Nat. Hazards Earth Syst. Sci., 9, 373–392, 2009
www.nat-hazards-earth-syst-sci.net/9/373/2009/

Natural Hazards and Earth System Sciences

Comparative analysis of model behaviour for flood prediction purposes using Self-Organizing Maps

M. Herbst[1], **M. C. Casper**[1], **J. Grundmann**[2], **and O. Buchholz**[3]

[1]Department of Physical Geography, University of Trier, Germany
[2]Department of Hydrology and Meteorology, University of Dresden, Germany
[3]Hydrotec GmbH, Aachen, Germany

Received: 19 December 2008 – Revised: 2 March 2009 – Accepted: 11 March 2009 – Published: 19 March 2009

Abstract. Distributed watershed models constitute a key component in flood forecasting systems. It is widely recognized that models because of their structural differences have varying capabilities of capturing different aspects of the system behaviour equally well. Of course, this also applies to the reproduction of peak discharges by a simulation model which is of particular interest regarding the flood forecasting problem.

In our study we use a Self-Organizing Map (SOM) in combination with index measures which are derived from the flow duration curve in order to examine the conditions under which three different distributed watershed models are capable of reproducing flood events present in the calibration data. These indices are specifically conceptualized to extract data on the peak discharge characteristics of model output time series which are obtained from Monte-Carlo simulations with the distributed watershed models NASIM, LARSIM and WaSIM-ETH. The SOM helps to analyze this data by producing a discretized mapping of their distribution in the index space onto a two dimensional plane such that their pattern and consequently the patterns of model behaviour can be conveyed in a comprehensive manner. It is demonstrated how the SOM provides useful information about details of model behaviour and also helps identifying the model parameters that are relevant for the reproduction of peak discharges and thus for flood prediction problems. It is further shown how the SOM can be used to identify those parameter sets from among the Monte-Carlo data that most closely approximate the peak discharges of a measured time series. The results represent the characteristics of the observed time series with partially superior accuracy than the reference simulation obtained by implementing a simple calibration strategy using the global optimization algorithm SCE-UA. The most prominent advantage of using SOM in the context of model analysis is that it allows to comparatively evaluating the data from two or more models. Our results highlight the individuality of the model realizations in terms of the index measures and shed a critical light on the use and implementation of simple and yet too rigorous calibration strategies.

Correspondence to: M. Herbst
(herbstm@uni-trier.de)

1 Introduction

In the course of climate change the expected increase in the occurrence of meteorological conditions that trigger extreme flood events has raised the demand for operational flood management and flood forecasting systems, also in small- to medium-sized catchments (Kundzewicz et al., 2007; Merz and Didzun, 2005). A key component of these systems is very often represented by spatially distributed deterministic hydrological modelling systems whose properties and concepts have been subject to extensive research during the HORIX project. This project aims at developing an operational expert system for flood risk management in mesoscale watersheds considering prediction uncertainty (Disse et al., 2008) and forms part of the national research programme RIMAX (RIsk MAnagement of eXtrme flood events) which is dedicated to developing and implementing instruments towards improved flood risk management (Merz et al., 2007). An important aspect of the HORIX project is also to examine to what extent and under which circumstances different hydrological modelling systems support the prediction of (extreme) flood events in river catchments (Disse et al., 2007).

The discharge simulations that are produced using deterministic hydrological models are subject to different types of

uncertainties which stem from the fact that every model is necessarily a conceptual and hence simplified representation of the natural system (e.g. Klemeš, 1983; Bossel, 2004; Sivapalan, 2005). As a consequence of this simplification, models are often not capable of covering the entire behavioural domain of the natural system with only one set of model parameters (Wagener et al., 2003). It is therefore recognized that models, because of their structural differences, have varying capabilities of capturing different aspects of the system behaviour equally well (Fenicia et al., 2007). In addition, the behavioural domain which can be reproduced by a model is further determined via calibration on historical discharge measurements which, in general, strives to account for the mean behaviour of the natural system although automatic calibration techniques (Duan et al., 2003) can emphasize, to some degree, different features of the data, depending on the performance measure which is chosen for the evaluation (Gupta et al., 1998). This, as a matter of course, excludes extreme events. Moore and Doherty (2005), however, have shown that the predictive capability of a model can be enhanced if weights are associated to those observations with the highest information content with respect to the required prediction. In order to maximize the probability that high discharge events can be simulated with a model it consequently appears reasonable to adapt the calibration strategy such that model performance in the domain of high discharges is emphasized.

In our study we use a Self-Organizing Map (SOM; Kohonen, 2001) in combination with index measures which are derived from the flow duration curve in order to examine the conditions under which three different distributed watershed models are capable of reproducing flood events present in the calibration data.

A Self-Organizing Map consists of an unsupervised learning neural network algorithm that performs a non-linear mapping of the dominant structures present in a high-dimensional data field onto a lower-dimensional grid. SOM has found almost countless applications in fields such as pattern recognition, image analysis (Kohonen, 2001) and exploratory data analysis (Kaski, 1997). However, applications related to hydrological modelling still seem to be the exception (see Minns and Hall, 2005). It has been used by Herbst et al. (2009) and Herbst and Casper (2008) for overall model evaluation and model identification purposes. Very recently, a SOM has been used by Reusser et al. (2008) to analyze the temporal dynamics of model behaviour. Kalteh et al. (2008) provide an overview of SOM applications in hydrological modelling.

In previous work in this field Herbst and Casper (2008) used the SOM to obtain a topologically ordered classification and clustering of the temporal patterns present in model outputs obtained from Monte-Carlo simulations. This clustering of entire time series allowed the authors to differentiate the spectrum of simulated time series with a high degree of discriminatory power and shows that the SOM can provide insights into parameter sensitivities, while helping to constrain the model parameter space to the region that best represents the measured time series. The major shortcoming of this approach, however, was that, in the hydrological context, the underlying criteria of this mapping ("pattern") did not provide meaningful information on the trade-offs of model behaviour. In order to improve the extraction of information in terms of interpretable time series features (see also Boyle et al., 2000) Herbst et al. (2009) linked the SOM approach to the Signature Index concept by Gupta et al. (2008). Using the Signature Indices presented by Yilmaz et al. (2008), the dissimilarities between measured and simulated time series could now be expressed in hydrologically meaningful terms referring e.g. to water balance, mean runoff reaction velocity and the volume associated to long term base flow. Consequently, the SOM of these Signature Indices provided a concise summary of model behaviour which can potentially be used for model diagnostics. The present study follows a similar approach, however, with a more specific focus: The index measures we use to compare the simulated and the observed runoff were designed with the sole purpose of extracting different characteristics in the reproduction of peak flow and do not strictly follow the Signature Index concept by Gupta et al. (2008). A SOM of these indices is used to represent the spectra of model realizations obtained from Monte-Carlo simulations with the distributed watershed models NASIM (Hydrotec, 2005), LARSIM (Bremicker, 2000) and WaSIM-ETH (Schulla and Jasper, 2001) and subsequently analyze the individual trade-offs of model behaviour in the peak flow domain. It is demonstrated how the SOM of indices provide useful information about specific details of model behaviour and also helps identifying the model parameters that are relevant for the reproduction of peak discharges. It is further shown how the SOM can be used to identify those parameter sets from among the Monte-Carlo data that most closely approximate the peak discharges of a measured time series. At the first stage of this work (Sect. 3.1) the proposed technique is applied to each of the three models individually. At the second stage (Sect. 3.2) we directly compare the model realizations which were obtained from the three models with respect to the proposed criteria. The discriminatory power of the SOM is again used to identify those model realizations that most closely match the given set of criteria; however these realizations are selected from three different modelling systems. In order to assess to what extent constraints on the parameters contribute to enhancing the predictive capabilities of the three models to discharges that exceed the range of the calibration data, the parameter sets obtained from the SOM are applied to an extreme historical flood event which has not been part of the calibration data. The paper concludes with a discussion of the potential and the shortcomings of the presented approach (Sect. 4).

2 Methods and material

2.1 Models and data

In the present study we examine the results of 12 000 Monte-Carlo simulations of hourly discharge over a period of approximately two years. The time series were generated using the distributed watershed models NASIM, LARSIM and WaSIM-ETH, i.e. for each of the models we carried out 4000 simulations for the same test catchment based on input data for the period from 1 November 1994 to 28 October 1996. The test watershed (Fig. 1) is the 129 km^2 low-mountain range catchment "Schwarze Pockau" in Saxony (Germany), a tributary of the Freiberger Mulde (Elbe sub-basin) situated near the border to Czech Republic. The catchment extends from the ridges of the Erzgebirge (Ore Mountains) at approximately 980 m.a.s.l. northward to the runoff gaging station "Zöblitz" at 440 m.a.s.l. The mean discharge at this station is 2.31 m^3/s while the highest discharge ever measured was recorded on 13.08.2002 with 160 m^3/s. The return period for events of this magnitude is estimated to 200 a. About 40% of the catchment is covered with forest. The dominant soil type is a sandy loamy cambisol. The availability of discharge measurements from this catchment, especially during the extreme event of August 2002, render this catchment a good data source to investigate the capabilities of hydrological models of reproducing extreme discharges.

The rainfall data consists of spatially interpolated, hourly precipitation fields with a resolution of 1 km^2 which were generated based on daily measurements from three gaging stations and hourly measurements from one gaging station within the area (Fig. 1). Additionally, gaging stations from outside the test-catchment (not shown in Fig. 1) were included in order to assure proper conditions at the boundaries of the field. First, a two-dimensional external drift kriging (EDK 2D) is carried out on the daily measurements to get the estimate of the daily areal precipitation. In order to account for the temporal characteristics of the precipitation field additionally a separate EDK 2D is performed on the hourly precipitation measurements. Subsequently, the daily measurements are disaggregated according to the temporal distribution of the interpolated hourly precipitation. In both interpolations the square root of the elevation was used as drift parameter. EDK 2D was also applied in order to generate the spatio-temporal fields of wind speed, however, in this case elevation data determined the drift in a linear way. For the interpolation of global radiation and relative air humidity measurements a two-dimensional ordinary kriging was used. Streamflow was measured at the outlet of the catchment at gaging station "Zöblitz".

Because appropriate prior information on parameter distributions was missing the Monte-Carlo simulations were run using uniform random sampling. The corresponding parameter ranges as well as the fixed parameters were set based on prior knowledge acquired via manual expert calibration

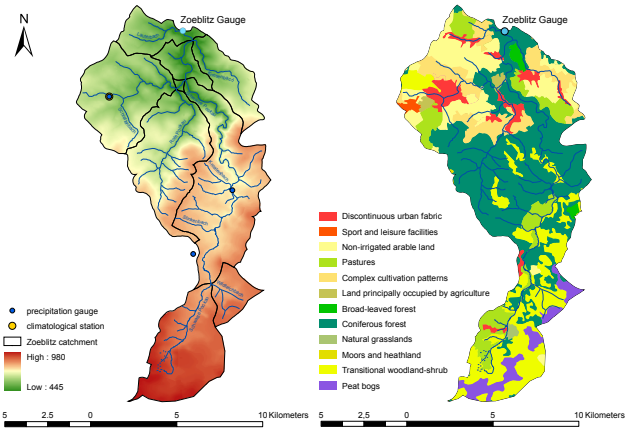

Fig. 1. The "Schwarze Pockau" low-mountain range catchment.

to the test watershed. It is assumed that these values represent the plausible parameter space for this watershed with very high probability. All parameters are related to the soil water balance and the vertical redistribution of flow components respectively. Parameters related to flood routing have not been considered for the Monte-Carlo simulation.

In the following, we give a brief outline of the model structures that were used for our study. Each of the models has found widespread application in different fields of hydrological modelling, including operational flood forecasting, throughout Germany and other countries. LARSIM and NASIM are distributed and operated commercially.

