Identifying the connective strength between model parameters and performance criteria

Björn Guse[1,*], Matthias Pfannerstill[1], Abror Gafurov[2], Jens Kiesel[3,1], Christian Lehr[4,5], and Nicola Fohrer[1]

[1]Christian-Albrechts-University of Kiel, Institute of Natural Resource Conservation, Department of Hydrology and Water Resources Management, Kiel, Germany
[2]GFZ German Research Centre for Geosciences, Potsdam, Germany
[3]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany
[4]Leibniz Centre for Agricultural Landscape Research (ZALF), Institute of Landscape Hydrology, Muencheberg, Germany
[5]University of Potsdam, Institute for Earth and Environmental Sciences, Potsdam, Germany

Correspondence to: Björn Guse (*bguse@hydrology.uni-kiel.de)

Abstract.

In hydrological models, parameters are used to **represent the time-invarying characteristics** of catchments. Hereby, **model** parameters need to be identified based on their role in controlling the hydrological behaviour. For parameter identification, multiple and complementary performance criteria are used, which have to capture the different aspects of hydrological response of catchments. A reliable parameter identification depends on how distinctly a model parameter can be assigned to one of the performance criteria.

We introduce an analysis that reveals the connective strength between model parameters and performance criteria. The connective strength assesses the intensity in the interrelationship between model parameters and performance criteria. In our analysis of connective strength, model simulations are carried out based on a Latin Hypercube sampling. Ten performance criteria including NSE, KGE and its three components (alpha, beta and r) as well as RSR for different segments of the flow duration curve (FDC) are calculated.

With a joint analysis of two regression trees (RT), it is derived how a model parameter is connected to different performance criteria. At first, RTs are constructed using each performance **criterion** as target variable to detect the most relevant model parameters for each performance criteria. **Secondly,** RT approach using each parameter as target variable detects which performance criterion is impacted by changes in parameter values. Based on this, appropriate performance criteria are identified for each model parameter.

In this study, a high bijective connective strength **is found** for low and mid flow conditions. Moreover, the RT analyses emphasise the benefit of an individual analysis of the three components of KGE and of the FDC segments. **Furthermore, the RT analyses highlight** under which conditions these performance criteria provide insights into a precise parameter identification. **Our results show that** separate performance criteria are required to identify dominant parameters on low and mid flow conditions, whilst the number of required performance criteria for high flows increases with **increasing** process complexity in the

catchment. Overall, the analysis of the connective strength using RTs contribute towards a better handling of parameters and performance criteria in hydrological modelling.

1 Introduction

In **rainfall-runoff** models, hydrological processes are represented in a simplified way. Fluxes and changes in states are described by mathematical equations. To adapt the model to the hydrological conditions of the study catchments, multiple parameters are included in the model structure. Each of them has a specific role by representing one or multiple processes.

For reliable model simulations, it is required to identify parameter values that lead to a reasonable reproduction of their corresponding hydrological processes (Wagener et al., 2003; Pfannerstill et al., 2015). Typically, model parameters are **identified** using performance criteria which minimise the differences between measured and modelled discharge. In this context, we use 'performance criteria' as an overall term both for statistical performance metrics and signature measures. In parameter identification, it is implicitly assumed that model parameters are precisely identified by the selected set of performance criteria. To investigate the **validity** of this assumption, the interrelationship between model parameters and performance criteria needs to be identified as an initial step towards accurate parameter identification. **However, the selection of the most appropriate performance criteria for a precise parameter identification is still a challenge. Moreover, the number and type of performance criteria which are required to explain the hydrological behaviour in a study catchment is unclear and depends on catchment characteristics and its underlying process complexity (Wagener and Montanari, 2011; Pokhrel et al., 2012).**

To ensure a good parameter identification, it is nowadays commonly agreed that multiple and contrasting performance criteria **are required to determine whether parameters are only relevant specifically for a certain performance criterion** (Gupta et al., 1998; Vrugt et al., 2003; Krause et al., 2005; Gupta et al., 2009; Reusser et al., 2009; Guse et al., 2014). Each performance criterion emphasises different hydrological conditions with respect to e.g. discharge dynamics, discharge magnitude, water balance, high **or low** flows (Madsen, 2000; Boyle et al., 2001; Wagener et al., 2001). By selecting a specific performance criteria, a certain part of the hydrograph is inevitably weighted higher than other parts **and thus,** different parts of the hydrograph are emphasised or neglected **during parameter identification** (Gupta et al., 1998; Yapo et al., 1998; Pokhrel et al., 2012; Pechlivanidis et al., 2014; Pfannerstill et al., 2014b; Haas et al., 2016).

