performance for longer simulation periods. Pebesma et al. (2005) analyzed the temporal dynamics of the difference between observed and predicted time series for single events and used linear models to predict these differences. For longer simulation periods, it has been shown 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 about 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 rationale behind this study is that we can obtain 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 (see Sect. 8). 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 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 modified natural reference time series has for example been proposed by Krause et al. (2005); Dawson et al. (2007). In contrast to the 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 79 models from a model family for their study.

The large amount of data produced in such an analysis quickly becomes overwhelming. 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.

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) and also in a comparable meteorological application where the bias of model results was determined conditional to the climatological input data (Abramowitz et al., 2008). The use of SOMs leads to a reduction of 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 N 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.

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 analysis is performed using the full set of measures. However, only a subset of the measures is reported for readability, excluding highly correlated measures.

In the present study we propose a novel combination of key aspects of the mentioned studies as well as the use of high resolution performance measure time series and provide evidence that this is a suitable approach for model evaluation for two very different model structures.

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 (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 and discussed in Sect. 7. Main findings and suggested future tasks are summarized in Sect. 8.

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. use of synthetic peak errors to interpret the values of the performance measures, i.e. to assess their error response;
4. use of SOMs and cluster analysis for data reduction and classification of error types;
5. analysis of temporal dynamics of error types with respect to measured and modelled time series;
6. removal of performance measures that have time series showing a high correlation with other time series for reporting the results;
7. analysis and characterization of error types using box plots and synthetic peak errors.

The analysis was performed with R (R Development Core Team, 2008) and the code is available as R-package (Reusser, 2009). A detailed description of the steps of the method is given below.

2.1 Performance measures

Dawson et al. (2007) assembled 20 performance measures used in hydrology into a test suite. This test suite includes the Nash-Sutcliffe coefficient of efficiency CE, several measures based on the absolute or squared error e.g. the mean absolute error MAE and the root mean squared error RMSE. The number of sign changes of the residuals NSC was introduced by Gupta et al. (1998). It is low if there is a bias. These and more 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. The measures have been implemented in the R package (Reusser, 2009).

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 (van den Boogaart et al., implemented in R;). Their measures are mainly based on MAE, MSE and RMSE defined as follows:

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

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

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

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

Table 1. List of performance measures, their abbreviations, error response group (ERG – see Sect. 5.2 for more details), lower (LB) and upper theoretical bound (UB) as well as the value obtained for a perfect match between model and measurement (no error).

Abr.	Full Name	ERG	LB	UB	No Error
from Dawson et al. (2007)					
MSE	mean squared error	1	-Inf	Inf	0
RMSE	root mean squared error	1	0	Inf	0
IRMSE	inertia root mean squared error	1	0	Inf	Inf[a]
R4MS4E	fourth root mean quadrupled error	1	0	Inf	0
CE	Nash-Sutcliffe efficiency	1	-Inf	1	1
PI	coefficient of persistence	1	-Inf	1	1
AME	absolute maximum error	1	0	Inf	0
PDIFF	peak difference	2	-Inf	Inf	0
MAE	mean absolute error	1	0	Inf	0
ME	mean error	3	-Inf	Inf	0
NSC	number of sign changes	9	0	LOT[b]	0
RAE	relative absolute error	1	0	Inf	0
PEP	percent error in peak	2	0	Inf	0
MARE	mean absolute relative error	1	0	Inf	0
MdAPE	median absolute percentage error	1	0	Inf	0
MRE	mean relative error	3	−Inf	Inf	0
MSRE	mean squared relative error	3	0	Inf	0
RVE	relative volume error	3	0	Inf	0
Rsqr	the square of the Pearson correlation	5	-1	1	1
IoAd	index of agreement	1	0	1	1
MSDE	mean squared derivative error	6	0	Inf	0
t_{test}	value of the paired t-test statistics	3	−Inf	Inf	0
from Jachner et al. (2007)					
CMAE	centred mean absolute error	7	0	Inf	0
CMSE	centred mean squared error	6	0	Inf	0
RCMSE	root centred mean squared error	7	0	Inf	0
RSMSE	root scaled mean squared error	5	0	Inf	0
MAPE	mean absolute percentage error	1	0	Inf	0
MALE	mean absolute log error[c]	1	0	Inf	0
MSLE	mean squared log error	1	0	Inf	0
RMSLE	root mean squared log error	1	0	Inf	0
MAGE	mean absolute geometric error	1	1	Inf	1
RMSGE	root mean squared geometric error	1	1	Inf	1
RMSOE	root mean squared ordinal error	5	0	Inf	0
MAOE	mean absolute ordinal error	5	0	Inf	0
MSOE	mean squared ordinal error	5	0	Inf	0
SMAE	scaled mean absolute error	5	0	Inf	0
SMSE	scaled mean squared error	4	0	Inf	0
SMALE	scaled mean absolute log error	1	0	Inf	0
SMSLE	scaled mean squared log error	7	0	Inf	0
SMAGE	scaled mean absolute geometric error	1	1	Inf	1
RSMSGE	root scaled mean squared geometric error	1	1	Inf	1
RSMSLE	root scaled mean squared log error	1	0	Inf	0
LCS	longest common sequence	5	0	1	1
additional measures					
t_L	lag time	8	−LOT	LOT	0
r_k	recession error	1	0	Inf	1
r_d	slope error	7	0	Inf	1
DE	direction error	8	0	LOT	0

[a] IRMSE becomes infinite for perfect match between model and observation. If the match is not perfect, small values are preferable
[b] determined by the length of the time series
[c] error of the log-transformed data.

both sequences. Thereby the method allows for deletions in one of the two series, i.e. characters can be ignored or missed (Jachner et al., 2007; van den Boogaart et al., for more details).

For this study, we complemented the above list of performance measures with the following set of four measures to obtain additional information: 1) The lag time t_L defined as the lag of the maximum in cross correlation, 2) 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, 3) the slope error r_d and 4) the recession error r_k based on the recession constant as derived by Blume et al. (2007). r_d and r_k are defined as:

$$r_d = \frac{\frac{dx_{\text{obs}}}{dt}}{\frac{dx_{\text{sim}}}{dt}} \tag{4}$$

$$r_k = \frac{k(x_{\text{obs}})}{k(x_{\text{sim}})} \text{ with } k(x) = -\frac{dx}{dt}\frac{1}{x} \tag{5}$$

The two measures were calculated as average over the time window used to calculate the other measures (see below). Measures 2–4) work best for "smoothed" time series where noise from the measurement on short time scales has been removed.

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 48 measures are calculated for a moving time window 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. The finger print will be similar for time windows where the difference between model and observation has similar characteristics. Identifying and characterizing periods with comparable finger prints gives a tool to:

- objectively separate periods of differing model performance;
- identify characteristics that are not easily found by visual inspection;
- find recurrent patterns of differences between model and observation in longer time series.

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). The selection is a compromise between looking for the local properties in the time series and having enough data to actually compute the values.

The vector $\boldsymbol{p}^{(t)}$ of the M performance measures was used as 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 48 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.

In order to avoid strong influence from extreme values, we transformed the values for each performance measure over all time windows to a uniform distribution in the range 0 to 1. In this transformed space, some performance measures are equivalent (e.g. MSE and RMSE). Because of this and as some performance measures behave very similarly and reporting 48 measures would make the study difficult to follow, we will report results only for a selection of the performance measures. Only one measure was used from each set of highly correlated performance measures ($|R|>0.85$ – see Sect. 5.1).