2.1.1 NASIM

NASIM (Hydrotec, 2005) is a conceptual distributed model that uses a spatial discretization based on sub-catchments. For the "Schwarze Pockau" watershed the pre-processing of spatial data resulted in 71 sub-catchments with a mean size of approximately 1.8 km^2. These are further subdivided into spatially homogeneous units with respect to soil and land use. Each of these elementary spatial units is again vertically divided into soil layers. All vertical processes that relate to soil and land use (soil moisture accounting, including interception, evapotranspiration, infiltration etc.) are calculated on the elementary unit scale. The resulting three lateral flow components are subsequently aggregated on the sub-cachment scale each passing an individual linear storage. Two of them, the interflow and surface flow, are in a prior step transformed by convolution with the time-area relationship to integrate sub catchment characteristics into the process of flow accumulation.

An outline of the principle elements of the model structure is given in Fig. 2. Note that the NASIM parameters examined in this study are unit less factors that modify the actual internal parameter values and act on the sub-catchment scale. The internal values are either based on global default values

Table 1. NASIM model parameters used for the Monte-Carlo simulation and their parameter (factor) ranges.

Name	Description	Factor Range	Internal Range	Unit
RetBasis	Storage coefficient factor for baseflow component	0.5–3.5	500[a]	[h]
RetInf	Storage coefficient factor for interflow component	2.0–6.0	50[a]	[h]
RetOf	Storage coefficient factor for surface runoff from unsealed surfaces	2.0–6.0	1.8–5.7[b]	[h]
StFFRet	Storage coefficient factor for surface runoff from urban areas	2.0–6.0	16[a]	[min]
hL	Horizontal hydraulic conductivity factor	2.0–8.0	1.5[a]	[mm]
maxInf	Maximum infiltration rate factor	0.025–1.025	11, 23, 30[c]	[mm/h]
vL	Vertical hydraulic conductivity factor	0.005–0.105	11, 23, 30[c]	[mm/h]

[a] fix value, not determined via pre-processing
[b] depending on sub-catchment slope
[c] depeding on soil type

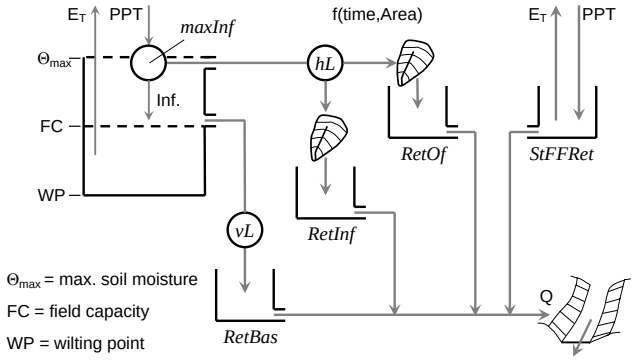

Fig. 2. Simplified schematic representation of the NASIM model structure; only those elements are reproduced that are considered in the scope of the present study. Parameters that have been subject to variation in the course of the Monte-Carlo simulation are printed in italic Times New Roman.

or have been determined individually for each sub-basin in the course of the spatial data pre-processing. The variation of these factors during the Monte-Carlo simulation, however, was performed with global values for all sub-catchments. Table 1 provides an overview of the calibration factors, internal values and their corresponding ranges. The parameter *maxInf* determines the maximum infiltration rate of the soil-moisture storage whereas the drainage is controlled by *vL*. The factors *RetOf*, *RetInf* and *RetBas* scale the storage coefficients for the quick, intermediate and the slow flow component, respectively. In the context of simulating flood events parameter *hL* potentially adopts a crucial role by determining the separation of exess flow into quick "overland flow" and intermediate "interflow" component. A special feature in NASIM is the representation of fast flow components from impervious urban areas whose retention is influenced by parameter *StFFRet*. However, in the Schwarze Pockau catchment only 6.6% of the area belongs to this land use type. Thus a dominant influence of this parameter is not expected. The internal values modified by *RetOf* are determined in the course of the pre-processing depending on the slope in each sub-basin, while the internal *RetInf*, *RetBas*, *StFFRet* are set to global values. The correspondents of *maxInf* as well as well as *vL* are determined according to soil type. The ranges of calibration factors and internal parameter values of the Monte-Carlo simulation with NASIM are given in Table 1.

2.1.2 LARSIM

LARSIM (Bremicker, 2000) is operated using the same spatially distributed input data. However, in our study, a raster based spatial discretization with a resolution of 1 km was chosen. LARSIM considers coupled land use and soil compartments on a regular grid but does not explicitly account for the spatial distribution of soil and land use related field capacities on the sub-catchment scale. Instead, the amount of water which is allowed to infiltrate per time step is given as the difference between effective rainfall and overland flow. The sum of field capacity and air capacity yield the maximum soil moisture content. LARSIM then simulates the soil moisture balance using a variable contributing area function, similar to the approach implemented by the Xinanjiang model (Zhao, 1977): The proportion of contributing saturated areas is calculated as a function of mean soil water content and a conceptual parameter *BSF* (which is an exponent that controls the shape of the contributing saturated area function, see Fig. 3). The resulting total amount of saturated flow is subsequently partitioned into a quick and a slow sub-component, Q_{of} and Q_{of2}, depending on the threshold parameter $A2$. Discharge from lateral drainage Q_i ("interflow") as well as vertical percolation Q_b is represented using non-linear, empirical relationships such that essentially all flow components are controlled by the soil moisture storage and the actual soil moisture content: in Fig. 3 W_z denotes the minimum soil moisture content to generate interflow (it is considered a constant and set to 0.7 mm). The parameters D_{min} and D_{max} determine the minimum and maximum

amount of lateral drainage from the soil-moisture storage (in mm/d) which is further governed by the actual soil-moisture content W_0. Percolation into groundwater per time step Q_b is linearly controlled by parameter β. All flow components are subsequently forwarded to linear storage elements before they reach the river channel. The parameters *EQD* and *EQD2* determine the storage coefficients that correspond to the sub-components of saturated flow Q_{of} and Q_{of2} respectively. They are linear scaling factors of the time of concentration in a sub-basin which is determined in the course of the pre-processing, i.e. they are proportional to the retention. For our study, the remaining storage coefficients are considered constant. The parameter ranges used to carry out the Monte-Carlo simulation are given in Table 2. The parameter values were identical for all sub-catchments during each run of the Monte-Carlo simulation.

2.1.3 WaSIM-ETH

WaSIM-ETH 6.4 version 2 (Schulla and Jasper, 1998) is operated with a raster based spatial discretization identical to the one used by LARSIM.

Infiltration of water into the soil is calculated for each grid cell following Peschke (1987). The remaining amount of water constitutes the surface flow component Q_d. Subsequently, soil water transport is simulated using the 1-D Richards differential equation on homogeneous soil columns which are determined by the spatial discretization. Soil hydraulic parameterization was carried out following the van Genuchten modelling scheme (Van Genuchten, 1976). The upper and lower boundary conditions are given by the amount of infiltrating water and the depth of the groundwater layer respectively. Lateral drainage Q_i results from the water balance calculations on the soil columns and is generated whenever the suction in the soil column falls below a given threshold (ψ_m=3.45 m). The drainage density parameter *dr* directly determines the amount of interflow which can be generated per time step. It expresses the drainage density of the (sub-)catchment as well as the anisotropy with regard to the vertical and horizontal hydraulic conductivities (Schulla and Jasper, 1998). For the simulations of our test catchment, however, no sub-basins were defined. A simple ground water model with a single linear storage approach is used to generate the slow discharge component Q_b. The subsequent concentration of the flow components is simulated using single linear storages and time-area functions on the catchment scale. The parameters *kd* and *ki* denote the storage coefficients of the surface runoff and the lateral flow, respectively. The resulting total discharge is calculated as the superposition of the flow components. A rough outline of the model is presented in Fig. 4. The parameter ranges for the Monte-Carlo simulation with WaSIM-ETH are reproduced in Table 3. Again, the parameter values remained identical for all sub-catchments during each simulation run.

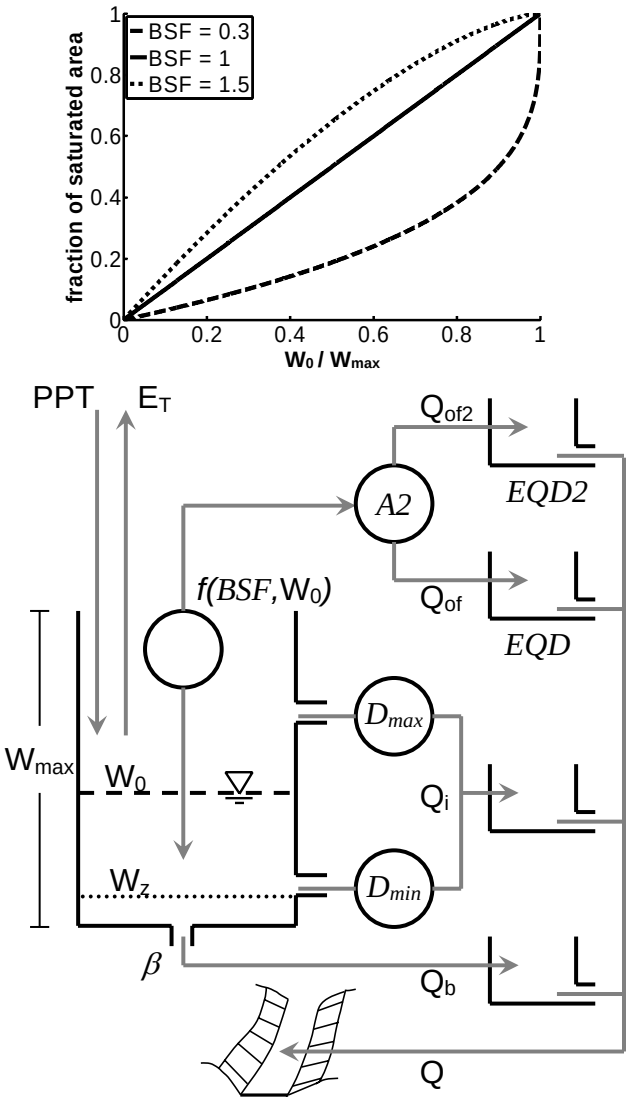

Fig. 3. Simplified schematic representation of the LARSIM model structure; only those elements are reproduced that are considered in the scope of the present study. Parameters that have been subject to variation in the course of the Monte-Carlo simulation are printed in italic Times New Roman.

As the focus of the present model evaluation lies on the reproduction of high discharges, the generation and concentration of flow through the model is considered to be the most important process here. Accordingly, the choice of model parameters for the Monte-Carlo simulation includes all parts of the particular model structures that seem to be most meaningful in this context. The resulting differences in the degrees of freedom between the models are considered here as an inevitable consequence of the particular model structure. In addition, it has to be taken into account that the number of available parameters can be strongly put into perspective by individual parameter sensitivities and by parameter interaction. In other words, model complexity is not a prerequisite

Table 2. LARSIM model parameters used for the Monte-Carlo simulation and their parameter ranges.

Name	Description	Unit	Range
EQD	Calibration factor for storage coefficient of fast runoff Q_{of}	[–]	100–5000
EQD2	Calibration factor for storage coefficient of fast runoff Q_{of2}	[–]	10–1000
BSF	Calibration factor of the "soil moisture" – saturated area function, variable contributing area approach	[–]	0.05–1.0
β	Drainage coefficient for deep storage	[1/d]	0.03–0.05
D_{min}	Minimum lateral drainage from soil storage	[mm/h]	0–5.0
D_{max}	Maximum lateral drainage from soil storage	[mm/h]	0–5.0
$A2$	Repartitioning factor for saturation overland flow and fast subsurface runoff	[mm/h]	0.8–3.0

Table 3. WaSIM-ETH model parameters used for the Monte-Carlo simulation and their parameter ranges.

Name	Description	Unit	Range
kd	Storage coefficient for surface runoff	[h]	0.1–40
ki	Storage coefficient for lateral flow	[h]	0.1–100
dr	Drainage density/anisotropy parameter	[1/m]	0.5–100

for good model performance (see e.g. Gan and Biftu, 2003). Thus, we see no strong reason to assume that a model would have less capabilities of reproducing certain runoff characteristics due to its degrees of freedom, even more as the present study focuses on a very specific aspect of model behaviour.

