Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification

Rita Ley, Hugo Hellebrand, Markus C. Casper and Fabrizio Fenicia

ABSTRACT

The ability of a hydrological model to reproduce observed streamflow can be represented by a large variety of performance measures. Although these metrics may suit different purposes, it is unclear which of them is most appropriate for a given application. Our objective is to investigate various performance measures to assess model structures as tools for catchment classification. For this purpose, 12 model structures are generated using the SUPERFLEX modelling framework, which are then applied to 53 meso-scale basins in the Rhineland-Palatinate (Germany). Statistical and hydrological performance measures are compared with signature indices derived from the flow duration curve and combined into a new performance measure, the standardized signature index sum (SIS). The performance measures are evaluated in their ability to distinguish the relative merits of various model alternatives. In many cases, classical and hydrological performance measures assign similar values to different hydrographs. These measures, therefore, are not well suited for model comparison. The proposed SIS is more effective in revealing differences between model results. It allows for a more distinctive identification of a best performing model for individual basins. A best performing model structure obtained through the SIS can be used as basin classifier.

Key words | catchment classification, flow duration curve, hydrological modelling, model evaluation, model performance, signature indices

Rita Ley (corresponding author)
Hugo Hellebrand
Markus C. Casper
University of Trier, Physical Geography,
Trier,
Germany
E-mail: *leyrita@uni-trier.de*

Fabrizio Fenicia
Swiss Federal Institute of Aquatic Science and Technology (Eawag),
Dübendorf,
Switzerland

INTRODUCTION

The selection of an appropriate model for a basin critically depends on the basin characteristics and its dominant runoff processes. This inevitably leads to modelling approaches that recognize the different characteristics of individual systems (e.g. Leavesley *et al.* 1996; Fenicia *et al.* 2011; Coxon *et al.* 2014). In conceptual hydrological modelling, flexible and multi-model frameworks have been already used to examine, for example patterns of structural errors across multiple basins (Clark *et al.* 2008); mean residence time and basin mixing mechanisms (McMillan *et al.* 2012; Hrachowitz *et al.* 2014); representations of plot-scale surface and groundwater dynamics (Krueger *et al.* 2010) and time scale control on model parameters and inferred complexity (Kavetski *et al.* 2011).

Previous work has shown that when a set of model structures is applied to multiple basins, their performance may rank differently (e.g. Duan *et al.* 2006; Fenicia *et al.* 2013; van Esse *et al.* 2013). As a result, the performance of the different structures may provide information on the similarities and differences between basins and may therefore be used as a catchment classifier. For this approach to succeed, it is necessary to select an appropriate measure for evaluating model performance to identify a best performing model structure for each basin.

doi: 10.2166/nh.2015.221

Streamflow simulations are typically assessed through a wide range of performance measures. These metrics include statistical performance measures (e.g. Pearson correlation coefficient (Pearson 1895); weighted R^2 (Krause *et al.* 2005)), hydrological performance measures (e.g. Nash and Sutcliffe Efficiency (NSE) (Nash & Sutcliffe 1970); volumetric efficiency (Criss & Winston 2008)), performance metrics that are derived from flow duration curves (FDCs) (Yilmaz *et al.* 2008) and other hydrological signatures (e.g. Westerberg *et al.* 2011; Coxon *et al.* 2014). The assessment of model performance may be conducted using these metrics in isolation or combined in a single objective function (e.g. Kling *et al.* 2012). Alternatively, these metrics may also be used simultaneously in a multi-objective framework (e.g. Yilmaz *et al.* 2008).

All known performance measures have specific strengths, weaknesses and sensitivities to various parts of the hydrograph (Legates & McCabe Jr 1999; Krause *et al.* 2005; Schaefli & Gupta 2007; Gupta & Kling 2011; Pushpalatha *et al.* 2012). For example, the NSE favours the simulation of the peaks, the R^2 reveals similarities in the dynamics, but disregards differences in the absolute values, whereas the FDC provides indications on the flow distribution, disregarding potential timing errors.

The FDC is perceived to represent a meaningful descriptor of catchment response. Blöschl *et al.* (2013) describe FDCs as 'a key signature of runoff variability' which 'can be used for evaluating rainfall–runoff model output and for calibrating such models'. The use of an FDC gives more information about the hydrological behaviour of the modelled basins (Hrachowitz *et al.* 2014) and their underlying hydrological processes (Yilmaz *et al.* 2008; Gupta *et al.* 2009; Wagener & Montanari 2011).

FDCs are often used in hydrology, e.g. for model evaluation (e.g. Yilmaz *et al.* 2008; Herbst *et al.* 2009; Westerberg *et al.* 2011) or catchment grouping (e.g. Carrillo *et al.* 2011; Sawicz *et al.* 2011). A comparison of FDCs can be done with one value for the whole FDC (Ganora *et al.* 2011; Sauquet & Catalogne 2011) or with multiple indices which consider where differences between two FDCs occur. Westerberg *et al.* (2011) and Coxon *et al.* (2014) use several evaluation points, whereas Yilmaz *et al.* (2008) propose indices describing meaningful parts of the FDC. Furthermore, the use of the FDC is not restricted to the entire curve. For certain research questions specific parts of the curve can be used as well (Herbst *et al.* 2009; Sawicz *et al.* 2011; Coxon *et al.* 2014).

The aim of this study is to test the appropriateness of statistical performance measures, hydrological performance measures and performance measures derived from the FDC to identify a best performing model out of various calibrated models for 53 basins with a view to basin classification. Basins that are characterized by the same model structure may build a class of similar basins. The structures of these models are generated within the SUPERFLEX modelling framework, which facilitates model development and enables controlled model comparison (Fenicia *et al.* 2011).

STUDY AREA AND DATA

The study area consists of 53 small to medium-sized gauged basin areas in Rhineland-Palatinate (RLP), Germany (Figure 1 and Appendix 1 (available in the online version of this paper)). The basins lie in low mountain ranges of the Rheinisches Schiefergebirge, the Saar-Nahe-Bergland and the Rhine Valley. Among these 53 basins, there are 35 headwater basins, of which three are triply nested. Basin sizes vary from 10 to 1,469 km^2; 48 basins are less than 400 km^2 and two are larger than 1,000 km^2. Elevation ranges between 100 and 818 m above sea level (a.s.l.) with a mean elevation of 341 m a.s.l. Geology differs from schist, greywacke, and quartzite to sedimentary rock with tertiary and quaternary volcanism (basaltic rocks, pumice stone and tuff). Almost all the basins are rural with little urbanization except for three basins, which are moderately urbanized (11–13%). Agricultural land use varies between 7 and 90%, for most of the basins between 40 and 80%. Some basins, especially in the southeast, support viticulture and orchards.

All basins within the study area belong to the same climatic region, but depending on altitude and location, show climatic peculiarities, caused by precipitation ranges from 530 mm/y in the southeast up to 1,108 mm/y in the west and differences in temperature and potential evaporation. Runoff behaviour varies between high reactivity and variability and high event runoff coefficients in wet basins with steep slopes and low storage and low reactivity and variability and low runoff coefficients in dry, mostly flat basins with high storage capacities (Ley *et al.* 2011; Ley 2014).