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

$$Q(t) = \begin{cases} Q_b & t < t_0 \\ Q_b * e^{(t-t_0)*k_c} & t_0 <= t < t_{\max} \\ Q_b + \left(Q_b * e^{t_{\max}*k_c} - Q_b\right) * e^{(t-t_{\max})*k_r} & t_{\max} <= t \end{cases} \tag{6}$$

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, event starting time and the peak time, 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

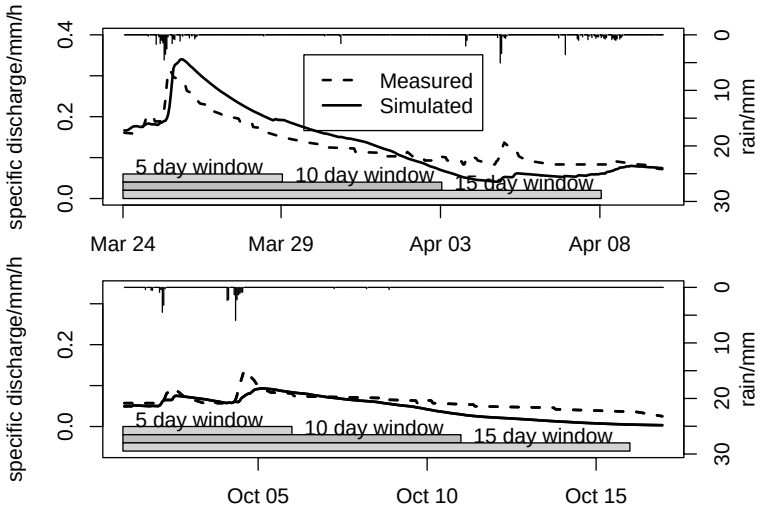


Fig. 1. Size of the selected time window with respect to two observed events (Case study Weisseritz catchment).

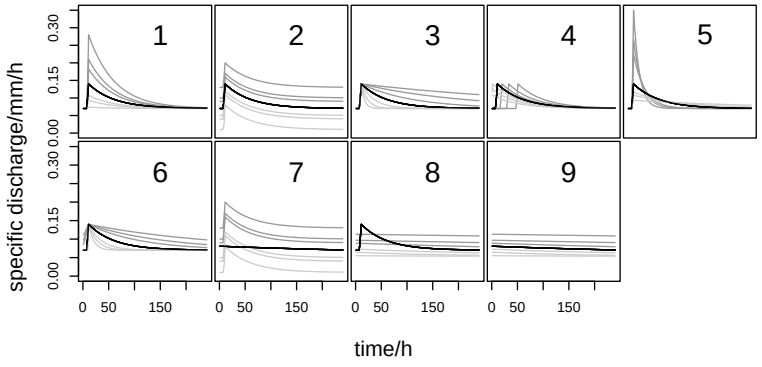


Fig. 2. Examples of synthetic errors for a single peak event: Peak over- or underestimation **(1)**, baseflow over- or underestimation **(2)**, recession too fast or too slow **(3)**, timing: too late or too early **(4)**, maximum peak flow over- or underestimation but with correct total volume **(5)**, peak too wide (start too early, recession too slow) or too narrow **(6)**, erroneously simulated peak **(7)** or missing peak **(8)**, and over- or underestimation during a late recession phase **(9)**. 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.

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 $\boldsymbol{v}=(v_j)_{j=1,\ldots,M}$ 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 $\boldsymbol{p}^{(t)}$ until the weight vectors $\boldsymbol{v}$ are stable:

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:

$$\boldsymbol{v}^{(i+1)}=\boldsymbol{v}^{i}+\sigma(x, y, \mathrm{BM}, i) * \alpha(i) * \left(\boldsymbol{p}^{(t)}-\boldsymbol{v}^{i}\right), \quad (7)$$

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 20×20 cells with the matrix $\mathbf{P}$ as input data (Yan, 2004; Weihs et al., 2005). As mentioned before, all measures where transformed to a uniform distribution in the range [0, 1] in order to reduce effects from the differing distribution shapes and scales.

The representation of the SOM (e.g. Fig. 5) 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; Dimitriadou et al., 2008). 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

$$J=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ki})^m||\boldsymbol{v}_k-\boldsymbol{w}_i||^2 \quad (8)$$

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:

$$V_{XB}=\frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ki})^m||\boldsymbol{v}_k-\boldsymbol{w}_i||^2}{c\left(\min_{i\neq k}||\boldsymbol{w}_i-\boldsymbol{w}_k||^2\right)} \quad (9)$$

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. 9) and as dissimilar as possible between classes (separation – denominator in Eq. 9). The optimal number of clusters is the one that minimizes V_{XB}.

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. 3). 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 cambisols. Land use is dominated by forests (≈30%) and agriculture (≈50%). The climate is moderate with mean temperatures of 11°C and 1°C for the periods April–September and October–March, respectively. 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. 9 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 the state office for environment and geology (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. 3.

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).

4.2 Catflow

The hillslope module of the physically based model Catflow (Zehe and Fluhler, 2001; Zehe and Blöschl, 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, evapotranspiration with the Penman-Monteith equation and 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 were 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 (Graeff et al., 2009; Lee et al., 2007; Lindenmaier et al., 2005; Zehe et al., 2001, 2005, 2006).

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.

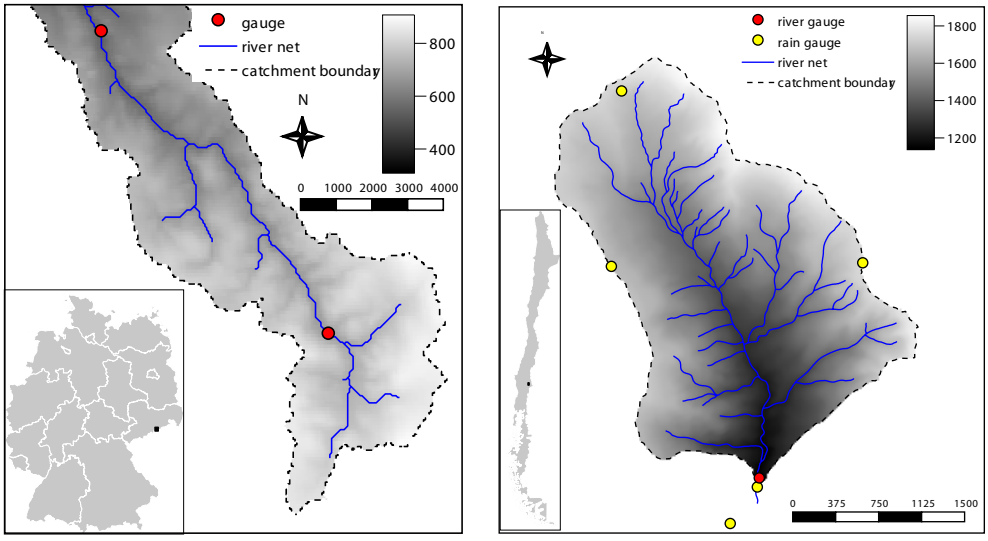
5 Weisseritz case study – results

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 (hourly time steps, 240 data points for each window, N=14 827). We repeated this 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 (Sect. 5.5). We will report only 19 performance measures (see Sect. 2.1 and Table 2). The summary of the measures shows that the ranges of the measures vary considerably (Table 3).