2.2 Derivation of index measures from the flow duration curve

In order to capture information on different characteristics of model behaviour within a specific domain of flow response we follow an approach which is adapted from the work of Gupta et al. (2008) and Yilmaz et al. (2008): Five index measures are derived based on the evaluation of simulated and observed flow duration curve properties. In contrast to commonly used statistical objective functions (e.g. see Legates and McCabe Jr., 1999) the "Signature Indices" presented by Yilmaz et al. (2008) constitute hydrologically meaningful measures of system response. In this respect, the indices we use differ from the concept proposed by Gupta et al. (2008) insofar as their diagnostic relation to different elements of the model structure as well as to the natural system is less obvious. In order to analyze the reproduction of flood events in detail the indices were conceptualized to focus solely on the characteristics of discharge events with an exceedance probability below a given threshold which is derived from the flow duration curve (Fig. 5). In our study this specific threshold is determined by visual examination of the slope of the observed flow duration curve which, in our example, shows a marked increase at 2%. The remaining section of the flow duration curve is further subdivided at 0.42%, following the same approach (Fig. 5). For each of these subsections individual index measures are calculated according to Eqs. (1) and (2). According to Yilmaz et al. (2008), the percent difference in slope of a flow duration curve segment relative to the observations is given as

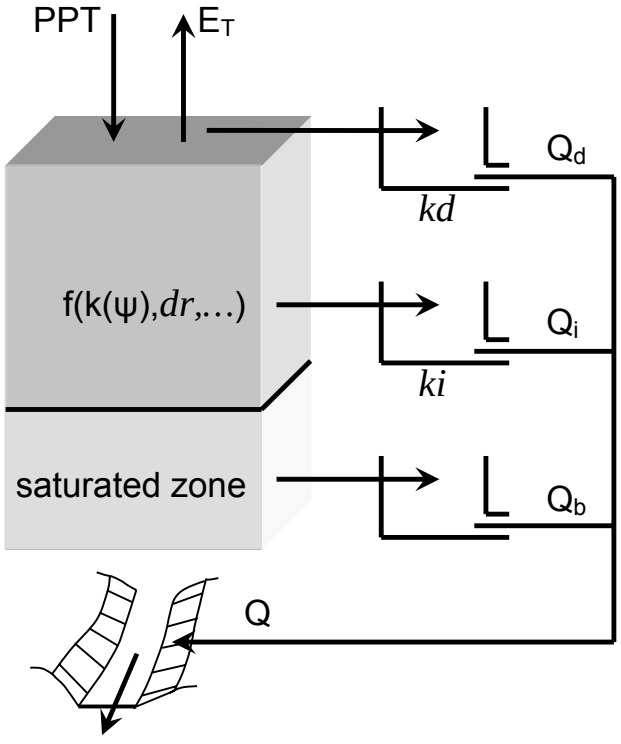

Fig. 4. Simplified schematic representation of the WaSIM-ETH model structure – only those elements are reproduced that are considered in the scope of the present study. Parameters that have been subject to variation in the course of the Monte-Carlo simulation are printed in italic Times New Roman.

$$\%BiasFDC = \frac{\left(\log(Q\mathrm{sim}_i) - \log(Q\mathrm{sim}_j)\right) - \left(\log(Q\mathrm{obs}_i) - \log(Q\mathrm{obs}_j)\right)}{\left(\log(Q\mathrm{obs}_i) - \log(Q\mathrm{obs}_j)\right)} \cdot 100 \quad (1)$$

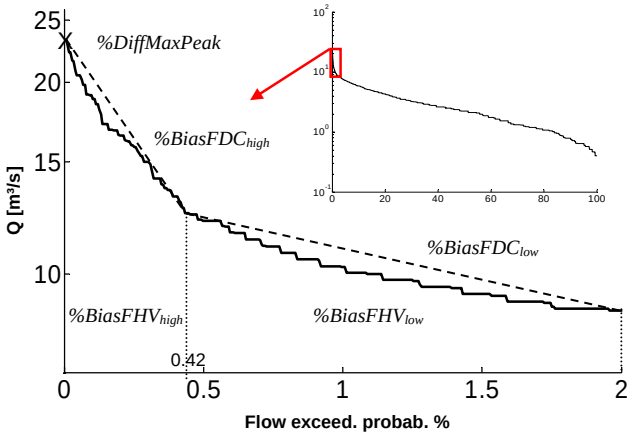

Fig. 5. Derivation of index measures from the upper 2% section of the flow duration curve (FDC).

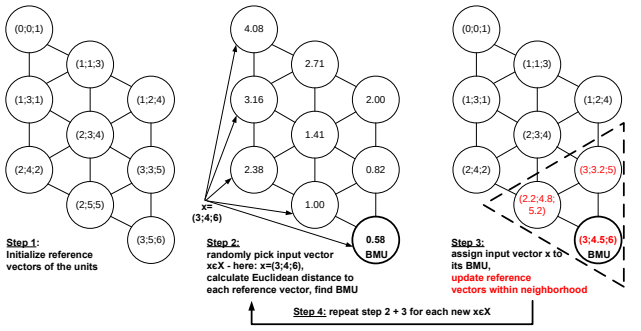

Fig. 6. The iterative training process of SOM (Herbst and Casper, 2008).

where i and j denote the thresholds that define a segment of the flow duration curve; Qsim being the simulated discharges and Qobs being the corresponding observations. Given the observed flow duration curve in Fig. 5 we define the slope of the lower section of the flow duration curve segment *%BiasFDC*$_{\text{low}}$ as *%BiasFDC* with i=2 and j=0.42. Accordingly, the slope of the upper section of the flow duration curve segment *%BiasFDC*$_{\text{high}}$ is defined as *%BiasFDC* with i=0.42 and j=0. Further, the percentage of bias in the flow duration curve high volume segment is calculated based on Yilmaz et al. (2008) as

$$\%BiasFHV = \frac{\sum_{h}(Q\text{sim}_h - Q\text{obs}_h)}{\sum_{h}(Q\text{obs}_h)} \cdot 100 \quad (2)$$

where h denotes the index of all discharge values with exceedance probabilities higher than i and lower than j. Again, we define the bias for the lower flow duration curve segment volume *%BiasFHV*$_{\text{low}}$ as *%BiasFHV* with i=2 and j=0.42. Correspondingly, we define *%BiasFHV*$_{\text{high}}$ as *%BiasFHV* with i=0.42 and j=0. In addition, the percentage of error in maximum peak discharge *%DiffMaxPeak* is determined after Eq. (3).

$$\%DiffMaxPeak = \frac{Q\text{sim}_H - Q\text{obs}_H}{Q\text{obs}_H} \cdot 100 \quad (3)$$

with the index number of the highest element of the flow duration curve being H.

As none of the model parameters that are subject to variation in the Monte-Carlo simulation is related to flood routing or exerts a significant influence on the timing of the discharge peaks we refrained from examining potential time lags between the simulated data and the observations. However, in a more general model evaluation problem, this might be a recommendable procedure.

2.3 Self-organizing maps

SOM is an unsupervised learning neural network algorithm that is applied to high-dimensional data sets in order to categorize the range of data patterns that occur in it and to extract a set of characteristics that describe its multidimensional distribution. A SOM essentially performs a non-linear mapping of vectorial input data items onto a discrete, low-dimensional grid. Most commonly a two-dimensional, rectangular grid with hexagonal topology is used. In contrast to common Vector Quantization methods or k-Means clustering, SOM is topology preserving, i.e. nearby locations on this mapping are attributed to similar data patterns. Likewise, the distance between two nodes on the mapping is proportional to the dissimilarity of the data items they represent. Each input data item $\boldsymbol{x}$ of the training data set X that has to be examined is considered as a vector $x=[x_1, x_2, \ldots, x_n]^T \in \Re^n$, with n being the dimension of the input data space. Let X represent a set of index vectors calculated according to Sect. 2.2, thus n=5. A SOM consists of a fixed number of k neurons that are arranged on a regular grid whose dimensions can be determined by means of heuristic algorithms, if no other preferences are made. Throughout this paper the terms "neuron", "node" and "map unit" are used synonymously. Figure 6 provides a schematic representation of the process of self-organization which, in the following, is explained based on the paper by Herbst et al. (2009).:

Each neuron i is represented through a reference vector

$$m_i = [\mu_{i1}, \mu_{i2}, \ldots, \mu_{in}]^T \in \Re^n \quad (4)$$

whose dimension n equals the number of elements in an input data vector $\boldsymbol{x} \in X$. Typically, the reference vectors $\boldsymbol{m}_i$ are initialized to small random values. However, in order to assure faster and more reliable convergence of the map, we initialize the $\boldsymbol{m}_i$ along the two greatest principal component eigenvectors of the data (Kohonen, 2001). In the classic sequential training the SOM is trained iteratively: In the first

step an input data item $\boldsymbol{x}\in X$ is randomly selected and the Euclidean distance

$$d_i = \sqrt{\sum_{j=1}^{n} \left(x_j - m_{ij}\right)^2} i = 1...k;\ j = 1...n \tag{5}$$

between $\boldsymbol{x}$ and each reference vector $\boldsymbol{m}_i$ is computed (theoretically any appropriate metric can be used as a measure of similarity). The "winning neuron" (also called the best-matching unit BMU of $\boldsymbol{x}$) is the map element c whose reference vector $\boldsymbol{m}_c$ has the smallest distance d_c to $\boldsymbol{x}$ with

$$d_c = \min_i \{\|\boldsymbol{x} - \boldsymbol{m}_i\|\}. \tag{6}$$

In the next step the reference vector $\boldsymbol{m}_i$ and all of its neighbouring neurons are updated according to

$$\boldsymbol{m}_i\,(t+1) = \boldsymbol{m}_i\,(t) + \alpha\,(t)\,h_{ci}\,(t)\,[\boldsymbol{x}\,(t) - \boldsymbol{m}_i\,(t)] \tag{7}$$

where $\boldsymbol{m}_i(t)$ is the current weight vector at iteration step t. Thus, the rate of change for each node of the map is scaled by three factors: a) the difference $(\mathbf{x}(t)-\boldsymbol{m}_i(t))$ between the input data set $\mathbf{x}$ and the prototype vector $\boldsymbol{m}_i$ b) the size of a neighbourhood function h_{ci} which decreases monotonically to zero with t and with distance from the winning neuron and c) a learning rate factor $\alpha(t)$ which gradually lowers the height of the neighbourhood function as the iteration advances. For h_{ci} it is common to use the Gaussian function

$$h_{ci}\,(t) = \exp\left(-\frac{\|\boldsymbol{r}_c - \boldsymbol{r}_i\|^2}{2\sigma^2\,(t)}\right) \tag{8}$$

where $\sigma(t)$ defines the width of the topological neighbourhood, and both $\sigma(t)$ and $\alpha(t)$ decrease monotonically with t. Note that an exact choice of the function $\alpha(t)$ is not required (Kohonen, 2001). Repeated cycling through the training steps causes different nodes and regions of the map to be "tuned" to specific domains of the input space. Importantly, the enforced local interaction between the SOM nodes results in the map gradually developing an ordered and smooth representation of the input data space (Kaski, 1997).

In this work, however, we used Kohonen's "batch-training" algorithm (Vesanto, 2000) to speed up the training process. Here, in each training step the data set is partitioned according to the Voronoi regions of the $\boldsymbol{m}_i$. Instead of sequentially running through all data items in each training cycle the whole data set X is presented to the map as a whole at each training cycle. The reference vectors are updated according to the weighted average of the data samples

$$\boldsymbol{m}_i(t+1) = \frac{\sum_{l=1}^{N} h_{ci}(t)\boldsymbol{x}_l}{\sum_{l=1}^{N} h_{ci}(t)} \tag{9}$$

where c is the index number of the BMU of data set $\boldsymbol{x}_l$, and N is the number of data samples. This variant of the training does not make use of the learning rate factor $\alpha(t)$.

In the course of the training the reference vectors are "tuned" to the different patterns contained in the input data. The final reference vectors form a discrete approximation of the input data distribution. Thus, patterns that occur more frequently in the input space are mapped onto a larger area. Note that, as the number of neurons – and consequently the number of reference vectors – is much smaller than the number of data items used for the training, SOM can also be seen as a data compression method.