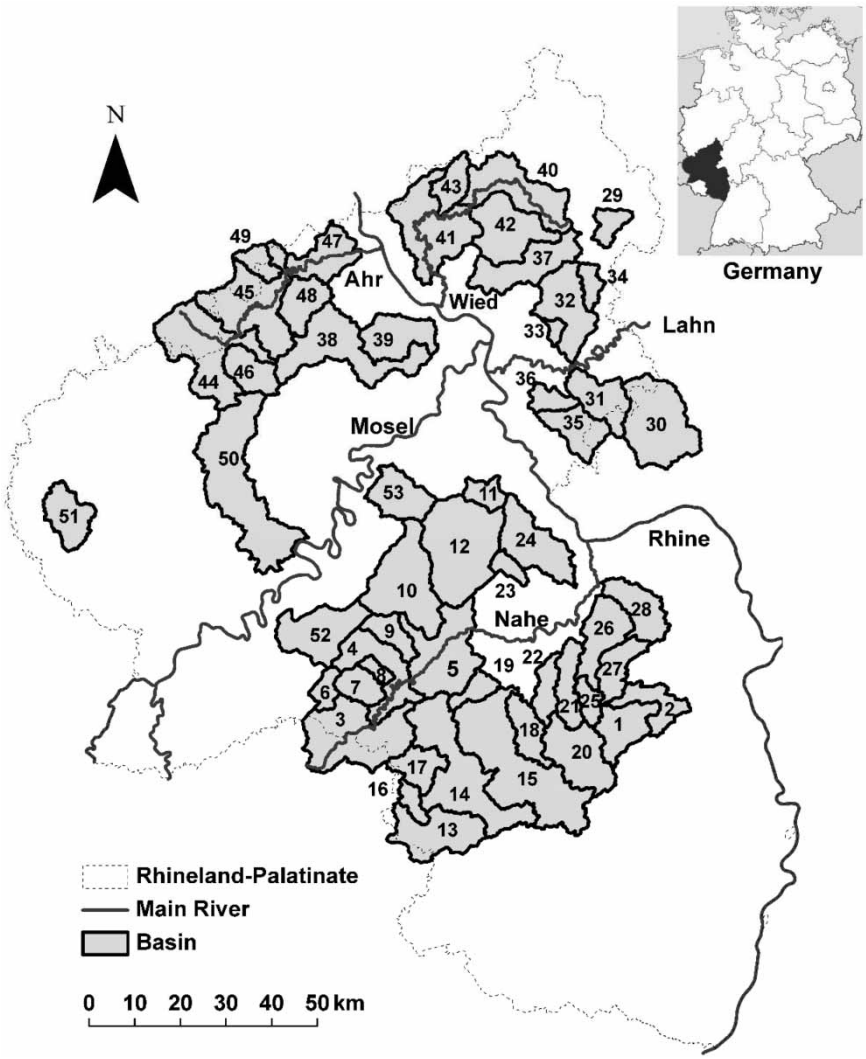

Figure 1 | The 53 basins of the study area in Germany. Nested basins are sorted by size: only the headwater basins are displayed completely. Numbers refer to list of basins (ID) in Appendix 1 (available in the online version of this paper). Small map of Germany: Wikipedia.org/wiki/Datei:Locator_map_Rhineland-Palatinate_in_Germany.svg (last accessed June 2014).

For the model application, hourly runoff, areal precipitation and temperature data for the period from January 1996 to December 2003 are used. These time series cover a wide range of diverse annual or seasonal precipitation and runoff events.

Areal precipitation was calculated with 'InterMet' (Gerlach 2006), which interpolates meteorological data using Kriging. To calculate areal precipitation for Rhineland-Palatinate and adjacent areas, InterMet takes into account data form about 200 rain gauges, meteorological data, prevailing atmospheric conditions, orography, and satellite and radar data. Typical rainfall fields extend in the range of most of the basin sizes. In summer, some mostly convective rainfall events affect only parts of the basins. Although snow events do occur in the study area, they are of minor importance since they are limited in amount, not prolonged and irregular. Snow processes are thus not considered in the modelling exercise.

METHODOLOGY

Model structures

The SUPERFLEX modelling framework can be used to perform model comparisons through constructing models that

differ in a controlled way. In the present case, it was applied using 12 model structures as proposed by Fenicia *et al.* (2013), which include serial, linear and parallel model structures with different numbers of reservoirs and parameters, thus covering a relatively broad range of conceptual model complexities (Figure 2). Starting from the simplest structure (model structure 1 (M01), Figure 2), the complexity gradually increases by adding reservoirs and lag-functions to the most complex structure (model structure 12 (M12), Figure 2). Although Fenicia *et al.* (2013) and van Esse *et al.* (2013) describe the models extensively, a short explanation of the structures follows.

Model structure 1 consists of a single reservoir with a nonlinear storage-discharge relationship characterized by a time constant and a power parameter. Model structure 2 also consists of a single reservoir, but with an upper threshold, and uses a linear function to describe outflow and a power function to describe flow exceeding the threshold. Model structures 3, 4, 5 and 6 show serial reservoir connections. These models differ in the constitutive functions used to describe the flows between the reservoirs and in the number of calibrated parameters. Model structure 3 describes the flow between the reservoirs as a threshold function while the model structures 4, 5 and 6 use power functions to describe outflows. Model structure 5 has a lag-function to represent hydrograph delay. Compared to model structure 5, model structure 6 has an interception reservoir and model structure 7 has a riparian reservoir. Model 8 is a simple parallel structure with two parallel reservoirs, namely a fast reservoir and a slow reservoir and a precipitation partitioning parameter. The model structures 9, 10, 11 and 12 build on model structure 8 with increasing