5.2 Synthetic errors

The synthetic peak errors are used to improve our understanding of the performance measures. In Fig. 4, nine plots show the response of some representative 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. Note that synthetic errors are generated to match the peaks of the case study (size, width, base flow). Therefore, Fig. 4 is valid for the Weisseritz case study and looks slightly different for the other case study. However, the following summary of the results also applies to the Malalcahuello case study. Some performance measures are very specific to a certain type of error. 23 out of 48 measures react to all peak errors, which is similar to the Nash-Sutcliffe efficiency CE in Fig. 4. We call this error response group (ERG) 1 (Table 1). This grouping is obtained by visual inspection of Fig. 4 and similar plots for all performance measures. The ERGs give a qualitative assessment of the measures used in this study. Measures from ERG 2 (e.g. PDIFF in Fig. 4) are insensitive to the error in recession (error 3), lag (error 4) and width (error 6). These



(a) Wilde Weisseritz (b) Malalcahuello

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

three error types do not change the maximum of the peak. Measures from ERG 3 (e.g. ME in Fig. 4) 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). SMSE (the only measure from ERG 4) is insensitive to errors related to shifts, the false peak, and peak size (errors 1, 2, 7, 9). Measures from ERG 5 (e.g. Rsqr in Fig. 4) are insensitive to errors related to shifts and peak size (errors 1, 2, 9). Measures from ERG 6 (e.g. MSDE in Fig. 4) are insensitive to errors related to shifts and shifts during the late recession phase (errors 2, 9). Measures from ERG 7 (e.g. SMALE in Fig. 4) are not sensitive for the shift only (error 2). Measures from ERG 8 (e.g. t_L in Fig. 4) are only sensitive to the lag time and the missing/false peak (errors 4, 7, 8). NSC (the only measure from ERG 9) has a value of 0 for most synthetic peak errors. Values above zero occur only if the sign of the error changes along the time series (errors 4, 5, 7, 8). The plots for all measures for both case studies are available from the first authors homepage.

5.3 Data reduction with SOM

Based on the transformed $\boldsymbol{p}^{(t)}$ of the model performance, a SOM was created. The representation according to Cottrell and de Bodt (1996) is shown in Fig. 5. Remember that the shape of the polygons indicates the distance between the cells and that the intensity of the colour is proportional to the number of $\boldsymbol{p}^{(t)}$ represented by a cell. No $\boldsymbol{p}^{(t)}$ are associated with white cells.

The 19 representations of the SOM in Fig. 6 help to identify a typology of the model performance finger prints. It is noteworthy that not all performance measures are shown (see Sect. 5.1). The value associated with each cell is colour coded using white for no error and black for 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. 6) allows identification of patterns that are related to certain errors. For example, positive lag times t_L are found in the top right corner of the SOM. In the center on the right hand side the model strongly overestimates observed peaks as indicated by negative values for t_{test} and ME, PEP, and PDIFF. However, a clear interpretation is 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 identify error classes on the SOM, 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. 7 . Based on the V_{XB}, we chose the solution with 6 clusters for further analysis. Note that the 2 and 5 cluster solutions have similar values for V_{XB}. The 2 cluster solution combines clusters A-C and D-F from the 6 cluster solution while the 5 cluster solutions combines clusters B and D from the 6 cluster solution. Therefore, the 6 cluster solution also represents the 2 and 3 cluster solutions. We also checked if the clustering algorithm could be applied to the $\boldsymbol{p}^{(t)}$ directly. For the two case studies presented here, we obtained equivalent results without SOMs. However, several test cases used during the development of the methodology suggested that the raw data is highly likely to not enable an identification of error clusters. In addition, the planned

Table 2. Performance measures to remove based on high correlation for the Weisseritz study. The table does not list all the remaining measures.

Measure to keep		Correlated measure ($\|R\|>0.85$) to be removed
RMSE	root mean squared error	AME, MAE, CMAE, R4MS4E, MSE
CE	Nash-Sutcliffe efficiency	RAE
PI	coefficient of persistence	IRMSE
MARE	mean absolute relative error	MdAPE, MRE, MSRE, RVE, MSLE, MAGE, MALE, MAPE, RMSGE RMSLE
MSDE	mean squared derivative error	CMSE, RCMSE, RSMSE, SMAE, SMSE
MAOE	mean absolute ordinal error	MSOE, RMSOE
RSMSGE	root scaled mean squared geometric error	RSMSLE, SMAGE, SMALE, SMSLE

Table 3. Summary of performance measures for the Weisseritz simulation.

Measure		Min	1st Q	Median	Mean	3rd Q	Max
PDIFF	peak difference	−0.355	−0.059	−0.014	−0.015	0.014	0.364
ME	mean error	−0.1052	−0.0287	−0.0119	−0.0172	−0.0020	0.0614
RMSE	root mean squared error	0.000	0.012	0.020	0.032	0.050	0.125
NSC	number of sign changes	0.0	0.0	1.0	1.9	4.0	11.0
PEP	percent error in peak	−343	−86	−27	−37	20	88
MARE	mean absolute relative error	6.1e-02	2.9e-01	5.0e-01	7.4e-01	1.1e+00	2.6e+00
Rsqr	square of the Pearson correlation	1.9e-08	3.1e-01	6.1e-01	5.5e-01	8.2e-01	9.8e-01
CE	Nash-Sutcliffe efficiency	−Inf	−18.27	−2.53	−Inf	−0.29	0.91
IoAd	index of agreement	0.00	0.27	0.48	0.48	0.71	0.98
PI	coefficient of persistence	−Inf	−1008.8	−269.3	−Inf	−83.4	−5.3
MSDE	mean squared derivative error	1.2e-09	8.2e-07	3.1e-06	1.1e-05	9.4e-06	1.6e-04
t_{test}	value of the paired t-test statistics	−3240.8	−44.6	−20.3	−39.7	−5.2	54.2
t_L	lag time	−20.0	0.0	1.0	2.2	5.0	20.0
r_d	slope error	−1.02	0.00	0.00	0.27	0.62	12.41
DE	direction error	0	10	24	29	41	134
r_k	recession error	0.00	0.48	1.36	1.89	2.62	14.16
MAOE	mean absolute ordinal error	0.000	0.066	0.123	0.150	0.217	0.502
LCS	longest common sequence	4.2e-03	5.4e-01	6.8e-01	6.8e-01	8.3e-01	1.0e+00
RSMSGE	root scaled mean squared geometric error	1.0	1.2	1.2	1.3	1.4	2.5

combination of the present method with a parameter sensitivity analysis (see also Sect. 8) will require an appropriate data reduction technique. We, thus, present here the full methodology including SOMs for data reduction.

The 6 clusters are represented with colour coding in the SOM in Fig. 8 . Uncoloured cells do not have any associated $\boldsymbol{p}^{(t)}$ vectors As expected, the clusters form connected regions on the SOM, since similar performance "finger prints" are placed close together on the SOM.

The temporal occurrence of the error classes is shown in Fig. 9 as colour bars in the discharge time series. The colour coding is equivalent to Fig. 8 . The plot shows clear patterns in the occurrence of the error classes, which are identified by visual inspection and described hereafter. Note that the cluster descriptions in parentheses will be further explained in the subsequent paragraphs. Cluster A (best fit, includes most synthetic peak errors) occurs mainly during late spring/early summer. Cluster B (underestimation, false peaks, differences for smaller values but good agreement for peaks) and C (dynamics well reproduced but overestimation) occur during snow melt events. Cluster D (bad reproduction of dynamics but small RMSE and maximum error) occurs mainly during late summer, fall and early winter. Cluster E (very bad agreement in terms of dynamics and volume, strong underestimation of peaks due to shift) occurs only a few times, mainly during the initial simulation period. Finally, cluster F (overestimation due to shift and false peaks, recession periods do not agree well, relative dynamics represented well) occurs during times where the model overestimates the observed data, mainly during summer and fall.