In our study we also make use of the fact that, once its training is finished, the SOM can be applied to project an input data vector $\boldsymbol{y}$ onto the map which has not been part of the training data set. This means that according to Eq. (6) the neuron $c(\boldsymbol{y})$ with reference vector $\boldsymbol{m}_{c(y)}$ is determined for which

$$\left\|\boldsymbol{y} - \boldsymbol{m}_{c(y)}\right\| = \min_i \{\|\boldsymbol{y} - \boldsymbol{m}_i\|\}. \tag{10}$$

Neuron $c(\boldsymbol{y})$ then represents the domain of input data patterns from X that is most similar to $\boldsymbol{y}$. It follows that the set of data items $\hat{X}\subset X$ which is attributed to $c(\boldsymbol{y})$ represents those training data items that are most similar to $\boldsymbol{y}$ with respect to the criterion given by Eqs. (5) and (10). The neuron c($\boldsymbol{y}$) is called the "best-matching unit" (BMU) of $\boldsymbol{y}$.

2.4 Data preparation and training of the SOM

For each of the 4000 time series obtained by running a Monte-Carlo simulation (Sect. 2.1) a set of five index measures was calculated according to Eqs. (1–3) (Sect. 2.2). The procedure was carried out for each of the three models.

Prior to the SOM training, each index was normalized to a value having zero mean and variance of one using a linear transformation such that high index values do not exert a disproportionate influence on the training. The side lengths of the map as well as the initial reference vectors were determined by means of a heuristic algorithm involving the calculation of the two biggest eigenvalues of the covariance matrix of the data (Vesanto et al., 2000). For more details on the data preparation and the training please see Herbst et al. (2009).

At the first stage of our study the three data sets were treated individually. Subsequently, the data preparation and the training were repeated with the combined data set of all three models.

For the SOM training as well as for a part of the evaluation procedures the "SOM Toolbox for MATLAB™" (Helsinki University of Technology, http://www.cis.hut.fi/projects/somtoolbox/) was used.

2.5 Evaluation of SOM results

Generally, the number of neurons on the maps is much smaller than the number of data sets used for the training. As

a consequence of this, every neuron represents a set of simulation runs and their respective index value pattern. In the following, we evaluated the index properties of the individual nodes by de-normalizing the reference vectors of the maps. Each of the nodes/reference vectors represent the mean index value properties of a small sub-set of model data used for the training. In the following, these reference vector index values are visualized by means of a small, coloured bar plot for each node. In a bar plot visualization the position of the BMU can easily be identified by the map unit with the "flattest" bars. Note that the height of the individual bars is scaled relative to the range of the corresponding index. Also colour coding of the index values is used, whereas the same map grid is reproduced five times with a colouring corresponding to the distribution of the individual index values (so-called component planes).

As a result of the self-organizing process that takes place in the course of the training, the data items which are grouped to such a sub-set have similar properties with regard to their five index values. Due to the topological properties of the mapping the distance between two nodes on the map is roughly a function of the dissimilarities between the data sets attributed to these nodes. Please note that, to some extent, the SOM embodies statistical properties, e.g. the number of reference vectors on a map that display a certain type of quality is proportional to the number of data sets with that property. As a simple measure of the quality of the mapping the "quantization error" (Kohonen, 2001) $\bar{d}$ is calculated using Eq. (10).

$$\bar{d} = \frac{1}{N} \sum_{p=1}^{N} \left\| \boldsymbol{x}_p - \boldsymbol{m}_{\boldsymbol{c}(p)} \right\| \tag{11}$$

It represents the average distance of each data vector $\boldsymbol{x}_p$ of the N input data items contained in the training data X to its associated BMU reference vector $\boldsymbol{m}_{c(p)}$, with p being the index of the data items (not to be confused with the index values of Sect. 2.2!).

We further take advantage of the possibility to label the input data items that are attributed to each neuron via the training. That way, each input data item is linked to a model parameter set and its original simulated time series. Thus, the neurons of the map can be evaluated with respect to the model parameters, e.g. by calculating the mean values of each parameter for the individual map units. The distribution of parameter values over the map is again visualized by means of colour coding. In doing so, the same map grid is reproduced for each parameter, however, with different colouring according to the distribution of parameter values. In the following, this type of visualization is referred to as parameter plane. Corresponding patterns on a component plane and parameter plane indicate that an index value is governed to a large extent by a particular parameter. Moreover, an irregular pattern on the parameter plane is indicative of parameter insensitivity, according to the components which were used to train the map.

For each map grid (i.e. for each model) the BMU of the measured discharge time series is determined according to Eq. (10). Following Sect. 2.2 (Eqs. 1–3) the time series of observed discharges Qobs maps as $\boldsymbol{y}$=[0 0 0 0 0]^T into the index space. We then calculate the quantization error of the BMU

$$\bar{d}_{\mathrm{BMU}} = \frac{1}{n} \sum_{r=1}^{\hat{N}} \left\| \hat{\boldsymbol{x}}_r - \boldsymbol{m}_{\boldsymbol{c}(y)} \right\| \tag{12}$$

in order to obtain a rough indicator of how close the data items $\hat{\boldsymbol{x}}_r \in \hat{X}$ with $r = 1 \ldots \hat{N}$ which are attributed to the BMU $c(\boldsymbol{y})$ of the observation (Eq. 10) approximate the observation (represented by $\boldsymbol{y}$=[0 0 0 0 0]^T). In Eq. 11 $\hat{N}$ denotes the number of data items in $\hat{X}$. Note, that it is possible to identify a BMU for any data set that has the same dimensionality as the input data, irrespective of its distance from the observations.

The data items on the BMU $c(\boldsymbol{y})$ of a map also correspond to one or more model parameter sets which are subsequently used to simulate an extreme flood event from August 2002 that has not been part of the Monte-Carlo data set. As a reference, we visually compare these simulations to the results we obtained by using the shuffled complex evolution optimization algorithm (SCE-UA, Duan et al., 1992) to find a parameter set that minimizes the root of the mean squared error (RMSE) for the same period of time for which a simple, model specific, weighting scheme after Casper et al. (2009) is used. This scheme basically applies a higher weight to all time steps with a discharge higher than three times the mean discharge (Q>3MQ).

3 Results

3.1 SOMs generated from the individual model data sets

Although from each of the models the same amount of data items was processed, the maps for NASIM, LARSIM and WaSIM-ETH slightly differ in number of neurons and side lengths which is caused by the initialization of the maps according to Sect. 2.4. The overall quantization error $\bar{d}$ (Eq. 11) for the three mappings ranges between 0.236 (NASIM) and 0.278 (WaSIM-ETH). Thus, it can be assumed that the SOM provides a good fit of the model data. The distributions of reference vector properties over the map support that the model data has been arranged by similarity over the maps. No void, i.e. interpolative, units are present. In the following, an account of the results for the individual models is given. Note that, as to the simulated time series illustrations, only a representative period of the entire simulated time series is reproduced in the following in order to assure better readability of the figures. Representations of the flow duration curve

Fig. 7. NASIM: Distribution of index properties on the map displayed as bar plots and the position of the best-matching unit (BMU). Note that the bars have individual relative scales.

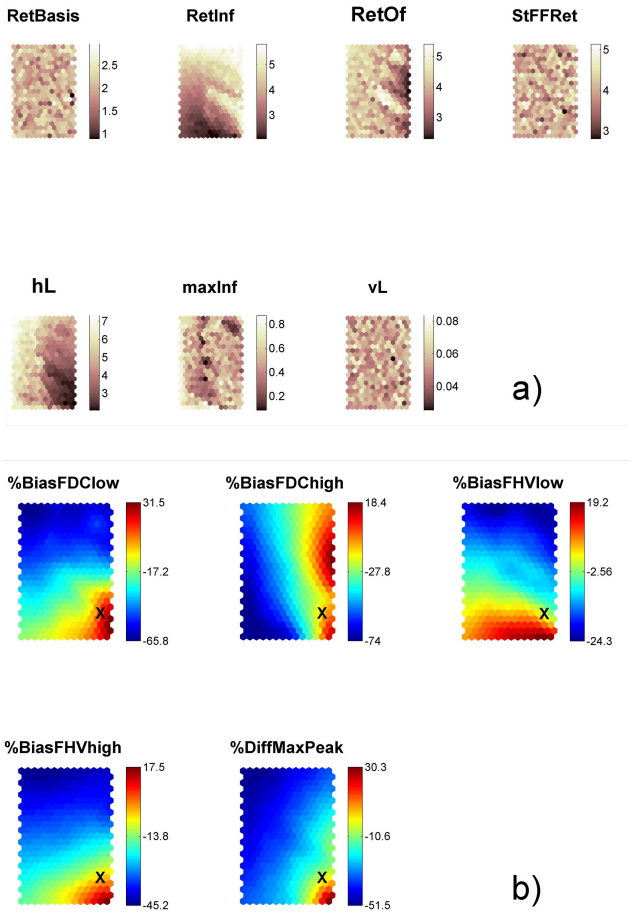

Fig. 8. NASIM: **(a)** Parameter plane, i.e. mean values of each model parameter for the simulations projected onto the individual map elements. **(b)** Distribution of reference vector (i.e. indices) properties on the map. The position of the BMU is marked.

(FDC), however, always refer to the full length of simulated discharge time series.

3.1.1 NASIM

The bar plot (Fig. 7) reveals that a significant proportion of simulation runs which were attributed to the upper half of the map underestimated all five properties that are represented by the indices. The increase of *%BiasFDC*$_{\text{high}}$ from the left to the right hand side implies a general increase in peak runoff reaction in this direction of the map. The remaining index components generally tend to smaller negative or positive values towards the lower part and the right hand side of the map. From the values of *%DiffMaxPeak* in Fig. 7 it immediately becomes obvious that only very few simulations exceeded the measured peak discharge.

A comparison of the parameter plane Fig. 8a with the component plane Fig. 8b indicates that the increase in peak runoff reaction *%BiasFDC*$_{\text{high}}$ is largely influenced by parameter *hL* in combination with the parameter for the retention of "overland flow", *RetOF*. Figure 8a also reflects that parameter *RetOf* remains insensitive with respect to the indices as long as *hL* has high values. According to Sect. 2.1, this behaviour is evident because the generation of "overland flow" is overridden by *hL*. Moreover, *RetInf*, which governs the retention of water allocated to interflow, with high probability exerts an influence on the increase in *%BiasFHV*$_{low}$, which consequently provides a rather simple explanation for the position of the BMU on the map. Further, it can be seen from Fig. 8a that parameter *maxInf* is – at best – only partially sensitive. *vL*, *RetBasis* and *StFFRet* are insensitive here because they are linked to the generation of "base flow" and runoff from urban or impervious areas which for the "Schwarze Pockau" catchment only comprise 6.6% of the total area.