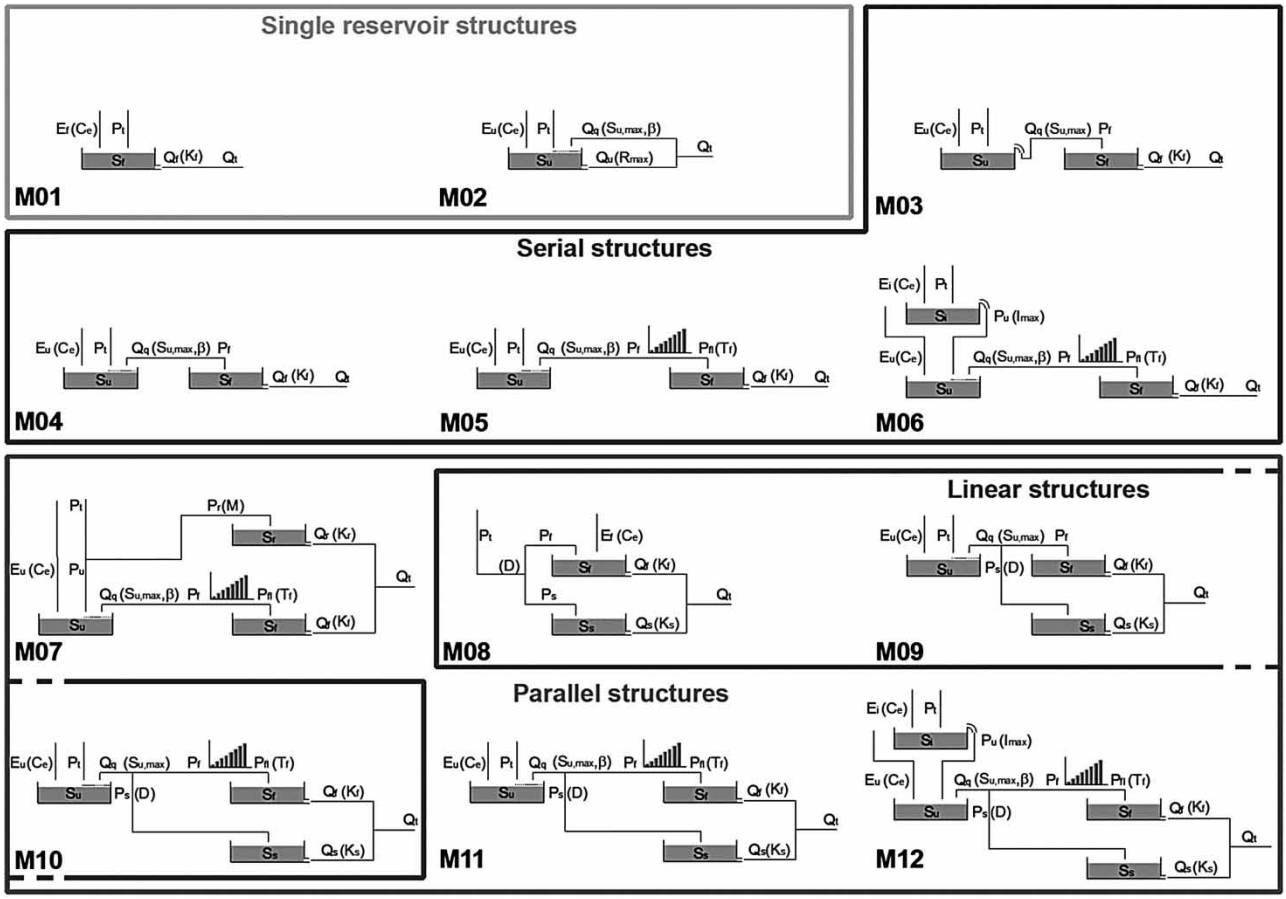

Figure 2 | Model structures of SUPERFLEX modelling framework. Adapted from Fenicia *et al.* (2013). D: partition between fast and slow reservoir; Ef (Ce): actual evaporation; Eu(Ce): unsaturated evaporation; K: storage coefficient, f = fast, s = slow; Pt: total precipitation; Qf: discharge from fast storage; Qq, Qu: discharge from unsaturated storage; Qs: discharge from slow storage; Qt: total modelled discharge; Sf: storage: f = fast, s = slow, u = unsaturated; Su,max: maximum storage unsaturated reservoir; β: power function.

complexity and an unsaturated reservoir preceding the parallel structure. Figure 2 details the 12 model structures.

Calibration approach

Fifty-three basins in RLP provide concurrent rainfall and runoff data. The warm-up period of the calibration consists of the first year of the data period (1996). Following a split-sample approach (Klemeš 1986), the remaining range (1997–2003) is subdivided into a calibration period and a validation period of equal length.

The calibration objective function is based on a weighted least squares approach, assuming independent Gaussian error with zero mean and standard deviation linearly proportional to the modelled discharge (Kavetski & Fenicia 2011). Optimization is carried out through a quasi-Newton method with 20 multi-starts randomly selected across the parameter space. The determination of one calibrated model for each basin and each structure results into 636 calibrated models and hence 636 validated models.

Diagnostics

'Hydrological model' and 'model structure' can be befuddling terms when not clearly defined. This study considers a hydrological model as a combination of a specific model structure with a particular parameter set. The calibration of this parameter set displays an optimal model for a given forcing data set (basin). If multiple model structures are used, the identification of the optimal model with the highest performance defines the best performing model structure.

The identification of a best performing model structure for a given basin commences with comparing observed with simulated runoff time series by means of a performance measure. This study tests three types of performance measures: (1) statistical performance measures, (2) hydrological performance measures and (3) performance metrics from the FDC. Appendix 2 (available in the online version of this paper) contains the mathematical formulations of all measures.

1. Statistical performance measures:
   - Root mean square error (RMSE)
   - Pearson product-moment correlation coefficient (Pearson 1895)
   - Weighted R^2 (Krause *et al.* 2005)
   - Spearman's rank correlation coefficient (Spearman 1904)
2. Hydrological performance measures:
   - NSE (Nash & Sutcliffe 1970)
   - Modified NSE (without squaring values) (Krause *et al.* 2005)
   - Index of agreement (Willmott 1981)
   - Modified index of agreement with (without squaring values) (Krause *et al.* 2005)
   - Kling–Gupta efficiency (Gupta *et al.* 2009; Kling *et al.* 2012)
   - Volumetric efficiency (Criss & Winston 2008)
3. Performance metrics from the FDC:
   - SIS: combination of four performance metrics:
     - FHV: very high flow (Yilmaz *et al.* 2008)
     - FMV: high flow
     - FMS: slope of the mid-segment FDC (Yilmaz *et al.* 2008)
     - FLV: low flow (Gronz 2013; Yilmaz *et al.* 2008)

The 12 selected model structures are calibrated for all basins and the above-listed performance measures for each individual model (i.e. structure + parameter set) and basin are calculated as well. All performance measures are analysed with a view to redundancies, explanatory power and suitability to identify a best performing model for a basin.

The statistical and hydrological performance measures describe the overall performance of a model with one value. The four performance metrics that are derived from the FDC examine the influence of specific aspects of the hydrograph on model performance. The FDC is the complement of the cumulative distribution function of stream flow (Vogel & Fennessey 1994). Despite the fact that FDCs include no information on timing of the flow, they are still a useful way of comparing observed and simulated runoff. A poorly reproduced FDC is an indication of poor model performance. Therefore, the comparison between simulated and observed FDC is a powerful descriptor of model performance.

To compare FDCs, the study adopts the approach proposed by Yilmaz *et al.* (2008) who developed so-called signature indices derived from FDCs. These indices

represent for specific parts of an FDC the bias between observed values and simulated values proportional to the observed FDC and have proven their usefulness in several applications (Casper *et al.* 2012; Gronz 2013; Herbst *et al.* 2009; Ley *et al.* 2011). The indices describe major behavioural functions of a basin: extreme high runoff (FHV), mid-slope of the FDC (FMS) and the low flow (FLV). Gronz (2013) modified the FLV index to prevent misleading indices. A fourth signature index for the high flow between extreme high and medium runoff (FMV) is added in order to consider the whole FDC (Figure 3). This study summarizes the above-listed indices into the term 'signature indices'.

Each of the four signature indices reflect the model's ability to reproduce specific parts of the hydrograph. A combination of the four indices into one value should reflect the ability of the model to reproduce the entire hydrograph. Since the four indices can have different orders of magnitude, an approach is needed that weights them equally when combined. A straightforward method to equally combine them is to standardize the indices and then to summarize. This results in a new performance measure: the standardized signature index sum (SIS).

The SIS is calculated as follows:

(1) Calibration of all model structures on all catchments, and calculation of the four signature indices x_{sia} (where s indicates the structure, i indicates the catchment, a the type of signature index, and x its value).
(2) Calculation of the absolute value of each signature index $|x_{\mathrm{sia}}|$ (since the sign is irrelevant, the absolute values treat under or overestimation equally).
(3) Calculation of the standard deviation σ_a and the mean $\bar{x}_a$ of $|x_{\mathrm{sia}}|$ for all i and s.
(4) Calculation of the standardized values (z-score); Equation (1).
(5) Combining the standardized values; Equation (2).