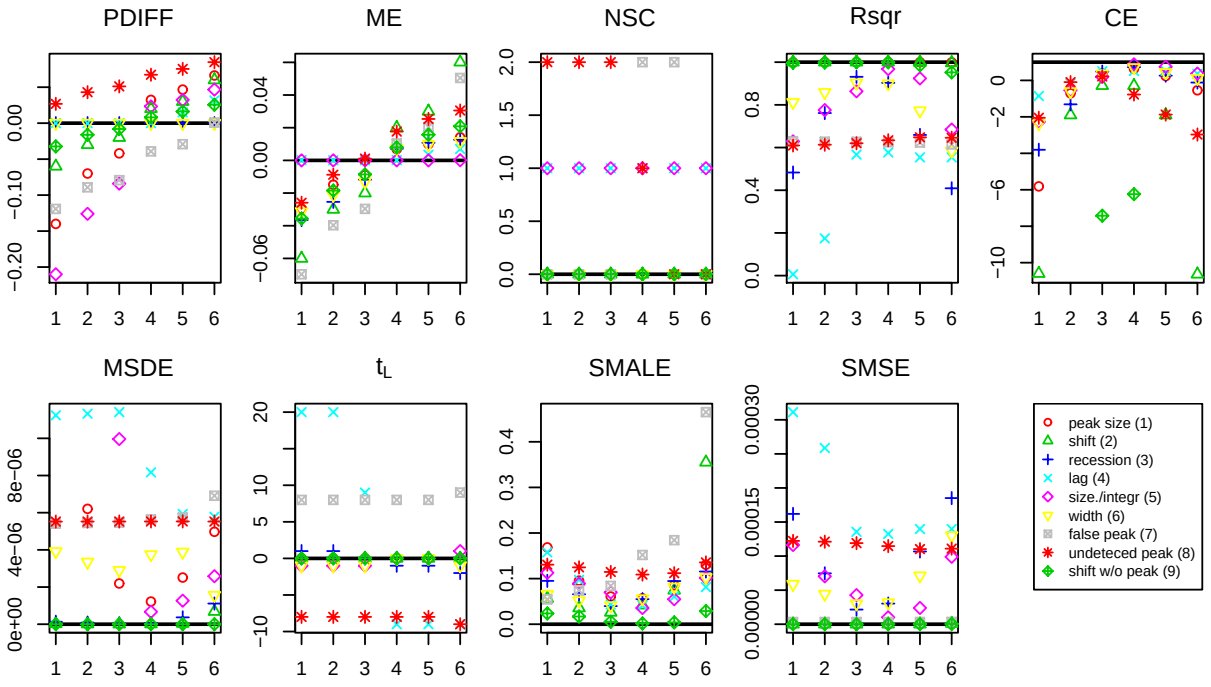


Fig. 4. Performance measures for synthetic peak errors. Along the x-axes, the degree of error varies, with index 1 to 3 indicating a peak that is much (some, little) too large (shift to too high discharges, too slow recession, too late, too wide) and 4 to 6 indicating too small peaks. The black line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

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 4 shows, to which clusters the synthetic peak errors are associated. Levels 1 to 3 correspond to overestimated values by the model compared to the observed data (the darker grey peaks in Fig. 2) while levels 4 to 6 correspond to underestimated values (to the lighter grey peaks). Cluster A includes most of the synthetic peak errors and especially the synthetic peak errors with small deviations. Cluster B includes the strong underestimation with a false peak. Cluster C includes strong overestimation due to the peak size error and errors due to undetected peaks. None of the errors were placed within Cluster D. Cluster E includes the strong underestimation of the peak due to shift. Cluster F corresponds to peaks with strong overestimation due to a shift and a shift during the late recession phase and due to false peaks. Note that cluster F is clearly related to overestimation, and Clusters B and E are clearly related to underestimation. Clusters A and C correspond to either over- or underestimation and no information is available about Cluster D from the synthetic peak errors.

Looking at the behavior of the performance measures within each cluster will provide us with more information. We therefore analyze box plots of the preformance measure values for each cluster. The box plots (Fig. 10) were created from the normalized weight vectors $\boldsymbol{v}$ of the cells in the SOM. The value for a perfect match between observation and model is shown as black line in the box plot. The normalized

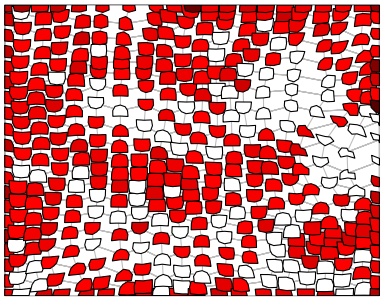


Fig. 5. Self organizing map of the performance „finger prints" (containing 48 measures) for all N=14 827 10-day time windows (Weisseritz case study).

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 box plots are read the following way: For example, looking at PDIFF, the black line indicating a perfect match between observation and model falls within the interquartile range for clusters A, B and D. Therefore, peaks are generally matched well for these clusters. However, as the interquartile range is large for cluster B, this cluster also includes cases with strong differences between peaks. Cluster E is found slightly below the black line, which indicates that peaks are generally slightly overestimated in this cluster. Clusters C and F are found far below the black line, which shows that peaks are strongly overestimated for these clusters.

Table 4. 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 Case Study		
Cluster	Error	Levels
A	peak size (1)	2 3 4 5 6
	shift (2)	2 3 4 5
	recession (3)	2 3 4 5 6
	lag (4)	1 2 3 4 5 6
	size./integr (5)	2 3 4 5 6
	width (6)	1 2 3 4 5 6
	undeteced peak (8)	2 3 4 5 6
	shift w/o peak (9)	2 3 4 5 6
B	false peak (7)	6
C	peak size (1)	1
	recession (3)	1
	size./integr (5)	1
	false peak (7)	4 5
	undeteced peak (8)	1
E	shift (2)	6
F	shift (2)	1
	false peak (7)	1 2 3
	shift w/o peak (9)	1
Malalcahuello Case Study		
Cluster	Error	Level
A	peak size (1)	1 2
	shift (2)	1 2 3
	recession (3)	3
	width (6)	1 2
	false peak (7)	1 2 3
	shift w/o peak (9)	1 2 3
B	shift (2)	5 6
	recession (3)	1 2 5 6
	lag (4)	6
	size./integr (5)	1
	width (6)	6
	false peak (7)	4 5
	undeteced peak (8)	1 2 3 4 5 6
C	shift w/o peak (9)	5 6
D	peak size (1)	5 6
	shift (2)	4
	recession (3)	4
	lag (4)	1 2 3 4 5
	size./integr (5)	2 3 5 6
	width (6)	3 4 5
E	false peak (7)	6
F	peak size (1)	3 4
	size./integr (5)	4
	shift w/o peak (9)	4

The findings from the box plots are summarized in Table 5. If the cluster median value was closest or the most distant from the perfect match value (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 and on the same side of the value for no error, it was also highlighted in the table. For the example from above, PDIFF is rated best for clusters B, D and E, and low for clusters C and F.

From the box plots (Fig. 10) and Table 5 we find that cluster A shows the best fit according to 9 performance measures. In this cluster there is thus a good agreement in (high flow) dynamics (CE, PI) and amounts (ME, RMSE, MARE, t_{test}) of simulated and observed stream flows. Peaks are late (t_L above target values) and the derivative is sometimes overestimated. LCS is the worst for cluster A. Since LCS is quite far from the optimal value for all clusters, this fact is negligible.

Cluster B has a good match between the observed and modelled time series in terms of high flows (PDIFF, CE, PI, t_{test}). Dynamics are not represented very well by the model (Rsqr, DE, MSDE), and data do not agree well after rescaling and ordering (MAOE, RSMSGE). Overall, this indicates differences for smaller values but good agreement for large values. For Cluster C, dynamics are matched reasonable (best values for PEP,Rsqr, IoAD, LCS, MAOE) but levels do not agree well (PDIFF). Also RMSE is high. For Cluster D on the other hand, the agreement is reasonable in terms of level (PDIFF, PEP, RMSE) but dynamics are not reproduced well (Rsqr, t_L, MAOE, LCS). Cluster E shows bad agreement between model and observation in terms of dynamics (Rsqr, CE, IoAd, PI, r_d, LCS) and level (t_{test}). The observed best values for PDIFF,RMSE, MSDE, t_L, DE and RSMSGE are initially somewhat surprising but can be explained by the fact that this cluster is related to low flow periods with little dynamics. In Cluster F, the level is not well represented as indicated by bad values for ME, RMSE, CE, PI, PDIFF and, PEP. Also, recession periods do not match well (r_k). Good values for r_d, DE and RSMSGE indicate that the relative dynamics are matched relatively well for cluster F.