A challenging issue in hydrological modelling remains the understanding of the relationship between model parameters and performance criteria. For reliable identification of a model parameter, a performance **criterion** is appropriate **if it** is related to the part of the hydrograph that is controlled by the selected model parameter. It is assumed that a performance criterion contributes to a better interpretation of the hydrological behaviour if it is related directly to the **corresponding** components of the model structure (Yilmaz et al., 2008; Gupta et al., 2009; Martinez and Gupta, 2010; Pechlivanidis et al., 2014). It is thus intended to establish a strong relationship between a model parameter and a performance criterion which is appropriate for the associated process (Fenicia et al., 2007). An appropriate set of performance criteria should be selected **such** that all **relevant** hydrological conditions **in a catchment** are represented by at least one performance **criterion**.

Thus, to capture magnitude and dynamic in the modelled discharge time series, a combination of statistical performance metrics and signature measures in the model evaluation is recommended (van Werkhoven et al., 2008, 2009; Pfannerstill et al., 2014b). Typical statistical performance metrics are the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) and as a further development, the Kling-Gupta-Efficiency (KGE) (Gupta et al., 2009; Kling et al., 2012) which separately considers the three components bias (KGE_beta), variability (KGE_alpha) and correlation (KGE_r) to improve the estimation of the performance error compared to the NSE. **Signature measures are** directly related to catchment functions with the aim to consider the relevance of a certain **hydrological component** individually (Yilmaz et al., 2008; van Werkhoven et al., 2009; Pokhrel et al., 2012). Signature measures based on flow duration curves (FDC) provide diagnostic information of how a model performs for different discharge magnitudes (Yilmaz et al., 2008; Cheng et al., 2012; Yaeger et al., 2012; Pfannerstill et al., 2014b). Pfannerstill et al. (2014b) showed that a separation of the flow duration curve into five segments improved the model results for different discharge magnitudes and reduced the trade-off between satisfying results both for high and low flows in the same model run. By using different signature measures, the hydrologic behaviour is represented better in the performance assessment (Martinez and Gupta, 2011; Singh et al., 2011; Euser et al., 2013) and a precise interpretation **of which hydrological components** are accurately reproduced is realised.

The relevance of model parameters is site-specific depending on the prevailing dominant processes (Gupta et al., 2014; Guse et al., 2016). Thus, representative investigations of the relationship of model parameters and performance criteria requires analyses in contrasting catchments with differences in the dominant processes.

The relevance of model parameters for a performance criterion can be derived using sensitivity analyses. In addition of using model results directly for a sensitivity analysis (Reusser et al., 2011; Guse et al., 2014), performance criteria are another way to detect the most relevant model parameters. Several studies showed that the relevance of model parameters changes **if different performance criteria were used** (van Werkhoven et al., 2008; Abebe et al., 2010; Herman et al., 2013). Gupta et al. (2009) emphasised the need to investigate how changes in model parameter values influence the three components of the Kling-Gupta-Efficiency (KGE). **Some performance criteria are more appropriate in representing the impact of a model parameter to performance criteria than the others.**

To our knowledge, the relationship of model parameter performance criteria was analysed up to now only in one direction, namely from model parameters to model outputs and performance criteria respectively. Thus, the opposite direction, that is the suitability of a certain performance criterion to identify a certain model parameter remains unconsidered and is still a challenging task. Gupta et al. (2008) and Gupta et al. (2009) argues that a better understanding of how model parameters and performance criteria are interrelated is a core idea of diagnostic model analysis. For a precise parameter identification, the most relevant performance criteria for each model parameter needs to be derived.