Correspondences in the distribution patterns in Fig. 8b reveal significant correlations between the indices with respect to the behaviour of NASIM. However, each component plane is scaled separately. Therefore, the individual optima of the

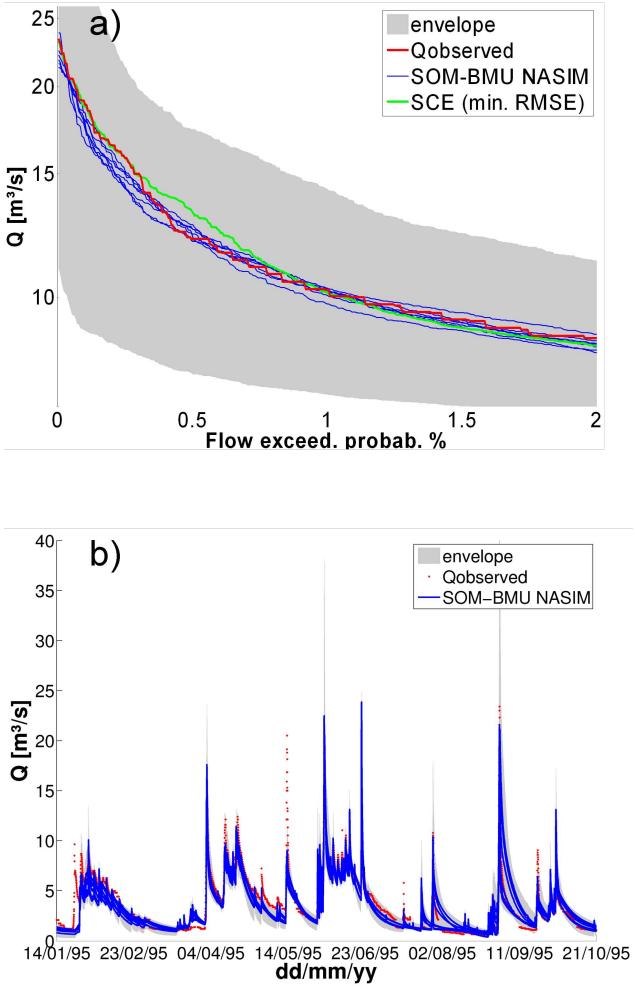

Fig. 9. NASIM: **(a)** Flow duration curves (upper 2% section): Simulations corresponding to the BMU compared to the observed discharge and the results of an optimization approach using SCE-UA. **(b)** Time series corresponding to the BMU compared to the observed discharge.

indices still do not coincide on the same map location. The scales of the component planes additionally give insight into the true lengths of the bar plots in Fig. 7.

The quantization error for the BMU of the observed discharge time series, calculated according to Eq. (11), yields $\bar{d}_{\mathrm{BMU}}$=0.85, which suggests that the observations are located somewhat off the model data obtained from the Monte-Carlo Simulation. Nevertheless, from the simulation results in Fig. 9a it can be seen that the parameter sets that were retrieved from the BMU associated to the observed runoff reproduce the characteristics of the measured FDC with very high accuracy. However, this requirement is also satisfied surprisingly well by the reference simulation which was obtained using the SCE-UA algorithm in combination with a simple weighting scheme. Notwithstanding, the time series of simulated discharge (Fig. 9b) shows that, with the exception of only a few events, the runoff peaks could be reproduced well. As the FDCs grow almost congruent towards the ordinate it does not surprise that a part of the simulation runs attributed to the BMU performs equally well during the validation event compared to the SCE-UA reference simulation, yet none of the model realizations is able to reach the peak flow (Fig. 10).

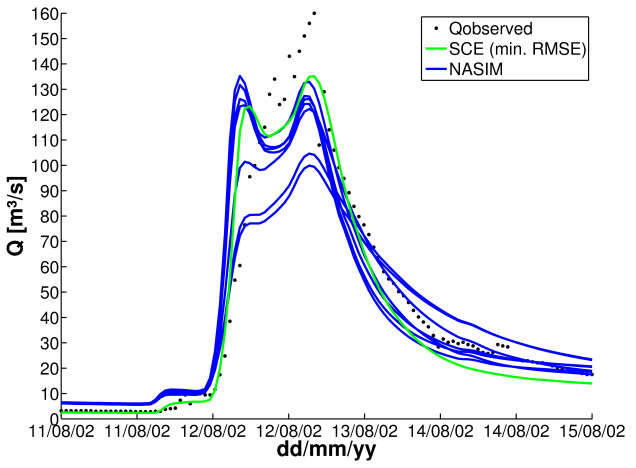

Fig. 10. NASIM: Results for the validation event August 2002 (BMU realizations and SCE-UA, RMSE).

3.1.2 LARSIM

Figure 11 shows that LARSIM, contrary to NASIM, tends to overestimate almost the entire set of indices whereas for at least the lower third of the map this tendency is very pronounced. Lower or negative index values as well as sporadic underestimation of runoff reaction ($\%BiasFDC_{\mathrm{high}}$) and –volume ($\%BiasFHV_{\mathrm{low}}$), are largely recorded in the upper regions. The only index, however, which is constantly overestimated throughout the data set is the runoff reaction during all time steps corresponding to the lower section of the FDC, $\%BiasFDC_{\mathrm{low}}$. The maximum peak discharge, expressed by *%DiffMaxPeak*, is also overestimated throughout large portions of the model data.

Towards the right hand corners a marked increase in $\%BiasFDC_{\mathrm{high}}$ is superimposed to the comparatively monotonous pattern of index combinations. From Fig. 12a it can be seen that parameter *EQD2* decreases in the same direction, which most likely reveals a main control for the runoff reaction in the fastest portion of flow. Likewise, the error in maximum peak flow, which is expressed by *%DiffPeakMax*, grows strongly positive towards the lower right hand corner. The volume allocated to flow corresponding to the lower branch of the FDC ($\%BiasFHV_{\mathrm{low}}$), and partially also $\%BiasFHV_{\mathrm{high}}$, increases towards the lower left corner of the map (Fig. 12b) which, according to Fig. 12a, indicates that these features are likely to be governed by parameter

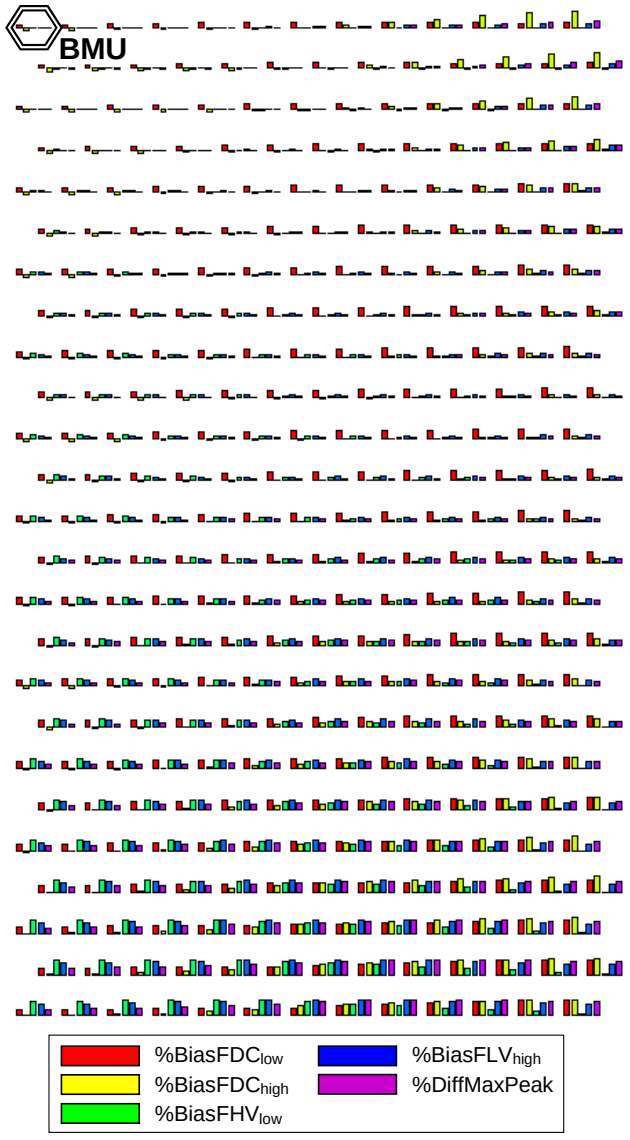

Fig. 11. LARSIM: Distribution of index properties on the map displayed as bar plots and the position of the best-matching unit (BMU). Note that the bars have individual relative scales.

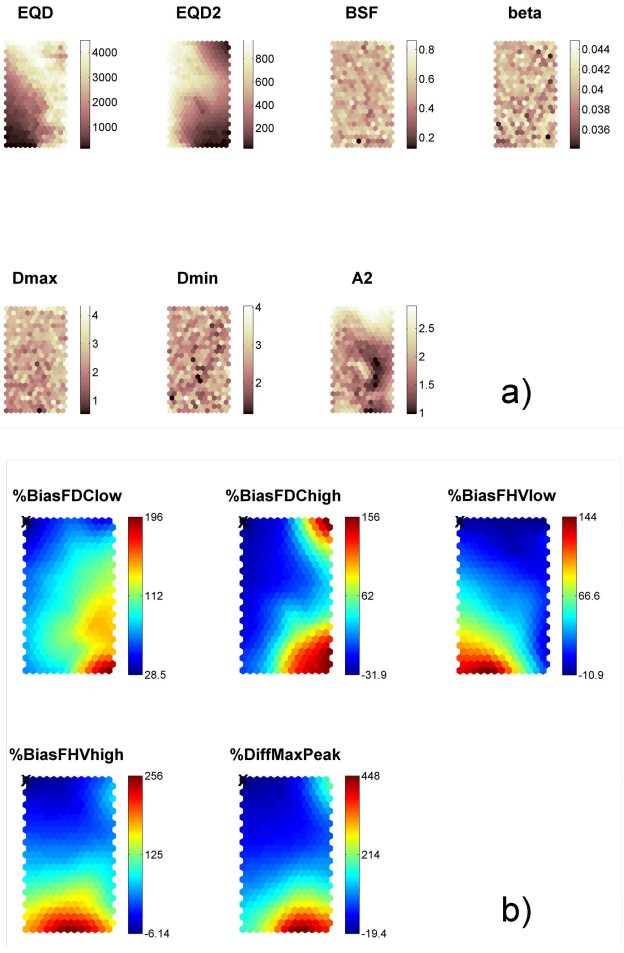

Fig. 12. LARSIM: **(a)** Parameter plane, i.e. mean values of each model parameter for the simulations projected onto the individual map elements. **(b)** Distribution of reference vector (i.e. indices) properties on the map. The position of the BMU is marked (top left corner of the map.

EQD. As to the remaining parameters, no further correlations with indices are clearly apparent, although at least parameter *A*2 shows a marked sensitivity with respect to data sets allocated to the right hand side of the mapping which points to some degree to an interaction between *EQD* and/or *EQD2* as well as *A*2.

The influence of *EQD2* on *%BiasFDC*$_{high}$ can easily be explained following Sect. 2.1: As *EQD2* is a scaling factor for the retention of the "fast" saturated flow component $Q_{of2,}$ it largely controls the volume per time step which is allocated to peak discharges. Its sensitivity, however, depends on the threshold parameter *A*2 which at the same time governs the function of storage coefficient *EQD*. As documented by the corresponding scales of the component planes in Fig. 12b, the sensitivity of *%DiffMaxPeak* and *%BiasFDC*$_{high}$ is extremely high. But also the volumes *%BiasFHV*$_{low}$ and *%BiasFHV*$_{high}$ seem to react with sharp gradients on changes of *EQD* and *EQD2*. Following Sect. 2.1 it does not surprise that the "base flow" storage coefficient β as well as D_{min} or even D_{max} do not show any apparent sensitivity according to Fig. 12a. However, the lack of sensitivity with respect to parameter *BSF*, which controls the generation of saturated flow via a variable contributing area approach, is unexpected and does not lend itself to a straightforward explanation.

The impression that LARSIM does not seem to be capable of simultaneously meeting the constraints imposed by the five indices is already conveyed by Fig. 11. The comparatively high quantization error for the BMU of the observed discharge time series $\bar{d}_{BMU}$=1.1 and the fact that this BMU is located on an extremely marginal position in the upper left

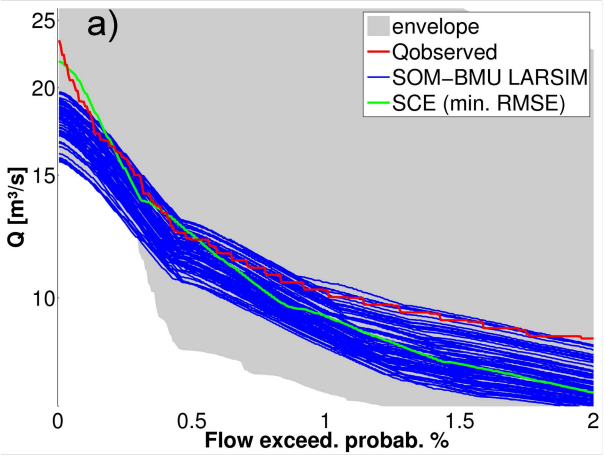

Fig. 13. LARSIM: **(a)** Flow duration curves (upper 2% section): Simulations corresponding to the BMU compared to the observed discharge and the results of an optimization approach using SCE-UA. **(b)** Time series corresponding to the BMU compared to the observed discharge.