5.5 Sensitivity for the size of the moving window and the size of the SOM

The entire case study was repeated two more times with a moving window of 5 days and 15 days, in order to test the sensitivity of the method for this choice. In short, the alternative window sizes resulted also in 6 clusters. The identified clusters had very similar error types and the temporal occurrence of the clusters was comparable to the 10 days window, the solution we retained for the present paper. In general, with smaller window sizes, the temporal occurrence of the error clusters becomes more fragmented.

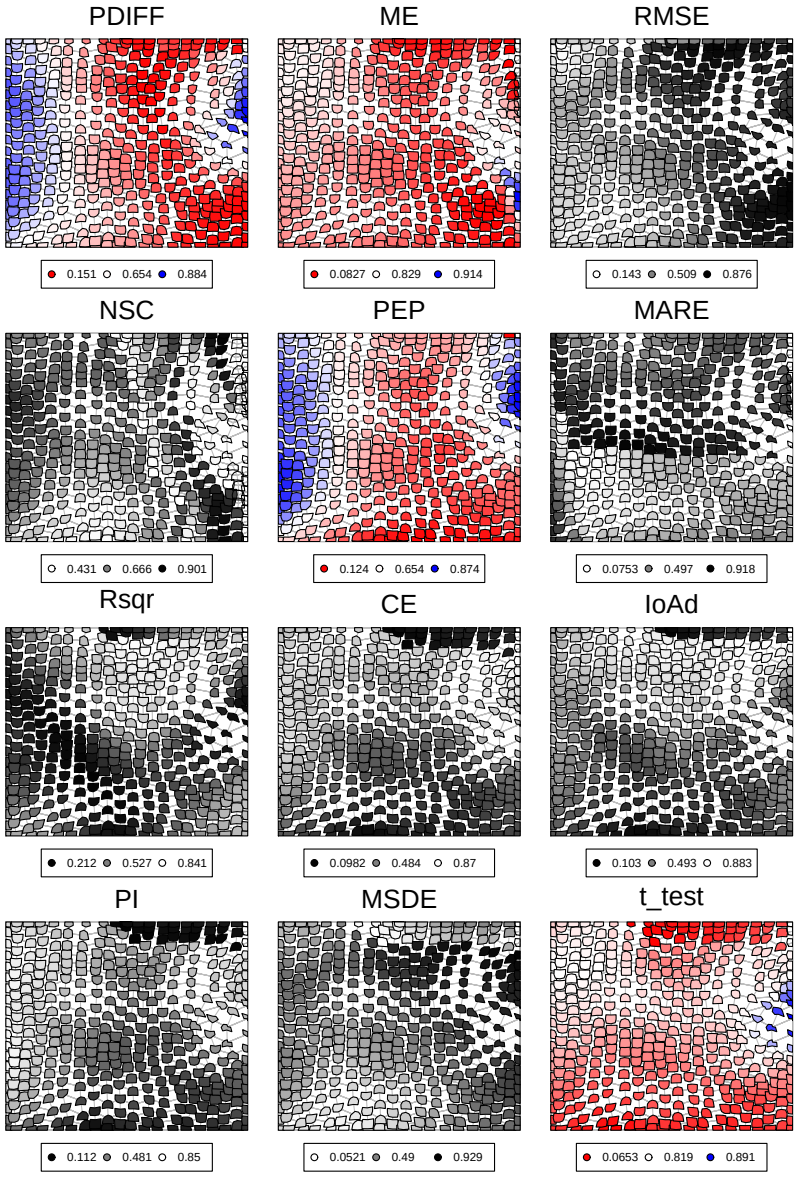


Fig. 6. Self organizing maps. The performance measure value of each cell of the SOM is color coded. White cells indicate no error, increasing saturation of grey (for single sided performance measures), and blue and red (for double sided performance measures) indicate increasing deviation from optimal performance (see Sect. 5.3 for more details.)

The entire case study was also repeated with SOM sizes of 10×10, 15×15, 25×25, 30×30, and 10×20. In this case, solutions were found for 5 or 6 clusters. The solutions with 5 clusters (30×30) combined two of the clusters presented above to a single cluster. Again, descriptions of the error types and temporal occurrence of the clusters were similar. The validity index and the interquartile ranges on the box plots (comparable to Fig. 10) were generally smaller for SOMs with a smaller number of cells because more variability was reduced during the generation of the SOM.

Detailed results (plots and tables) are available on the corresponding authors homepage at http://www.uni-potsdam.de/u/Geooekologie/institut/wasserhaushalt/hessd_homep.

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; hourly time step, 120 points) was chosen as streamflow here is faster in response and dynamics than in the Weisseritz catchment. After excluding correlated measures, a set of 16 performance measures (N=3241) remained. All of these measures were also used in the Weisseritz case study. The 9 synthetic errors proposed in Sect. 2.2 were adapted for the time window as well as the range in flows.

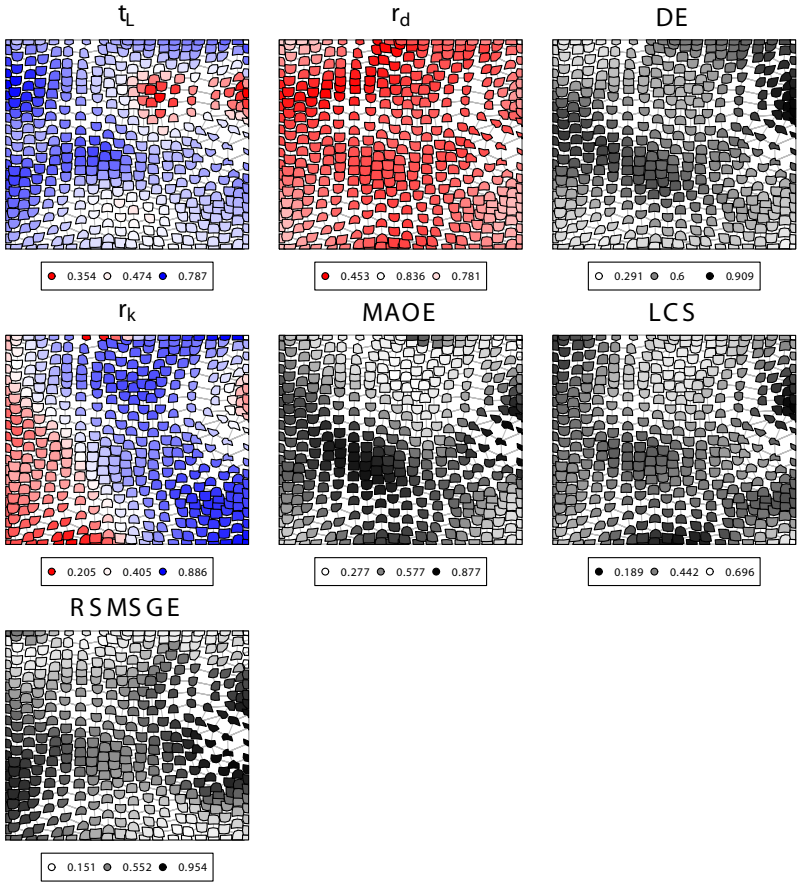


Fig. 6 . Continued.