The sum of the four standardized signature indices of one model for a given basin describes the deviation for the entire FDC and therefore the performance for a particular model and basin. The lowest SIS value for a given basin identifies the best performing model for this basin.

$$Z_{\mathrm{sia}} = \frac{|x_{\mathrm{sia}}| - \bar{x}_a}{\sigma_a} \tag{1}$$

$$\mathrm{SIS}_{\mathrm{si}} = Z_{\mathrm{siFHV}} + Z_{\mathrm{siFMV}} + Z_{\mathrm{siFMS}} + Z_{\mathrm{siFLV}} \tag{2}$$

Index	FDC	meaning
FHV	Difference of the volume of the very high flow (exceed. prob. < 2%)	very high flow: reaction to large precipitation events (Yilmaz et al., 2008)
FMV	Difference of the volume of the high flow (exceed. prob. 2-20%)	high flow: variability of reaction to heavy rainfall
FMS	Difference between the slopes of the mid segment FDC (exceed. prob. 20-70%)	medium flow: vertical redistribution of water; reactivity of the basin. (Yilmaz et al., 2008)
FLV	Difference of the volume of the low flow (exceed. prob. > 70 %)	low flow; base flow (Yilmaz et al., 2008)

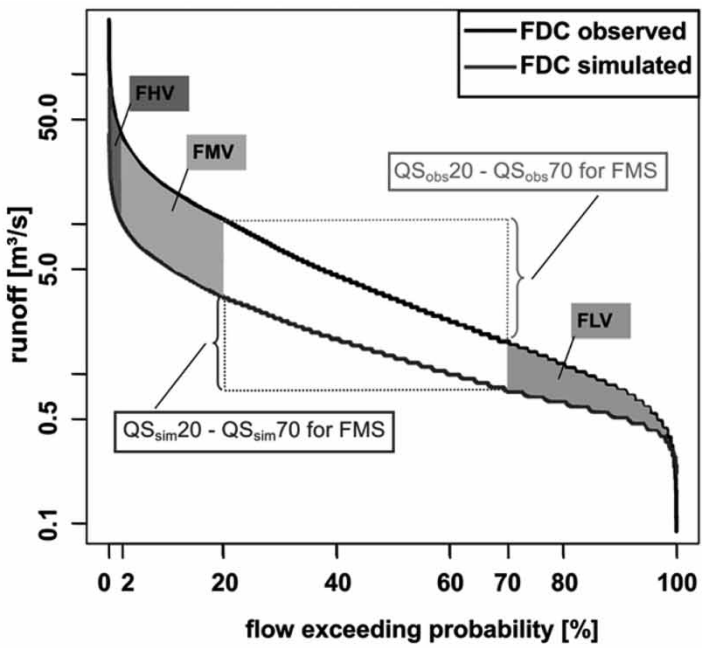

Figure 3 | Meaning and position of the signature indices at the FDC.

where SIS = standardized signature indices sum; a = signature index (i.e. FHV, FMV, FMS or FLV); s = model structure; i = basin; x = value of a signature index; $\bar{x}_a$ = mean of all values of one signature index a; σ_a = standard deviation all values of one signature index a.

RESULTS AND DISCUSSION

Concerning model assessment, validation is an important step, since it provides independent information on model consistency (Klemeš 1986; Andréassian *et al.* 2009). Comparing the results of the calibrated models with the validated models, only minor differences occur with reference to model performance. The analysis therefore uses the simulated results of the calibrated models.

Identification by classical statistical performance measures and hydrological performance measures

The patterns of model performance calculated with classical statistical or hydrological performance measures are very similar. As an example of model performance on a given catchment, Figure 4 displays the results of the different performance measures for the simulated runoff of the basin 'Flaumbach' at the gauging station 'Kloster Engelport'. The calibrated models based on structures 1, 2 and 8 demonstrate a worse performance than the other models, regardless of the performance measure. The performances of the other nine models demonstrate almost similar performances. The simulated runoff of the other basins displays a similar pattern as depicted in Figure 4. Owing to this pattern, the choice of a classical statistical or hydrological performance measure seems to be less important for the identification of a best performing model structure. Therefore, the NSE is chosen for further analysis.

Fenicia *et al.* (2013) as well as van Esse *et al.* (2013) found patterns between model structure and model performance similar to our results. Enlarging the parameter space (to avoid too small parameter boundaries affecting model performance) ensued in better results for the model structures 9, 10, 11 and 12 than in van Esse *et al.* (2013). Hydrological differences in the respective study areas could attribute to this, but it may well be that limitations in parameter space in the French modelling exercise cause the differences.

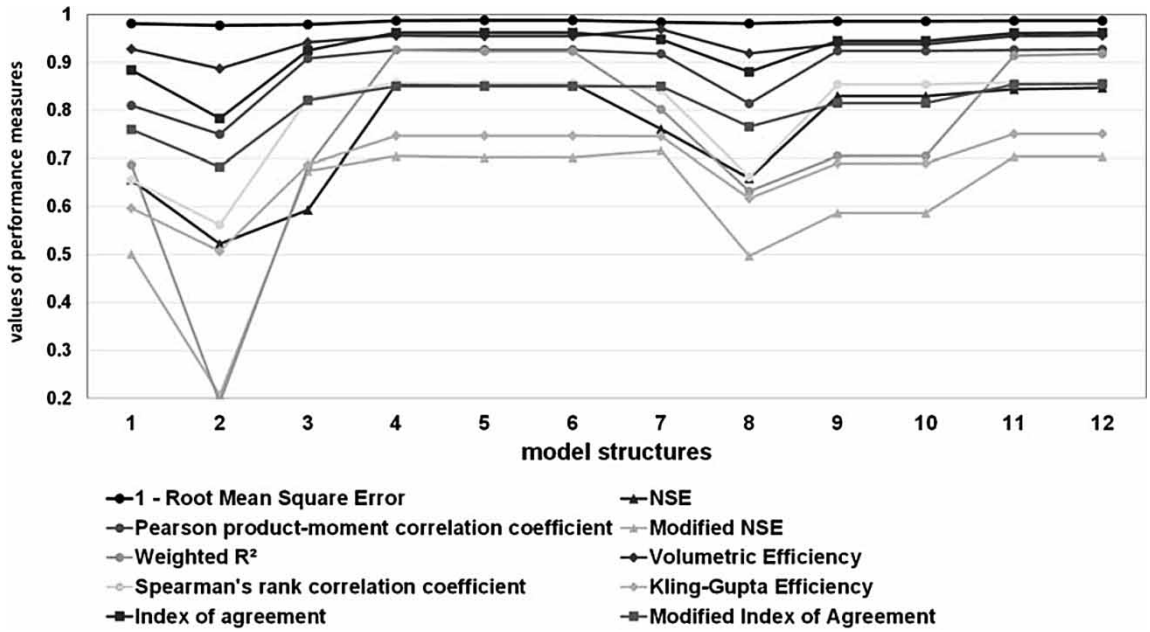

Figure 4 | Different performance measures for the simulated runoff at gaging station 'Kloster Engelport' demonstrate very similar patterns. Each line represents one performance measure. With 1-RMSE, all shown performance measures have a maximum value of 1. Their ranges vary between –Inf to 1, from –1 to 1, or from 0 to 1. For the shown basin and models, the values of the performance measures range between 0 and 1 and therefore, all measurements can be displayed on a single *Y*-axis.

The sequence of the NSE between model performance and model structure for individual basins determines a best performing model structure for each basin. For most of the basins, several model structures display almost identical values for the NSE (Figure 5). The conceptual differences between the model structures 4, 5 and 6; 9 and 10 and 11 and 12 are minor (Figure 2) and apparently result in very similar NSEs for these model structures. Often, also conceptually different model structures display similar NSE values with differences of less than 0.05. This effect obscures the identification of a decidedly best performing model structure for a single basin.

Only model structure 3 displays a wide range of performance values for the NSE, which allows for the identification of basins with a good performance. The performance of model structure 3 is worse for basins with low precipitation or low total runoff coefficients. Probably, the threshold overflow between the two reservoirs, which is a specific characteristic of structure 3, causes this. This may give indications about the process representation by this structure, e.g. an indication of a threshold-like response. However, this is a distinction in wet and dry basins, which is a trivial result.

Identification by signature indices

Figure 6 displays the FDCs (observed and simulated with different structures) for the two gaging stations Weinähr and Wernerseck. These stations have similar NSE values for more than one model, but differ when it comes to their simulated FDCs. This shows that simulations for one basin with different models and with similar NSEs need not have similar hydrographs.