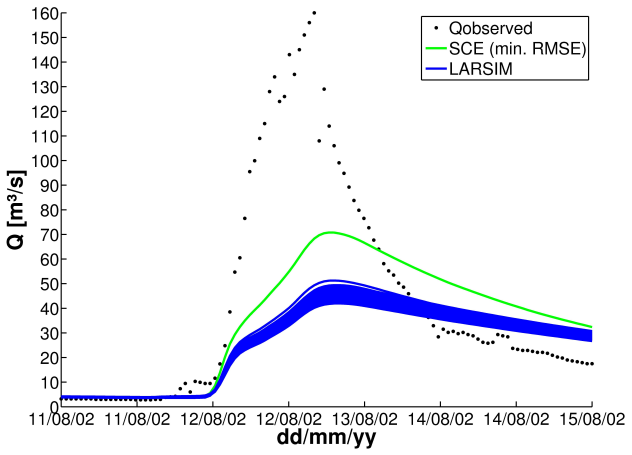

Fig. 14. LARSIM: Results for the validation event August 2002 (BMU realizations and SCE-UA, RMSE).

hand corner of the map further corroborates this finding. The resulting model behaviour is illustrated in Fig. 13a and b. The envelope of the simulations in Fig. 13a shows that, in accordance with Fig. 11, a significant proportion of simulation runs (i.e. parameter combinations) is potentially capable of reaching sufficiently high and even excessive peak discharge values. However, the constraints linked to the lower (0.42–2%) part of the FDC counteract this behaviour with very high probability and force the position of the BMU (Fig. 11) towards the upper left hand side of the map. This finding is further supported by the results obtained by using the SCE-UA optimization algorithm (Duan et al., 1992) to minimize the RMSE of the simulated time series. In addition, the sharp gradients and extreme overestimation of *%DiffMaxPeak* and $\%BiasFDC_{\text{high}}$ in reaction to changes of *EQD* and *EQD2* as well as the position of the simulation envelope in relation to the observations (Fig. 13a) point at that the ranges allowed to these parameters in the course of the Monte-Carlo simulation are disproportionate. Consequently, the BMU parameter sets comprise high values for both *EQD* and *EQD2*, apparently in order to compensate for an excess in fast runoff components. This excess could have been triggered by of the settings for parameter *A*2 and/or overly high values for parameter *BSF*. The model realizations selected by using the BMU criterion display some kind of "plateau behaviour" which is illustrated by the decrease in slope towards the upper end of the FDC (Fig. 13a) and indicates deficits in discharge volume generation for runoff peaks. These deficits also manifest themselves with regard to the time series results for the training period (Fig. 13b) and, even more, for the extreme event of August 2002 (Fig. 14). Here, the model realizations that were attributed to the BMU perform even worse than the reference simulation which was obtained using the SCE-UA algorithm and hardly reach about 50 m^3/s peak flow compared to approximately 160 m^3/s of observed peak discharge.

3.1.3 WaSIM-ETH

Regarding the WaSIM-ETH model realizations, the indices used to examine the model behaviour show a marked correlation and a general gradient extending from the upper left to the lower right corner (Fig. 15). Nevertheless, the index ranges covered by the Monte-Carlo simulation with WaSIM-ETH are quite individual and comprise negative as well as positive values. Thus, the upper third of the map is generally characterized by underestimation of the indices which gradually fades to high index values towards the lower right hand side of the map such that the lowest index values (i.e. the model realizations that best "fit" the observations), and thus the BMU, can be found only a few nodes below the centre of the map.

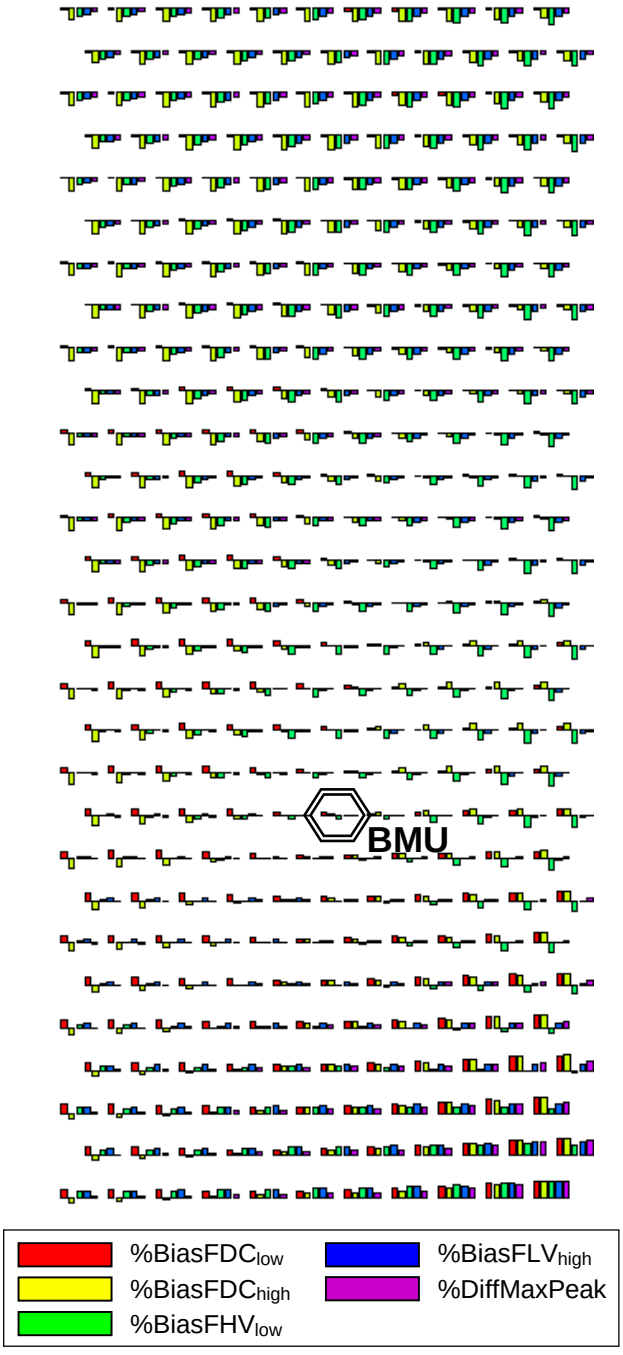

Fig. 15. WaSIM: Distribution of index properties on the map displayed as bar plots and the position of the best-matching unit (BMU). Note that the bars have individual relative scales.

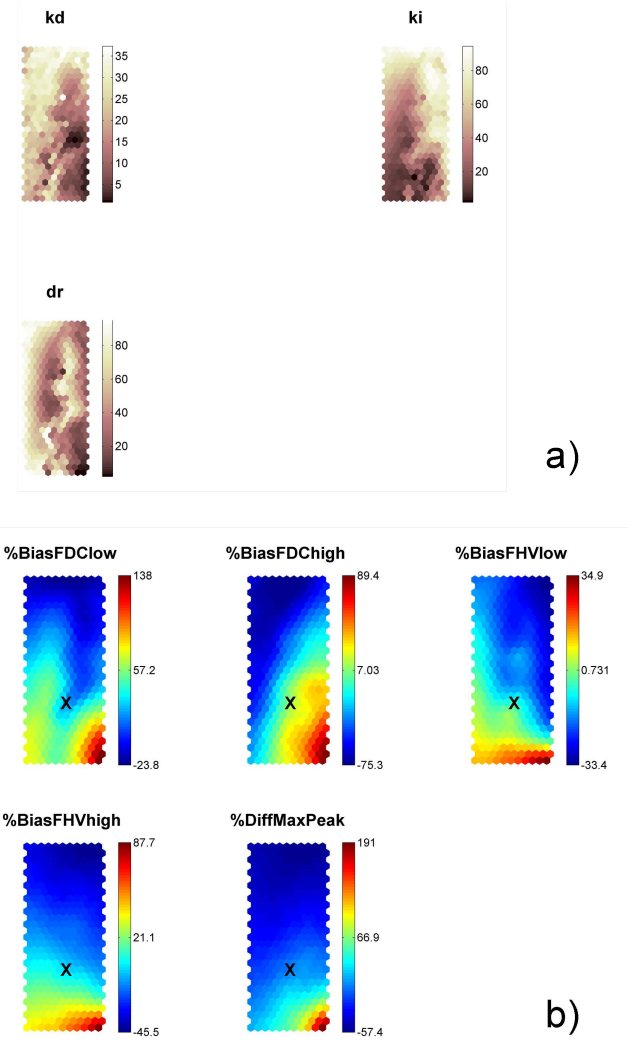

Fig. 16. WaSIM: **(a)** Parameter plane, i.e. mean values of each model parameter for the simulations projected onto the individual map elements. **(b)** Distribution of reference vector (i.e. indices) properties on the map. The position of the BMU is marked.

Contrary to the other models we examined, from a visual comparison of Fig. 16a and the component planes (Fig. 16b) it is not possible to isolate any straightforward relationship between individual model parameters and indices, although each of the parameters we included in the Monte-Carlo simulation shows a marked sensitivity with respect to the index measures. However, to some extent, the parameter planes themselves (Fig. 16a) display correlated patterns, e.g. the map units with high values for *ki* suspiciously coincide with the map units for which parameter *kd* acquires predominantly low values. As to the general behaviour of the WaSIM-ETH model structure with respect to the indices, it can be stated that low values for parameter *kd* and *dr* in combination with intermediate *ki* values redound to an increase of peak flow.

From these findings we infer that all three model parameters are equally important for matching the five index measures and that parts of the model structure exhibit a strongly interacting, maybe even equifinal behaviour. That way, the effect of one parameter can be compensated to some extent by a combination of the other parameters. This would also provide some explanation for the fact that the BMU is located

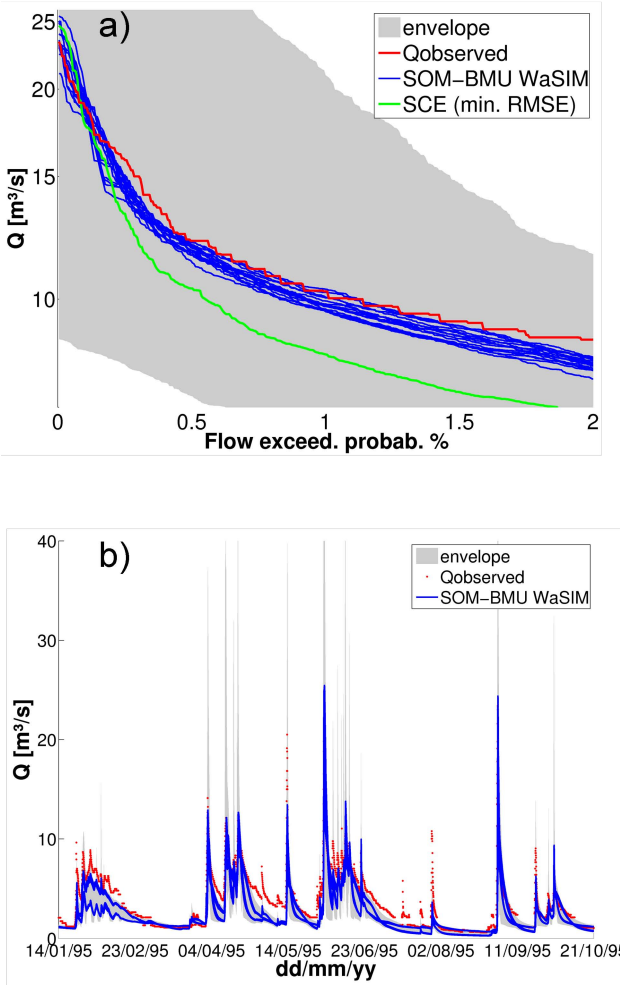

Fig. 17. WaSIM: **(a)** Flow duration curves (upper 2% section): Simulations corresponding to the BMU compared to the observed discharge and the results of an optimization approach using SCE-UA. **(b)** Time series corresponding to the BMU compared to the observed discharge.