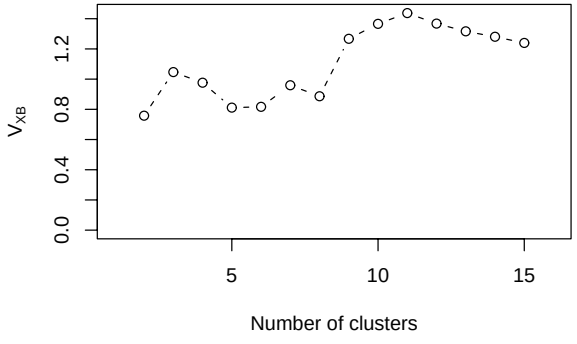


Fig. 7 . Validity index for the identification of the optimal cluster number for c-means clustering (Weisseritz case study).

6.2 SOM and fuzzy clustering

As in the Weisseritz case study, data reduction was achieved by producing a self-organizing map. 6 error clusters were identified. Looking at the distribution of the error clusters over the time series (Fig. 9) 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. Again, the descriptions in parenthesis will be further explained below. Cluster B (strong differences in peak width – including recession errors, false and undetected peaks – large errors also for rescaled data, bad performance in terms of derivatives) is allocated to a series of peaks in June. Times attributed to cluster C (small RMSE but dynamics not reproduced well, underestimation of recession phase) 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 (dynamics well reproduced, low mean errors, time lags) occurs in shorter time blocks in May and late June/beginning of July. Cluster E (worst performance, underestimation with false peaks) is attributed to the late recessions in June and August. Some of the 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. The early recession phases in May and July/August are attributed to cluster F (good reproduction of long term behaviour/balance, bad scores for the ratio of the recession constant).

Table 5. Characterization of performance measures clusters derived from visual inspection of the box plots in Fig. 10 a and 10 b.

Cluster	Description	
Weisseritz Case Study		
A	best:	ME, RMSE, MARE, CE, IoAd, PI, t_{test}, DE, r_k, RSMSGE
	worst:	t_L, r_d, LCS
B	best:	PDIFF, t_{test}, t_L, r_k
	worst:	RMSE, NSC, Rsqr, MSDE, r_d, DE, MAOE, LCS, RSMSGE
C	best:	PEP, Rsqr, IoAd, MAOE, LCS
	worst:	RMSE, r_d
	low:	PDIFF
D	best:	PDIFF, RMSE, PEP
	worst:	Rsqr, t_L, r_d, MAOE, LCS
E	best:	PDIFF, RMSE, NSC, MSDE, t_L, DE, RSMSGE
	worst:	MARE, Rsqr, CE, IoAd, PI, t_{test}, r_d, MAOE, LCS
	low:	PEP
F	best:	NSC, r_d, DE, RSMSGE
	worst:	ME, RMSE, CE, PI, LCS
	low:	PDIFF, PEP
	high:	r_k
Malalcahuello Case Study		
A	best:	Rsqr, DE, MAOE, LCS
	worst:	MARE
	low:	PDIFF, ME, t_{test}
B	best:	ME, t_{test}
	worst:	RMSE, MSDE, r_d, r_k, RSMSGE
C	best:	RMSE, NSC, Rsqr, MSDE, t_L, r_d, r_k, MAOE, RSMSGE
	worst:	CE, DE, LCS
	high:	PDIFF, ME, t_{test}
D	best:	ME, MARE, CE
	worst:	NSC, r_d, r_k
	high:	PDIFF, t_L
E	best:	NSC
	worst:	MARE, Rsqr, DE, MAOE
	low:	t_L
	high:	PDIFF, ME
F	best:	PDIFF, ME, RMSE, MARE, Rsqr, MAOE
	worst:	r_d

Locating the synthetic peak errors (corresponding to Fig. 4) on the SOM (see Table 4) leads to the following characterization: Cluster A contains most of the overestimating synthetic errors. Cluster B includes the slight underestimation due to a false peak (error 7) and the extreme peaks related to wrong recessions (error 3). In addition, the most extreme too early lag time error (error 4) and the most extreme overestimating errors due to peak size with correct integral and undetected peaks are found in this cluster. Most of these synthetic errors are related to a strong difference in peak width. Cluster C contains the most extreme error shifting the modelled below the measured time series in absence of a peak (error 9). Cluster D includes a number of intermediate/underestimating errors and all but one error related to lag times. Cluster E includes the underestimating error due to a false peak (baseline shifted far below the reference). Cluster F contains the intermediate errors related to peak size with and without correct total volume and shift during the late recession phase.

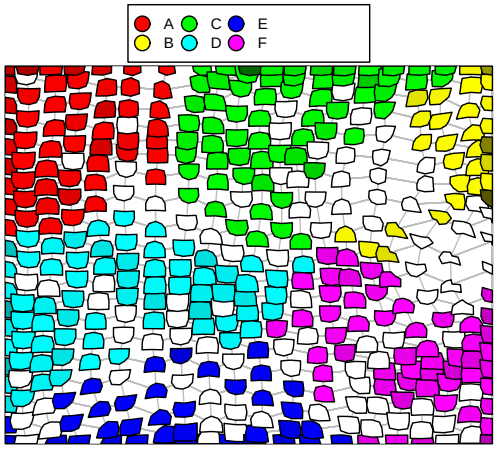


Fig. 8 . Self organizing map with color coded error cluster assignment (see Sect. 5.4)