From Figure 6 it is difficult to decide which simulated FDC performs better. Figure 7 displays the values of the signature indices for all basins as box-and-whisker plots. For most of the basins, the four signature indices show clear differences in performance. Except for the models that are based on structure 3, all models underestimate the very high flow (FHV). For the other three signatures, most of the structures perform well. In general, the models of structure 8 have very low values for the NSE, caused by high deviations for the very high flow. However, when it comes to the signature indices, structure 8 shows a good performance for the other parts of the FDC. Analogous to the large variation in NSE values for structure 3, the four indices of the FDC for this structure display a wide range as well.

As for the above-mentioned statistical and hydrological performance measures, most of the simulated FDCs indicate a better agreement for structures 4, 5, 6, 7, 11 and 12 than for the other model structures (Figure 7). Although the NSEs for structures 9 and 10 are good, their simulated FDCs show for most basins a levelled curvature and thus obtain a bad signature index performance. The Steinbach gaging station provides a good example for the different relationships between the NSE and the signature indices of the FDC. Steinbach has NSEs between 0.63 and 0.69 for structures 4, 5, 6, 7, 9, 10, 11 and 12 and poor NSEs of

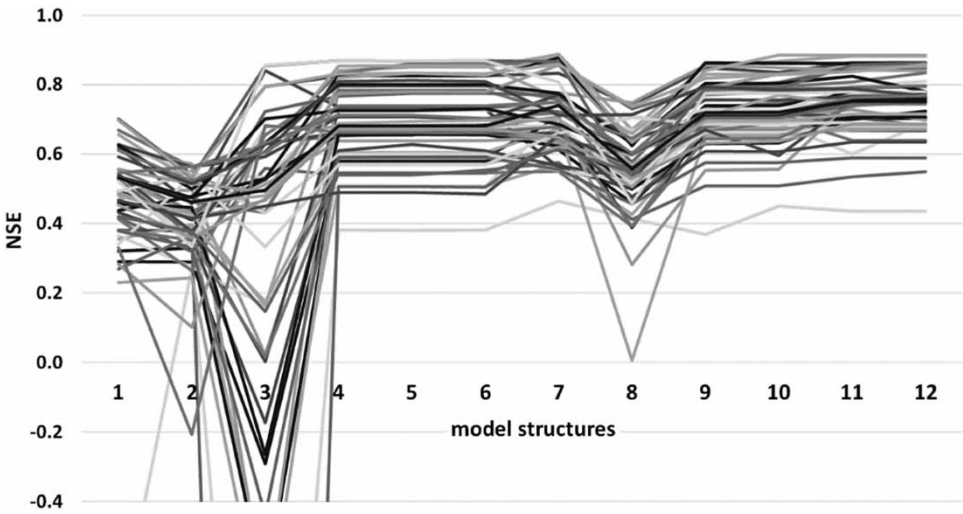

Figure 5 | NSE of all basins for all models. Each line represents the NSE of one basin for all models. Very low NSEs (<0.4) are not show here for clarity.

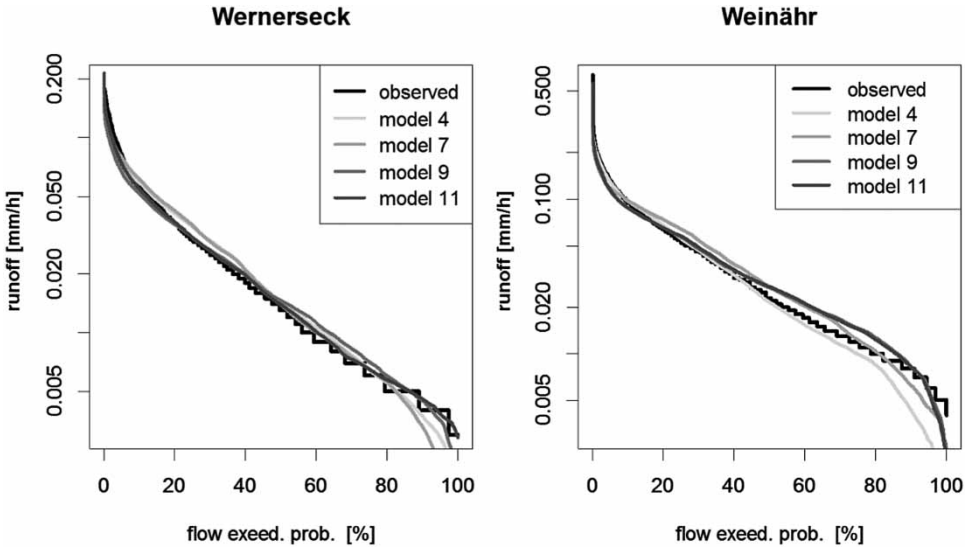

Figure 6 | Observed and simulated FDCs for two representative catchments. The FDCs of models based on structures 4, 9 and 11 are similar to 5, 6, 10, respectively, 12. NSE for the models based on structures 4, 7, 9 and 11: Weinähr: 0.81/0.71/0.79/0.79; and Wernerseck 0.69/0.70/0.69/0.72. The NSE for the models based on structures 1, 2, 3 and 8 are much lower.

0.45 and 0.47 for structures 1 and 3. Figure 8 displays the measured and simulated FDCs for each structure in detail, reflecting the index values depicted in Figure 7. However, model structure 3 shows a distinctively better agreement between observed and simulated FDC than for the other model structures. Despite the fact that model structures 1 and 3 show similar NSEs, the simulated FDC of model structure 1 is apparently worse than the simulated FDC of model structure 3. The diverging results with the NSE may be caused by a high sensitivity of the NSE to overestimated extreme high discharge.

The SIS (Equation (1)) indicates an overall performance for a single basin. Negative values point to an above average good performance and the lowest value identifies the best performing model. Figure 9 displays the signature indices of the three gaging stations Weinähr, Seelbach and Wernerseck, listing the sum of the SIS and NSE as well. In contrast to the NSE, the SIS identifies one model as undeniably best performing. As for the NSE, the model structures 4, 5 and 6, 9 and 10, and 11 and 12 often have minor differences between their SIS. In these cases, the simpler model structures (4, 9, respectively, 11) are set as best performing.

From Figure 9 the following can be observed:

- For the gaging station Weinähr (basin size 215 km^2), the model based on structure 7 has the lowest SIS, which is due to a very low divergence from the observed FDC for high and mid runoff and a moderate divergence for low flow. Only the very high flow (FHV) shows a considerable bias, which is weighted lower by standardizing the biases for SIS.
- For the gaging station Seelbach (basin size 193 km^2), the model based on structure 4 has the lowest SIS. The models based on structures 7 and 12 show a slightly better NSE, which is caused by lower biases for FHV and disregarding better adaptions for the high and mean part of the FDC.
- For gaging station Wernerseck (basin size 242 km^2) the model based on structure 12 has the lowest SIS. The difference in NSE between the models based on structures 11 and 12 is 0.001 and is in this context negligible. With signature indices however, the differences in performance for these structures becomes apparent.