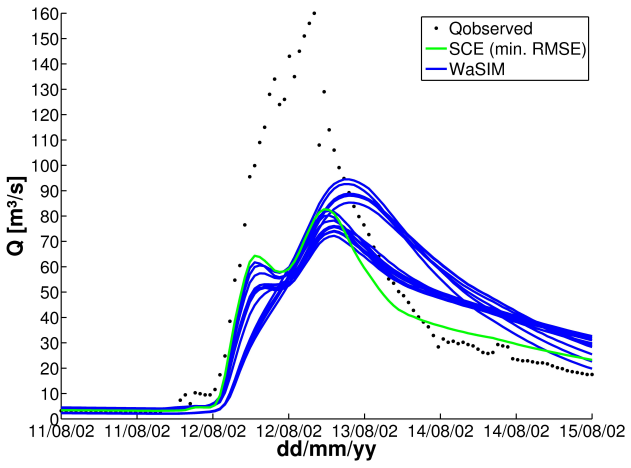

Fig. 18. WaSIM: Results for the validation event August 2002 (BMU realizations and SCE-UA, RMSE).

in a map region where all three parameters are subject to considerable alterations. It finally results that the parameter sets associated to the BMU can be divided into two groups with strongly contrasting values: The first group has low values with respect to *kd* and simultaneously high values for *dr* while the second group displays high *kd* values in combination with low values for *dr*. The respective index measures, however, turned out to be quite similar.

With $\bar{d}_{\mathrm{BMU}}$=0.76 the quantization error of the BMU model realizations that correspond to the time series of measured discharges is the smallest among the three models and indicates that these parameter sets match the observed behaviour of the system relatively well. This is further supported by the rather central position of the BMU on the map. The model realizations which are retrieved from the BMU thus result in a rather accurate representation of the FDC characteristics, especially regarding the highest runoff values (Fig. 17a) whereas the FDC of the SCE-UA reference simulation adopts a steeper trajectory and hits the ordinate somewhat above the FDC of the observations. The model time series (Fig. 17b) finally exemplify that the BMU model realizations were obtained using parameter sets from disjoint regions of the parameter space. The peaks, however, are reproduced very well, with the exception of only a few minor runoff events. In contrast, the peak of the validation event (Fig. 18) is strongly underestimated by all model realizations, whereas only one half of the BMU model realizations exceed the peak flow of the SCE-UA reference simulation.

3.2 Results generated from SOM of the joint model data set

According to the initialization procedure (Sect. 2.4) the number of neurons on a map does not increase linearly with the number of data items used for the training. Therefore, the proportion of data items to the number of map units results somewhat higher for the SOM trained on the joint data set than for the SOMs we discussed in the previous section. The overall quantization error $\bar{d}$=0.38 (Eq. 10) of this SOM result appears to be sufficiently low so as to characterize the mapping as a very good approximation of the model data.

In Fig. 19a the neurons of the map are reproduced as pie charts in order to illustrate the distribution of data items from the different models on the SOM. These represent the percentage of data from each model that has been attributed to the neurons via the training. In Fig. 19b the same pie charts are scaled using the number of data items on each map unit in order to simultaneously visualize the distribution of data quality and quantity on the map. The void regions on the map indicate interpolative units where the data items are clearly disjoint and characterized by marked differences with regard

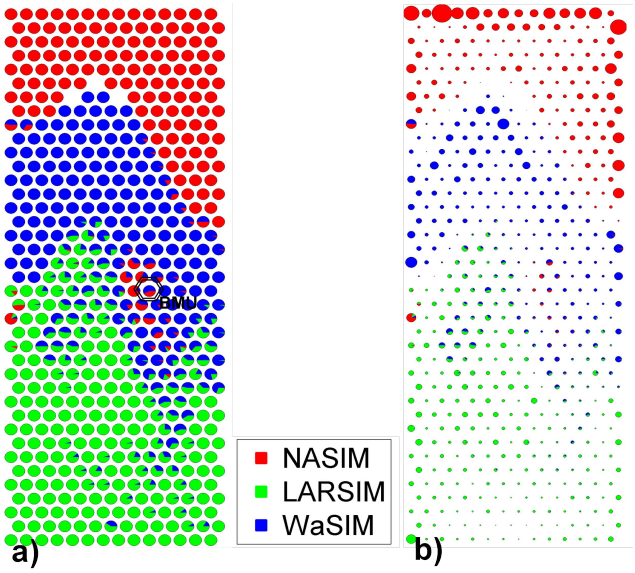

Fig. 19. Comaparison of NASIM, LARSIM and WaSIM: **(a)** The neurons of the SOM are reproduced as pie charts that represent the percentage of data from each model that has been attributed to the neurons via the training. **(b)** same as (a) but displayed as a "hit histogram", i.e. the size of the pie charts is proportional to the number of data items attributed to the corresponding neuron.

to their indices. With the exception of some very isolated occurrences there are no nodes on the map that are simultaneously populated with model realizations from all three models. The same holds true for simultaneous occurrences of NASIM and LARSIM realizations on a node. Close to the center of the map the nodes are predominantly populated with mixtures of model realizations from LARSIM and WaSIM-ETH as well as WaSIM-ETH and NASIM. Following the theory of Self-Organizing Maps (Sects. 2.3 and 2.5), it can be assumed that these model realizations display equivalent characteristics with respect to the indices we used to describe them. The distances between the nodes further allow inferring that the differences between NASIM and LARSIM in terms of index characteristics are stronger than the differences between realizations of NASIM and WaSIM-ETH or LARSIM and WaSIM-ETH, which is most importantly highlighted by the fact that hardly any node is populated at the same time with realizations from NASIM and LARSIM. The neurons close to the bottom of the map, on which simultaneous occurrences of LARSIM and WaSIM-ETH can be found, point at sporadic extremes of WaSIM-ETH model behaviour. Figure 19b further exemplifies that similarities between model realizations from the three models only occur with rather low probability. Moreover, it can be seen that the data items are distributed with a higher density around the upper margin of the map. There are two potential possibilities that lend themselves to explain this phenomenon: When the multidimensional data distribution is rapidly thinning to-

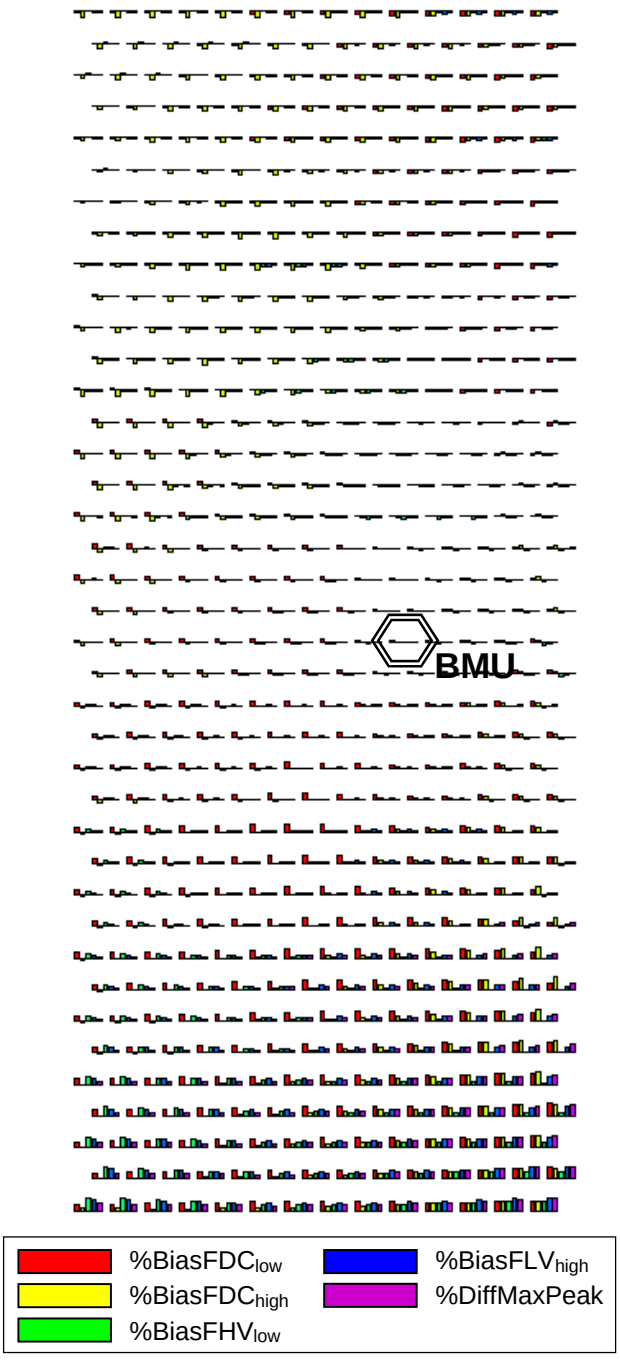

Fig. 20. SOM of the combined data from all models in bar plot illustration.

wards its margins it can not always be covered entirely by the SOM reference vectors. Consequently, the remaining data sets are attributed to the nearest, marginal nodes. Otherwise, this type of distribution could indicate that the corresponding model realizations are indeed very similar compared to the remaining data set and thus attributed to the same neuron.

The aforementioned individuality of model realizations is further demonstrated by the bar plot of the map (Fig. 20)

which compares the index characteristics that distinguishes the models. The mapping shows a very distinct organization of data properties: Predominantly underestimated index values on the upper part gradually fade to overestimated indices in the lower part. The model characteristics that are captured in this representation of the map correspond exactly with Figs. 7, 11 and 15. The position of the BMU which represents the most likely location of the measured time series on the map, is already roughly identifiable from the bar plots in Fig. 20 and coincides, according to Fig. 19a, with the map region in which the neurons are simultaneously populated with data from the models NASIM and WaSIM-ETH. Thus, these models can be characterized as equivalent, of course, only according to the criteria which have been imposed in our study to discriminate between individual model realizations.

The model realizations allocated to the BMU partly correspond with the BMU realizations which have been identified using the mappings of the individual data sets. However, as a consequence of the map dimensions, a somewhat higher number of model realizations are attributed to this BMU. Nevertheless, it can be seen from the FDC plot in Fig. 21a and from Fig. 21b and c that these model realizations still represent the characteristics of the observed discharge time series very well. A distinctive feature of the model realizations obtained from NASIM (Fig. 21b) is the partial overestimation of peak discharge which is why a better performance of these realizations with respect to the extreme flood event from August 2008 can be expected, compared to the results from Sect. 3.1. This effect, however, can not be influenced deliberately and must be attributed to the wider range of data items on the BMU.

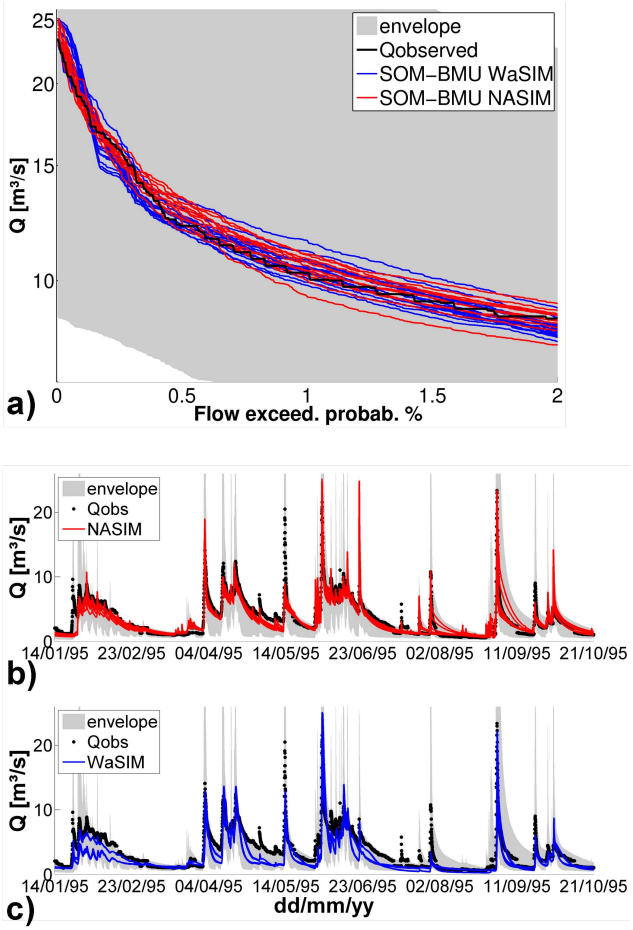

Fig. 21. (a) Flow duration curves (upper 2% section): Simulations corresponding to the BMU of the SOM trained on the combined model data compared to the observed discharge and the results of an optimization approach using SCE-UA. **(b)** Time series of the model NASIM corresponding to the BMU of the SOM trained on the combined model data compared to the observed discharge. **(c)** Time series of the model WaSiM-ETH corresponding to the BMU of the SOM trained on the combined model data compared to the observed discharge.