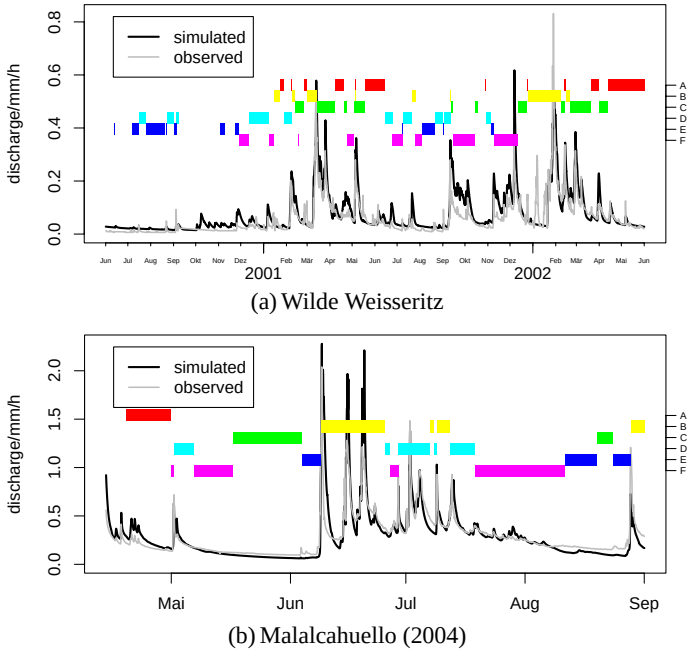


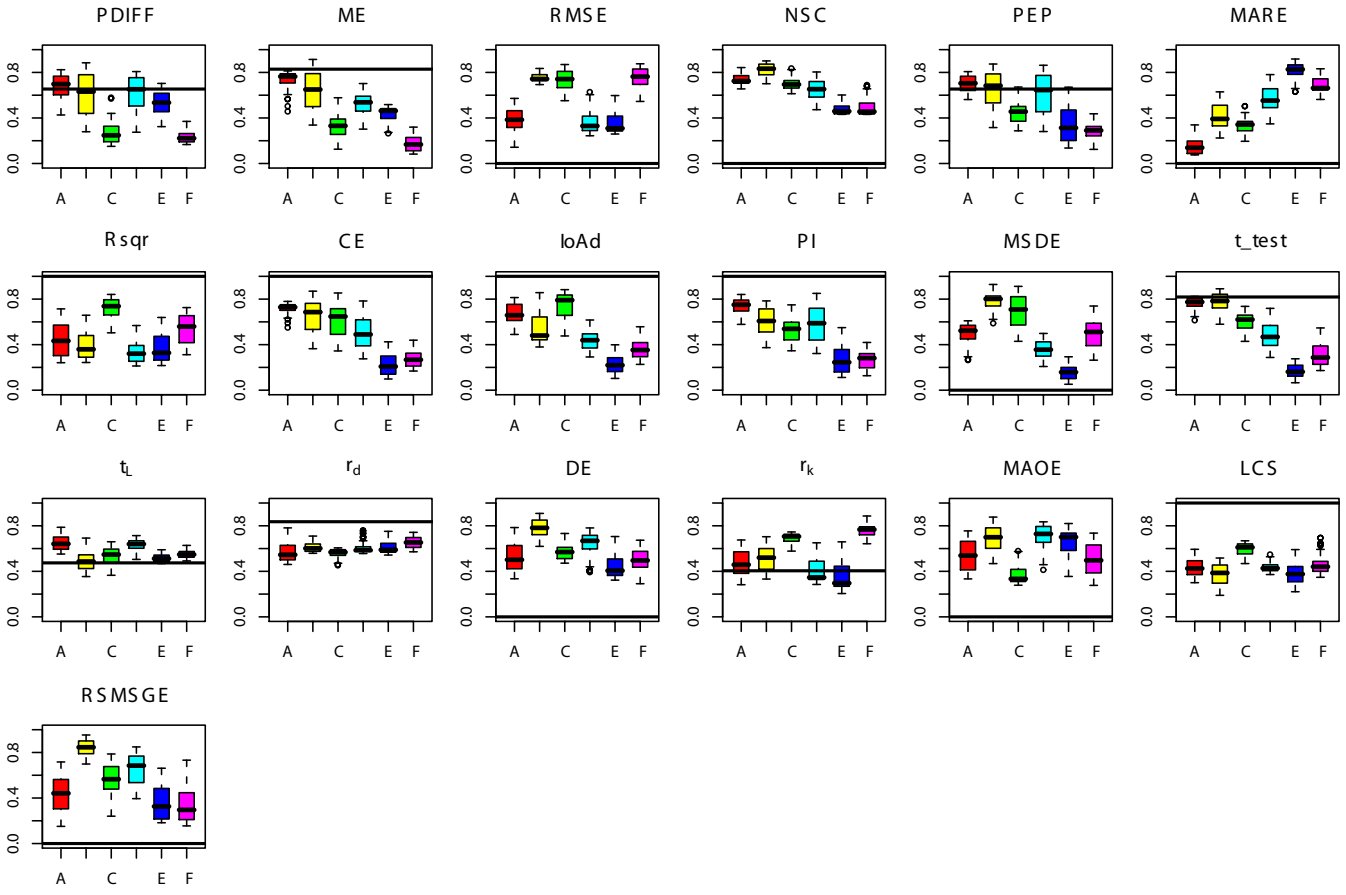
Fig. 9 . Simulated and observed discharge series. The colour bars indicate the error class during this time period.

The box plots for each performance measures and clusters are shown in Fig. 10 . A summary of the specific characteristics of each cluster is given in Table 5. Cluster A shows the best performance for those measures looking at the correlation of the time series (Rsqr, DE, LCS, MAOE) but has the characteristic values for overestimating the time series in general (ME and t_{test} below aim). Peaks are also overestimated (PDIFF below aim). Cluster B strongly overestimates the peaks (RMSE, PDIFF low) and fits the worst after rescaling (RSMSGE). Also, derivative based measures are worst for this cluster (r_k, r_d MSDE). Good values for t_{test} and ME and intermediate values for CE and Rsqr indicate that the dynamics are still reproduced quite well. Cluster C shows good performance for derivative based measures and a small RMSE but dynamics (CE, LCS) and peaks (PDIFF, ME and t_{test}) are badly reproduced. For Cluster D, dynamics (CE) and overall volume (ME, t_{test}) agree well. However, derivative based measures (r_d, r_k) show bad values. A high NSC indicates that the modelled time series changes often between lying above and below the measured time series. Cluster D thus describes times where the model has only slight over and underestimation in peaks, quite good correlation and low mean errors. Cluster E can easily be identified as having the worst performance measures (scores worst on 7 of the performance measures and best only for the NSC). Peaks as well as the overall time series are underestimated (PDIFF and ME above target value). The correlation between modelled and measured time series is low as it has the worst scores on Rsqr, MARE, MAOE, and DE. Finally, cluster F might be regarded as the best performing cluster. However, it corresponds to recession periods with little dynamics, therefore CE values are only intermediate. Scores are good for mean and mean relative errors (ME, MARE) and RMSE. However, the derivatives r_d do not match well.

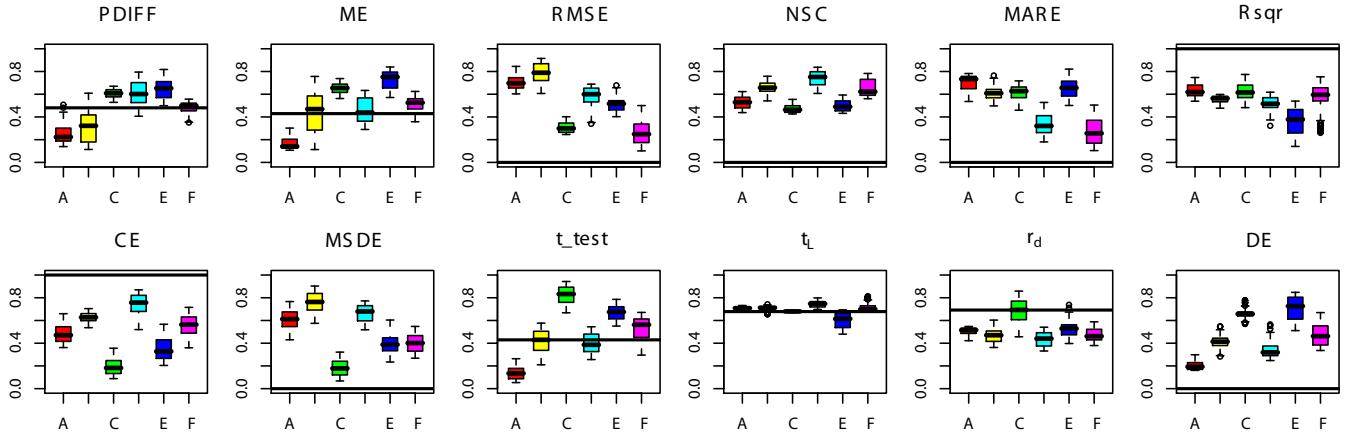
7 Discussion

In both case studies we found 6 classes or clusters of model performance (Fig. 10). A temporal pattern of the occurrence could be identified in both cases, indicating that the model has different deviations during different phases. For the Weisseritz simulation we found the following weaknesses:

- Times of "best" performance (cluster A) still show a great range of variability (most synthetic peak errors attributed to this period).
- Completely missing peaks during snow season (cluster B). More detailed analysis showed that these were events occurring at times with reported temperatures well below freezing – which must be clearly radiation induced melt events. This process is missing in the model.
- Major snow melt events are generally overestimated.
- Periods during summer/fall, where observed peaks are completely missing.
- Strong underestimation of low flow during late summer, together with



(a) Wilde Weisseritz



(b) Malalcahuello

Fig. 10 . Matrix of box plots comparing the normalized error measure values $\boldsymbol{v}$ (see Sect. 2.3). The black line indicates the "perfect fit" for each of the performance measures.