With respect to the 12 model structures, the simple models 3 and 4 show the best performance based on the SIS for 38% of the basins. These two models differ in that the outflow from the unsaturated reservoir in model 4 is a power function rather than a threshold function as in model 3 (Fenicia *et al.* 2013). The extension of model 4 with a lag function (model 5) and an interception reservoir

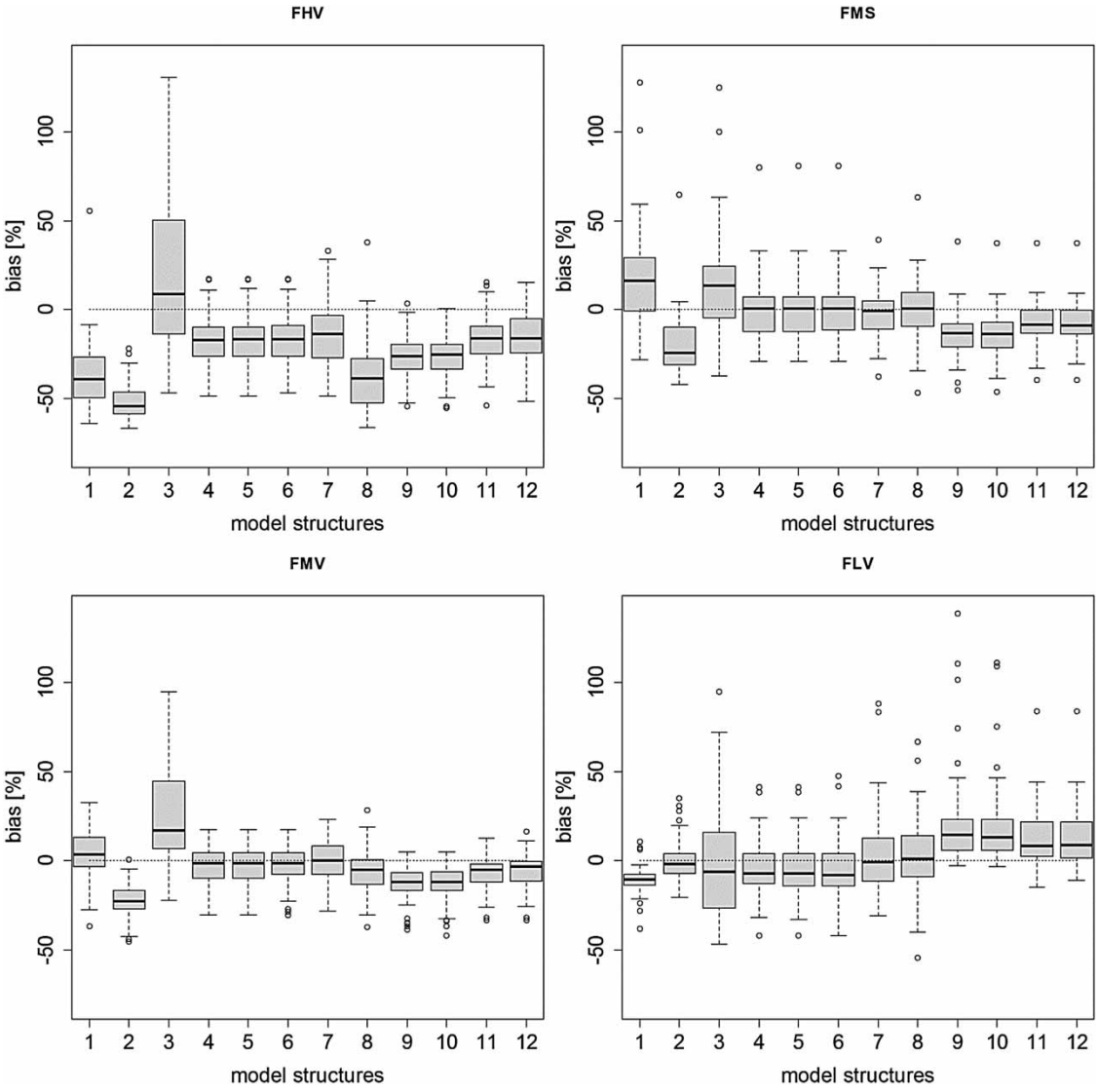

Figure 7 | Box-and-whisker plots of the signature indices of all basins and for all models. The boxes show the 25th and 75th percentile, the black line in the box the median, the end of the whiskers the maximum, minimum or, in the case of outliers (circles), the 1.5-fold interquartile range. Three outliers with signature indices larger than 150 (FHV model 3, FMS model 1 and FLV model 8) are not shown here for clarity. The dotted line marks zero bias, the best agreement between simulated and observed FDC.

(model 6) rarely leads to a better performance: model 6 outperforms model 4 only for one basin.

Model 7 is an extension of model 5 with an additional riparian zone reservoir that receives a constant fraction of the total precipitation (Fenicia *et al.* 2013). Although this model performs best in many cases, the performances of its simpler variants (4, 5) are often almost equally good. The same holds for the models 10 and 12 and their less complex counterparts 9 and 11, respectively: the gain in performance for these models (i.e. 10 and 12) is only marginal when compared to the performance of the less complex ones (i.e. 9 and 11). Therefore, the less complex models are preferable as catchment representation.

Although single indices indicate a very good performance for special parts of the FDC, the SIS recognizes the overall performance with a compensation of extreme values and considers equally all parts of the FDC to describe the overall performance. Furthermore, the SIS value allows evaluating the similarity of the performance of different