4 Discussion and conclusion

Similar to Herbst et al. (2009) our study is based on a combined approach: While the indices adopt the function of performance measures (rather than Signature Measures, according to their underlying theory, see Gupta et al., 2008) the Self-Organizing Map serves as a tool to analyze and visualize the data which is obtained through them.

The indices we used were conceptualized to extract data on very specific characteristics of model behaviour according to the focus of our study. These characteristics are represented by a choice of FDC-based indices that are intended to focus on the reproduction of peak discharges. They consequently have to be understood as an example of a model evaluation problem. Of course, the choice of indices can be tailored, according to the individual goals of the model analysis. This also includes a weighting of individual indices or measures in order to put more emphasis on specific time series features. However, we did not make use of this option and preferred to weight all indices with 1 instead. Thus, our study shares the underlying assumptions of the work by Herbst et al. (2009), namely that the index or performance measures are equally relevant and that the model is capable of reproducing them.

The SOM helps to analyze the data obtained via the indices by producing a discretized (and thus data-compressed) mapping of their distribution in the index space onto a two dimensional plane such that their pattern and consequently the patterns of model behaviour can be conveyed in a comprehensive manner. This is achieved by different visualization techniques (see also Vesanto, 1999) and importantly by linking the model properties to the corresponding parameter space. In a sense, the SOM helps to turn the data extracted via the indices into information on model behaviour which subsequently can be used in the decision making process.

The results from Sects. 3.1 and 3.2 clearly demonstrate that a SOM can be used to cluster model output data according to different (time series) characteristics. Although the indices used in our study are not fully independent (finally they are all derived from the FDC) they effectively helped differentiating the simulation results obtained from the watershed models NASIM, LARSIM and WaSIM-ETH. It has been demonstrated that the clustering of model output data provides useful insights, such as a preliminary sensitivity analysis and a general characterization of model behaviour regarding the reproduction of peak discharges. In addition, the presented approach allows identifying the model parameters and -time series that best approximate the observations with respect to a given set of constraints embodied by the indices. This is achieved by determining the BMU of the index vector that corresponds to the observations, which involves identifying the reference vector that minimizes the Euclidean distance to the observations vector. This procedure, commonly used in many SOM applications, deserves critical attention as it implies converting a multi-objective optimization to a single-objective problem (e.g. Madsen 2003) which does not always permit to find the optimal solution of a multi-objective problem (Zadeh, 1963; see also Gupta et al., 1998).

Notwithstanding, the results obtained for the BMU of the NASIM map and the WaSIM-ETH map represent the characteristics of the observed time series with similar or partially superior accuracy compared to the results we obtained by implementing a simple calibration strategy by means of the optimization algorithm SCE-UA. This finding, on one hand, is partly owed to the fact that the SCE-UA optimization algorithm, in contrast to the SOM approach, allows to globally searching the parameter space with potentially infinite resolution. On the other hand, the poor results with respect to the BMU of the LARSIM map are attributable to the constraints that were imposed by applying the five index measures. These constraints turned out to be overly rigorous to be simultaneously satisfied by the LARSIM model and are probably incompatible with its general behaviour. Consequently, the given set of indices avoided that its potential could be exploited to the full extent. At this point it has to be stressed that an accurate reproduction of runoff events in the first place depends on the quality of the precipitation input data. However, as to this aspect, the results obtained with the three models give no reason for concern. Besides the information on model behaviour that can be extracted using the SOM approach, one of its strengths has to be seen especially in the ability to extract a set of model parameters that meet a set of very specific criteria (which e.g. could have been imposed by decision makers). The corresponding parameter ranges, in turn, constitute a potential key for the assessment of parameter uncertainties.

Regarding the simulation of the extreme flood event (11–15 August 2008) our approach did not yield clear improvements compared to the much simpler SCE-UA calibration strategy. Thus, it has to be put into question whether the predictive abilities of hydrological models can be enhanced using this approach. In contrast, the results rather indicate that the ability of a model to "extrapolate" to behavioural domains beyond the calibration data can be exploited with a higher probability if the model realizations match or overestimate the highest section of the FDC. Another, rather self-evident conclusion from the LARSIM result is that a successful calibration strategy always has to consider the peculiarities of model behaviour. However, this behaviour is also largely determined by the parameters which, in the scope of our study, are considered as constant, among other important influences such as the input data. Thus, the parameterization used for the LARSIM model deserves further critical analysis in order to elucidate its behaviour.

The most prominent advantage of the SOM in the context of model analysis is that it allows to simultaneously evaluating the data from two or more models. Using a SOM in combination with an appropriate set of "measures" that help extracting specific information from time series, model realizations that satisfy a given set of criteria can be selected from among various model structures at a time. As only similar data items are attributed to the same map unit (see Sects. 2.3 and 2.5) the distribution in Fig. 19a, on one hand, highlights the individuality of the different model data sets with regard to their behaviour which is expressed through characteristic combinations of index values. On the other hand, it provides a vivid evidence of the high discriminatory power of the SOM approach.

The possibility of a direct comparison of model behaviour properties lends itself for a series of potential applications, e.g. in a model ensemble framework: The proportion of "equivalent" model realizations in a data set obtained from the results of an ensemble simulation, in turn, could serve as a "proxy" for the independence of model structures. That way, a set of model realizations or model structures that together cover a broader range of measured system behaviour than each individual model (e.g. model realizations that emphasize different sections of the FDC) could be determined and constitute the base of a (multi-)model ensemble application (Fenicia et al., 2007).

Acknowledgements. Financial support for this research provided by the German Federal Ministry of Education and Research (BMBF, grant 0330699D) is gratefully acknowledged.

Edited by: A. Schumann
Reviewed by: two anonymous referees

References

Bossel, H.: Systeme, Dynamik, Simulation – Modellbildung, Analyse und Simulation komplexer Systeme, Books on Demand, Norderstedt, 2004.

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.

Bremicker, M.: Das Wasserhaushaltsmodell LARSIM. Modellgrundlagen und Anwendungsbeispiele, Freiburger Schriften zur Hydrologie, 2000.

Casper, M. C., Herbst, M., Grundmann, J., Buchholz, O., and Bliefernicht, J.: Einfluss der Niederschlagsvariabilität auf die Simulation extremer Abflüsse, Hydrology and Water Resources Management Germany, accepted, 2009.

Disse, M., Pakosch, S., and Yörük, A.: Development of an operational expert system for flood forecasts considering prediction uncertainty, Hydrologie und Wasserbewirtschaftung, 51, 210–215, 2007.

Duan, Q.: Global optimization for watershed model calibration, in: Calibration of Watershed Models, edited by: Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N., and Turcotte, R., Water Science and Application, AGU, Washington D.C., 89–104, 2003.

Duan, Q., Sorooshian, S., and Gupta, V. K.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015–1031, doi:10.1029/91WR02985, 1992.

Fenicia, F., Solomatine, D. P., Savenije, H. H. G., and Matgen, P.: Soft combination of local models in a multi-objective framework, Hydrol. Earth Syst. Sci., 11, 1797–1809, 2007, http://www.hydrol-earth-syst-sci.net/11/1797/2007/.

Gan, T. Y. and Biftu, G. F.: Effects of model complexity and structure, parameter interactions and data on watershed modelling, in: Calibration of Watershed Models, edited by: Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N., and Turcotte, R., Water Science and Application, AGU, Washington D.C., 317–329, 2003.

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, 751–764, 1998.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, doi:10.1002/hyp.6989, 2008.

Haykin, S.: Neural networks – a comprehensive foundation, 2nd ed., New Jersey, 842 pp., 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/.

Herbst, M., Gupta, H. V., and Casper, M. C.: Mapping model behaviour using Self-Organizing Maps, Hydrol. Earth Syst. Sci., 13, 395–409, 2009, http://www.hydrol-earth-syst-sci.net/13/395/2009/.

Hydrotec: Rainfall-Runoff-Model NASIM – program documentation (in German), Hydrotec GmbH, Aachen, 579 pp., 2005.

Kalteh, A. M., Hjorth, P., and Berndtsson, R.: Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application, Environmental Modelling & Software, 23, 835–845, doi:10.1016/j.envsoft.2007.10.001, 2008.

Kaski, S.: Data Exploration Using Self Organizing Maps, Dr. thesis, Department of Computer Science and Engineering, Helsinki University of Technology, Helsinki, 57 pp., 1997.

Klemeš, V.: Conceptualization and scale in hydrology, Journal of Hydrology, 65, 1–23, 1983.

Kohonen, T.: Self-Organizing Maps, 3rd ed., Information Sciences, Springer, Berlin, Heidelberg, New York, 501 pp., 2001.

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

Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Döll, P., Kabat, P., Jiménez, B., Miller, K. A., Oki, T., Sen, Z., and Shiklomanov, I. A.: Freshwater resources and their management, in: Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Parry, M. L., Canziani, O. F., Palutikof, J. P., van der Linden, P. J., and Hanson, C. E., Cambridge University Press, Cambridge, 173–210, 2007.

Madsen, H.: Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, Adv. Water Resour., 26, 205–216, 2003.

Merz, B. and Didszun, J.: Risikomanagement extremer Hochwasserereignisse, USWF – Zeitschrift für Umweltchemie und Ökotoxikologie, 17, 191–192, 2005.

Merz, B., Didszun, J., and Ziemke, B.: RIMAX Risikomanagement extremer Hochwasserereignisse, 2. Auflage ed., GeoForschungsZentrum Potsdam (GFZ), Potsdam, 51, 51 pp., 2007.

Minns, A. W. and Hall, M. J.: Artificial Neural Network Concepts in Hydrology, in Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., 307–320, 2005.

Moore, C. and Doherty, J.: Role of the calibration process in reducing model predictive error, Water Resour. Res., 41, W05020, doi:10.1029/2004WR003501, 2005.

Peschke, G.: Soil Moisture and Runoff Components from a Physically Founded Approach, Acta hydrophysica, 31(3/4), 191–205, 1987.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth Syst. Sci. Discuss., 5, 3169–3211, 2008, http://www.hydrol-earth-syst-sci-discuss.net/5/3169/2008/.

Sivapalan, M.: Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, in: Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., Wiley, 193–220, 2005.

Schulla, J. and Jasper, K.: Model description WaSIM-ETH, Zürich, 167 pp., 2001.

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, 2003.

Van Genuchten, M. T.: A Closed-Form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils, Soil Sci. Soc. Am. J., 44(5), 892–898, 1976.

Vesanto, J.: SOM-based data visualization methods, Intelligent Data Analysis, 3(2), 111–126, 1999.

Vesanto, J.: Neural network tool for data mining: SOM toolbox, Symposium on Tool Environments and Development Methods

for Intelligent Systems (TOOLMET2000), Oulu, Finnland, 184–196, 2000.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J.: SOM Toolbox for Matlab 5, Helsinki University of Technology, Report A57, Espoo, Finland, 60 pp., 2000.

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

Zadeh, L.: Optimality and non-scalar-valued performance criteria, IEEE T. Automat. Contr., 8, 59–60, 1963.

Zhao, R. J.: Flood forecasting method for humid regions of China, East China College of Hydraulic Engineering, Nanjing, 1977.