– strong overestimation of recession periods occurring during autumn, which indicates that soil and interflow storage is not well parametrized.

From this analysis, we suggest to test the following model improvements. The snow melt component may be better suited for this catchment after including radiation induced snow melt. We will check the data again very carefully for the peaks that are completely missing during summer periods. If the data is valid, we are likely to miss an important process in the model. We will also try to further improve the parametrization of the soil and interflow storage. However, as model runs take about 20 min, classical calibration methods with more than 1000 required runs are time consuming. Strong storage parameter interactions in WaSiM-ETH with the Topmodel soil storage additionally complicate calibration attempts.

For the Malacahuello case study the main findings are:

– During the first month, the model overestimates the observed discharge, indicating too high initial filling of the soil storage.

– In the recession period in August, the model completely fails to reproduce stream flow dynamics.

– The three major events in June form a distinct group as they are strongly overestimated by the model. Both the missed dynamics in August as well as this strong overestimation are likely to be the result of the lacking representation of snow dynamics in the model.

– Flow was found to be underestimated during the longer recession periods.

The first step for model improvement will be to include a snow module. The long-term storage behaviour could probably be improved by coupling the model with a ground water model. Moreover, the evaluation exercise shows that the observed discharge data needs to be preprocessed in order to remove variability/noise on the very short time scales.

While some of the identified errors are already apparent in a first visual inspection of the model output, others are less obvious and might be overlooked – especially for longer simulation periods.

8 Conclusions

This paper presents a new method to analyse the temporal dynamics of the performance of hydrological models and to characterize the types of errors. This new method is consistent with the diagnostic evaluation approach presented by Gupta et al. (2008). They suggest to use "signature indices that measure theoretically relevant system process behaviors" and argue that a single criterion is not sufficient for diagnosis of current environmental models. Instead, multiple diagnostic signatures should be derived from theory and used to compare modelled and observed behavior. This corresponds to the main idea of the performance finger prints presented in this paper.

The developed methodology combining time-resolved performance analysis and data reduction techniques is applied successfully in two case studies. These two case studies differ strongly in both, model type and runoff generation processes and thus the method seems to be applicable for a wide range of research areas and modelling approaches.

In the two case studies, a set of uncorrelated performance measures calculated for a moving 5 or 10 day window is used to characterize the temporal dynamics of the model performance (model performance finger print). As the results show, 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 calibration.

Self organizing maps (SOM) are used to reduce the amount of data and in a subsequent step, different clusters of performance finger prints are identified. These clusters are in fact not readily identifiable in the raw data data (before data reduction).

To test the sensitivity of the performance measures as well as to characterize the error clusters, the presented model diagnostics methodology includes synthetic peak errors. They show that some performance measures are very specific for a certain type of errors while others react to all types of error. Some of these errors are visible in visual inspection of the simulated and the observed reference time series. However, as illustrated for the two case studies, analyzing the temporal patterns of the identified error types gives valuable additional insights into model structural deficiencies.

In summary, the proposed methodology has the following main benefits:

– Identification and separation of time periods with different model performance characteristics are achieved in an objective way.

– Long simulation periods, for which analysis of single events becomes almost impossible can be processed. Recurrent patterns become apparent.

– Subtle but important differences between observation and model can be detected.

Especially the patterns of error repetition are likely to contain valuable information if they can be connected to 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 these periods. Overall the methodology presented here proves to be viable and valuable for the analysis of the temporal dynamics of model performance.

Acknowledgements. We would like to thank E. Pebesma, M. Clark and P. Bernardara for their valuable suggestions during the review process. This study has been funded as part of OPAQUE (operational discharge and flooding predictions in head catchments), a project within the BMBF-Förderaktivität "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 open source statistical software R and contributed packages, we would like to thank its community.

Edited by: F. Laio

References

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A.: Evaluating the Performance of Land Surface Models, J. Climate, 21, 5468–5481, 2008.

Beven, K. and Kirby, M.: A physically based variable contributing area model of basin hydrology, Hydrological Sciences Bulletin, 24, 43–69, 1979.

Bezdek, J.: Pettern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981.

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.

Blume, T., Zehe, E., and Bronstert, A.: Rainfall runoff response, event-based runoff coefficients and hydrograph separation, Hydrolog. Sci. J., 52(5), 843–862, 2007.

Blume, T., Zehe, E., Reusser, D. E., Iroume, 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.

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.

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

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.

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.

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.

Cottrell, M. and de Bodt, E.: A Kohonen map representation to avoid misleading interpretations, in: 4th European Symposium on Artificial Neural Networks, http://www.dice.ucl.ac.be/esann/proceedings/papers.php?ann=1996, 1996.

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.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A.: e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, r package version 1.5-18, 2008.

DWD: Deutscher Wetter Dienst (German Weather Service) Climatological data for 11 cllimate stations around the Weisseritz catchment, data, 2007.

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.

Graeff, T., Zehe, E., Reusser, D., Lück, E., Schräder, B., Bronstert, A., Wenk, G., and John, H.: Process identification through rejection of model structures in a mid-mountainous rural catchment: observations of rainfall-runoff response, geophysical conditions and model inter-comparison, Hydrol. Process., 23(5), 702–718, 2009.

Gupta, H., Beven, K., and Wagener, T.: Encyclopedia of Hydrological Sciences, Model Calibration and Uncertainty Estimation, John Wiley & Sons, chap. 131, 1–17, 2005.

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

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

Haykin, S.: Neural networks – A comprehensive foundation, Self-organizing maps, Prentice-Hall, 2nd edn., chap. 9, 425–474, 1999.

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/.

Iroumé, A.: Transporte de sedimentos en una cuenca de montańa en la Cordillera de los Andes de la Novena Regíon de Chile, Bosque, 24, 125–135, 2003.

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.

Kohonen, T.: Self-Organizing Maps, in: Series in Information Sciences, vol. 30, Springer, Heidelberg, 2nd edn., 1995.

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/.

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/.

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.

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.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, http://www.sciencedirect.com/science/article/B6V6C-487FF7C-1XH/1/75ac51a8910cad95dda46f4756e7a800, 1970.

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, http://www.sciencedirect.com/science/article/B6V6C-46HBKF8-2/2/e7d43db548caa8d7c0ee195052aa4e98, 2002.

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.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: http://www.R-project.org, ISBN 3-900051-07-0, 2008.

Rango, A. and Martinec, J.: Revisiting The Degree-Day Method For Snowmelt Computations, Water Resour. Bull., 31, 657–669, 1995.

Reusser, D.: tiger: Analysing TIme series of Grouped ERrors, r package version 0.1, 2009.

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

Schulla, J. and Jasper, K.: Model Description WaSiM-ETH, 2001.

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.

SRTM: Shuttle Radar Topography Mission (SRTM) Elevation Data Set, dataset, 2002.

van den Boogaart, K., Jachner, S., and Petzoldt, T.: qualV: Qualitative Validation Methods, r package version 0.2-3.

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, 1214, doi:10.1029/2002WR001746, 2003.

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, http://dx.doi.org/10.1002/hyp.1135, 2003.

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, 2006.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N.: klaR Analyzing German Business Cycles, in: Data Analysis and Decision Support, edited by: Baier, D., Decker, R., and Schmidt-Thieme, L., Springer-Verlag, Berlin, 335–343, 2005.

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

Yan, J.: som: Self-Organizing Map, r package version 0.3-4, 2004.

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

Zehe, E. and Blöschl, 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.

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.

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.

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.

Zehe, E., Lee, H., and Sivapalan, M.: Dynamical process upscaling for deriving catchment scale state variables and constitutive relations for meso-scale process models, Hydrol. Earth Syst. Sci., 10, 981–996, 2006, http://www.hydrol-earth-syst-sci.net/10/981/2006/.