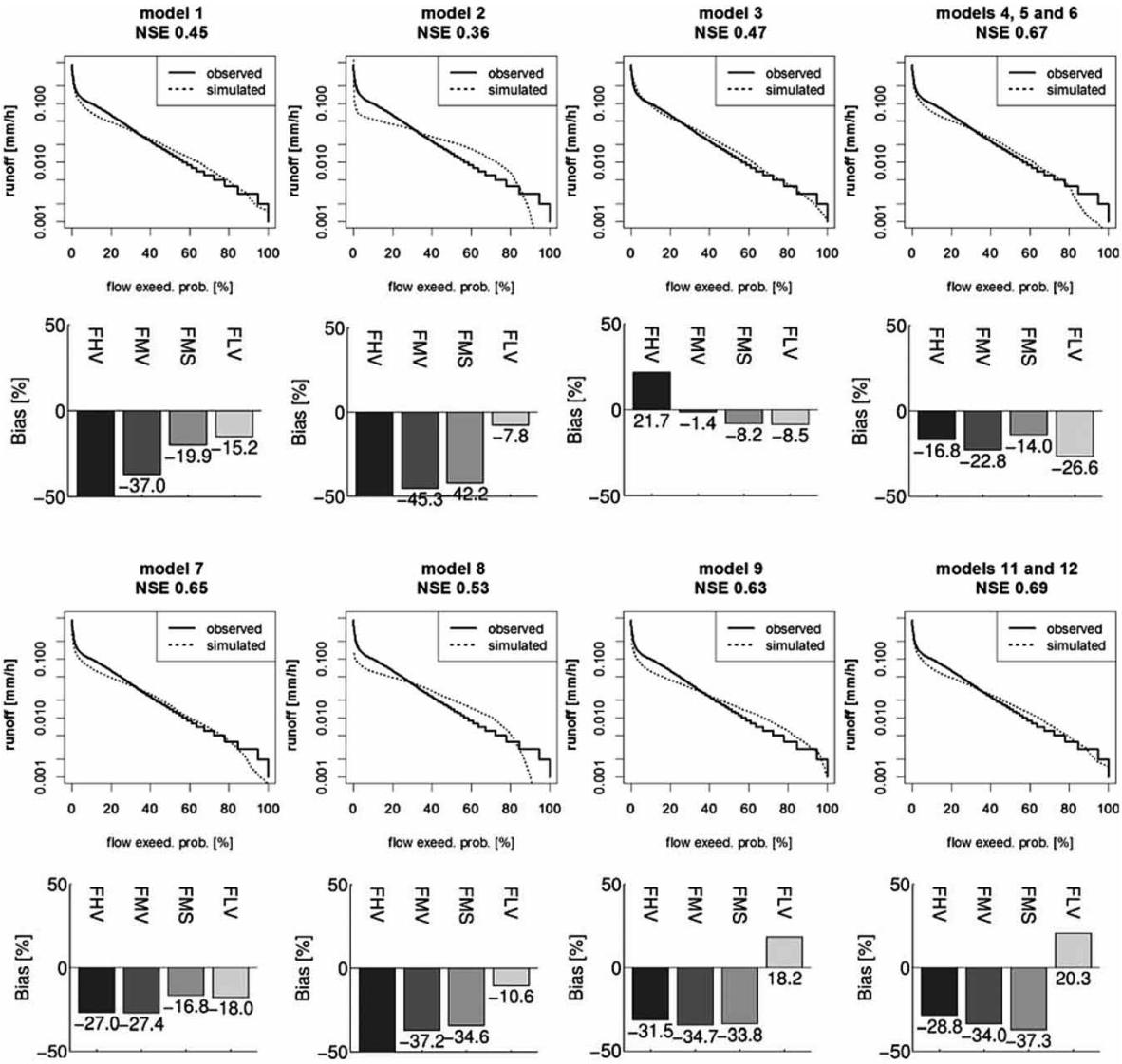

Figure 8 | FDC and biases of signature indices for gaging station Steinbach (46 km²). FDCs for models 4, 5 and 6 are nearly equal, also FDCs of models 11 and 12. FDCs for models based on structures 9 and 10 have very low differences, therefore only the results of model 9 are shown here.

models. This enables us to better differentiate, e.g. between 'good' or 'bad' performing models. In combination with the single indices, a decision for a best performing model consistent with special aspects of a given research question is now possible.

Since many active components of the hydrological cycle occur below the sub-surface (Beven & Freer 2001), it makes them difficult to observe. Thus far, human observation is largely barred from observing these sub-surface processes properly and this has a psychological consequence which Kahneman & Tverky (1982) called 'the perceptual best bet'. In hydrology this means that experience of above ground phenomena shape the expectation of the hidden sub-surface processes (Hellebrand 2010). Basin classification with hydrological modelling requires further research, where a larger number of model structures need testing to find optimal structures. However, to prevent the conception of 'perceptual best bet' models (i.e. models that are based upon our unobserved perception of the sub-surface), it would be of interest to automatically generate model structures by means of genetic programming, which would provide the modeller with new and unthought-of structures (hypotheses) that can be tested.

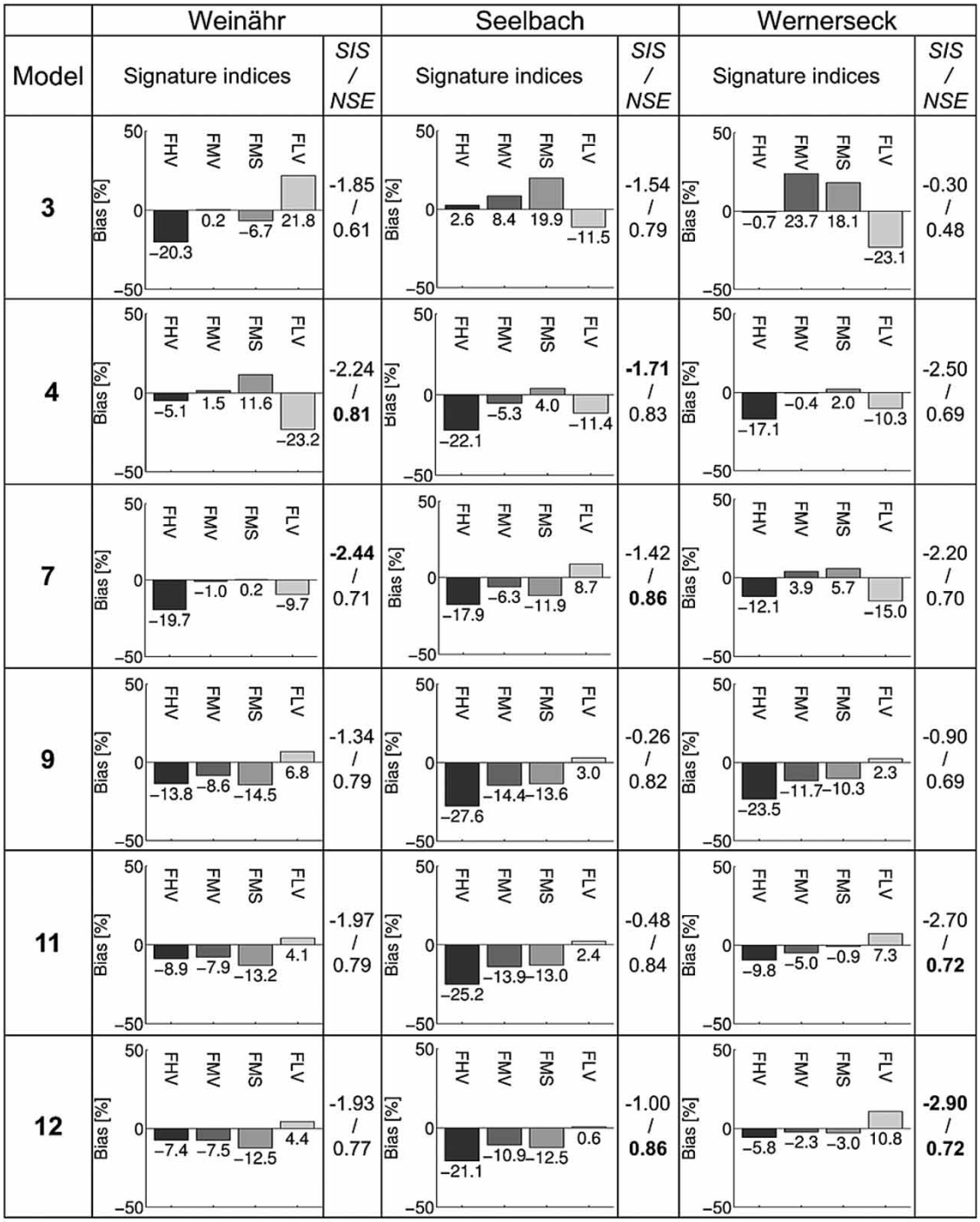

Figure 9 | Signature indices, SIS and NSE at the gaging stations Weinähr, Seelbach and Wernerseck for the models based on structures 3, 4, 7, 9 and 11 (values for model 4 are similar to model 5 and model 6, and values for model 9 are similar to model 10). The models 1, 2 and 8 show a bad performance for all basins and are therefore not listed here. The best values of SIS and NSE for each basin are printed in bold. Please note that for SIS the best value is the lowest one and for NSE is the highest value the best one.

CONCLUSIONS

This study compares different performance metrics to assess model performance with a view to catchment classification. If an insufficient number of classical performance measures are used simultaneously, they fail to discriminate between different model structures, providing similar values for seemingly different hydrographs. Signature indices derived from the FDC instead succeed in capturing differences between model results. Although standard hydrological performance measures are suitable to divide well from less well performing models, they show hardly any differentiation between good performing models. Since several of these performance measures are sensitive for special parts of the hydrograph, a bias between observed and modelled hydrographs for certain flow types (e.g. high flows) can mask a good performance for other parts of the hydrograph.

The four signature indices that calculate biases between the observed FDC and the simulated FDC, provide more differentiated results: they clearly identify at which part of the hydrograph these biases occur. The combination of the indices allows a decision for a best performing model that is consistent with specific aspects of a research question. The proposed SIS treats all parts of the FDC equally and makes a reasoned identification of a best performing model for single basins possible.

The use of signature indices for the evaluation of a best performing model structure is a promising way forward for basin classification by means of multiple hydrological model structures. There is clearly a need to expand the range of different types of hydrological model structures as well as to test this approach to different meso-scale research areas.

ACKNOWLEDGEMENTS

We acknowledge the financial support of the Deutsche Forschungsgemeinschaft (DFG) through grant CA728/5-1. Furthermore, we would like to thank the LUWG, Mainz (D) for providing the data. We also would like to thank both reviewers for their constructive comments.

REFERENCES

Andréassian, V., Perrin, C., Berthel, L., Moine, N. L., Lerat, J., Loumagne, C. & Valéry, A. 2009 Crash tests for a standardized evaluation of hydrological models. *Hydrol. Earth Syst. Sci.* **13**, 1757–1764.

Beven, K. & Freer, J. 2001 Equifinality, data assimilation and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **249**, 11–29.

Blöschl, G., Sivapalan, M., Wagener, T. & Viglione, A. (eds) 2013 *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*. Cambridge University Press, Cambridge.

Carrillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C. & Sawicz, K. 2011 Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient. *Hydrol. Earth Syst. Sci.* **15**, 3411–3430.

Casper, M. C., Grigoryan, G., Gronz, O., Heinemann, G., Ley, R. & Rock, A. 2012 Analysis of projected hydrological behavior of catchments based on signature indices. *Hydrol. Earth Syst. Sci.* **16**, 409–421.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Jasper, A. V., Gupta, H. V. & Hay, L. E. 2008 Framework for understanding structural errors (FUSE): a modular framework to diagnose differences between hydrological models. *Water Resour. Res.* **44**, doi: 10.1029/2007WR006735.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A. & Clark, M. 2014 Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrol. Process.* **28**, 6135–6150.

Criss, R. E. & Winston, W. E. 2008 Do Nash values have value? Discussion and alternate proposals. *Hydrol. Process.* **22**, 2723–2725.

Duan, Q., Schaake, J. V., Andréassian, S. F., Goteti, G., Gupta, H. V., Gusev, Y. M. & Wood, E. F. 2006 Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320**, 3–17.

Fenicia, F., Kavetski, D. & Savenije, H. H. G. 2011 Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resour. Res.* **47**, doi: 10.1029/2010WR010174.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L. & Freer, J. 2013 Catchment properties, function, and conceptual model representation: is there a correspondence? *Hydrol. Process.* **28**, 2451–2467.

Ganora, D., Claps, P., Laio, F. & Viglione, A. 2011 An approach to estimate nonparametric flow duration curves in ungauged basins. *Water Resour. Res.* **47**, doi: 10.1029/2008WR007472.

Gerlach, N. 2006 INTERMET Interpolation meteorologischer Größen. In Niederschlag-Abfluss-Modellierung zur Verlängerung des Vorhersagezeitraumes operationeller

Wasserstands- und Abflussvorhersagen, Kolloquium am 27. September 2005 in Koblenz (p. 98). Bundesanstalt für Gewässerkunde, Koblenz.

Gronz, O. 2013 Nutzung von Abflussprozessinformation in LARSIM. PhD Thesis. Universität Trier, Germany.

Gupta, H. V. & Kling, H. 2011 On typical range, sensitivity, and normalization of mean squared error and Nash–Sutcliffe Efficiency type metrics. *Water Resour. Res.* **47**, W10601.

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

Hellebrand, H. 2010 An applied hydrological spatio-temporal assessment of meso-scale basins with a view to regionalization. PhD Thesis. Delft University of Technology, Delft.

Herbst, M., Casper, M. M., Grundmann, J. & Buchholz, O. 2009 Comparative analysis of model behaviour for flood prediction purposes using self-organizing maps. *Nat. Hazards Earth Syst. Sci.* **9**, 373–392.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R. & Gascuel-Odoux, C. 2014 Process consistency in models: the importance of system signatures, expert knowledge, and process complexity. *Water Resour. Res.* **40** (9). doi: 10.1002/2014WR015484.

Kahneman, D. & Tverky, A. 1982 Variants of uncertainty. *Cognition* **11**, 143–157.

Kavetski, D. & Fenicia, F. 2011 Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resour. Res.* **47**, doi: 10.1029/2011WR010748.

Kavetski, D., Fenicia, F. & Clark, M. P. 2011 Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: insights from an experimental catchment. *Water Resour. Res.* **47**, W05501.

Klemeš, V. 1986 Operational testing of hydrological simulation-models. *Hydrol. Sci. J.* **13**, 13–24.

Kling, H., Fuchs, M. & Paulin, M. 2012 Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.* **424–425**, 264–277.

Krause, P., Boyle, D. P. & Bäse, F. 2005 Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* **5**, 89–97.

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E. & Haygarth, P. M. 2010 Ensemble evaluation of hydrological model hypotheses. *Water Resour. Res.* **46**, W07516.

Leavesley, G. H., Markstrom, S. L., Brewer, M. S. & Viger, R. J. 1996 *The Modular Modeling System (MMS) – the Physical Process Modeling Component of a Database-Centered Decision Support System for Water and Power Management*. US Geological Survey, Denver.

Legates, D. R. & McCabe Jr, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **35** (1), 233–241.

Ley, R. 2014 Klassifikation von Pegel-Einzugsgebieten und Regionalisierung von Abfluss- und Modellparametern unter Berücksichtigung des Abflussverhaltens, hydroklimatischer und physiogeografischer Gebietsmerkmale. PhD Thesis. Universität Trier, Germany.

Ley, R., Casper, M. C., Hellebrand, H. & Merz, R. 2011 Catchment classification by runoff behaviour with self-organizing maps SOM. *Hydrol. Earth Syst. Sci.* **115**, 2947–2962.

McMillan, H. K., Tetzlaff, D., Clark, M. P. & Soulsby, C. 2012 Do time-variable tracers aid the evaluation of hydrological model structure? A multimodel approach. *Water Resour. Res.* **48**, doi: 10.1029/2011WR011688.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I: a discussion of principles. *J. Hydrol.* **10** (3), 282–290.

Pearson, K. 1895 Notes on regression and inheritance in the case of two parents. *Proc. R. Soc. London* **58**, 246–263.

Pushpalatha, R., Perrin, C., Moine, N. L. & Andreassian, V. 2012 A review of efficiency cirteria suitable for evaluating low-flow simulations. *J. Hydrol.* **420–421**, 171–182.

Sauquet, E. & Catalogne, C. 2011 Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France. *Hydrol. Earth Syst. Sci.* **15**, 2421–2435.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A. & Garillo, G. 2011 Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrol. Earth Syst. Sci.* **15**, 2895–2911.

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

Spearman, C. 1904 The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101.

Van Esse, W. R., Perrin, C., Booij, M. J., Augutsijen, D. C. M., Fenicia, F. & Lobligeois, F. 2013 The influence of conceptual model structure on model performance: a comparative study for 237 French catchments. *Hydrol. Earth Syst. Sci.* **17**, 4227–4239.

Vogel, R. M. & Fennessey, N. M. 1994 Flow-Duration Curves. I: new interpretation and confidence intervals. *J. Water Resour. Plann. Manage.* **120**, 485–504.

Wagener, T. & Montanari, A. 2011 Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resour. Res.* **47**, doi: 10.1029/2010WR009469.

Westerberg, I. K., Guerro, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S. & Xu, C. Y. 2011 Calibration of hydrological models using flow-duration curves. *Hydrol. Earth Syst. Sci.* **15**, 2205–2227.

Willmott, C. J. 1981 On the validation of models. *Phys. Geog.* **2**, 184–194.

Yilmaz, K. K., Gupta, H. V. & Wagener, T. 2008 A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. *Water Resour. Res.* **44**, W09417.

First received 11 December 2014; accepted in revised form 24 March 2015. Available online 12 May 2015