To investigate the interrelationship between performance criteria and model parameters in a bi-directional manner, we built on the approach of Singh et al. (2014a) and Pechlivanidis and Arheimer (2015) who used Classification and Regression Trees (CART) to classify performance criteria with respect to appropriateness for different catchment characteristics. Both studies showed how different catchment characteristics and derived signatures resulted in a typical model performance. This idea can **be transferred** to the relationship between performance **criteria** and model parameters. To our knowledge, **this is the first attempt to analyse** how performance criteria are controlled by different model parameters using regression trees and how this relationship changes for different types of performance criteria. **To do so, we suggest the new concept of bijective connective strength between model parameters and performance criteria with the aim to obtain a more precise parameter identification**.

The connective strength assesses how strongly model parameters and performance criteria are interrelated using regression trees (RT). At first, it is investigated how the most influencing parameters vary for the selected performance criteria to analyse how strongly a set of model parameters affects different performance criteria. Second, looking from the side of the model parameters, it is analysed which performance criteria are impacted by changes in a certain model parameter. **In this way, performance criteria are detected which are** able to represent changes in a certain parameter. A high connective strength is given in the case that first a performance **criterion** is controlled by one model parameter and second this model parameter influences the same performance **criterion** to a relevant extent. This means that the model parameter has no relevant impact on other performance **criteria**. **With this study, we present a way to detect the appropriateness** of performance criteria as an initial step **towards** a precise identification of model parameters.

2 Methods and Materials

2.1 Study catchments

In contrasting catchments, different hydrological processes are of major relevance (Atkinson et al., 2002; Merz and Blöschl, 2004; Jothityangkoon and Sivapalan, 2009; Guse et al., 2016) **and thus, the ability** of a certain performance **criterion in identifying** a certain model parameter varies **as well**. With increasing relevance of a process, an accurate reproduction **in the model** becomes more important. **Threrefore,** two catchments with different catchment characteristics **were** selected **in this study to check the applicability of the proposed approach. For the analysis,** measured daily discharge time series from their catchment outlets **were** used **to assess model performance**.

2.1.1 Treene

The Treene catchment (up to the hydrological station Treia, 481 km^2), **located in the northern part of Germany is** a typical lowland catchment **with** strong groundwater **dominance to total discharge** even under high flow conditions (Guse et al., 2014; Pfannerstill et al., 2015; Guse et al., 2016). Moreover, tile flow is a relevant process **since large parts of its agriculture area are drained** (Kiesel et al., 2010). Other fast runoff components are of minor relevance as it is expected from the low topographic gradient in the catchment (**maximum elevation of** 80 meters). The Treene catchment is **dominated** by agricultural areas whilst only a minor part is covered by forests and urban areas (Guse et al., 2015). **Mean annual precipitation is about 995 mm/a with the highest values in summer months and monthly average temperature ranges from 1.5°C (January) to 17.6°C (July) with an average annual temperature being 9.2°C. Mean discharge at the catchment outlet is 6.23 m^3/s.**

2.1.2 Upper Saale

The Upper Saale catchment (hydrological station Blankenstein, about 1000 km^2) is located in the mid-range mountains **in the southeast** of Germany. This catchment is characterised by a higher diversity in dominant processes compared to the Treene catchment with **temporal** changes in the relevance of snowmelt, surface runoff as well as groundwater flow (Guse et al., 2016). The landscape is covered mostly by forests (upper parts) and **agricultural fields** (lower parts). **In contrary** to the Treene catchment, **the altitude in the Upper Saale catchment** is higher (between 415 and 856 meters) and slopes are steeper. Thus, fast runoff components are of higher relevance **in this catchment**. **Mean annual precipitation is 929 mm/a and monthly average temperature ranges from -1.7°C (January) to 17.2°C (July) with an annual mean temperature being 7.9°C. Mean discharge at the catchment outlet is 13.04 m^3/s.**

2.2 Soil and Water Assessment Tool (SWAT)

The conceptual and process-based eco-hydrological model SWAT (Soil and Water Assessment Tool, Arnold et al. (1998)) is used in this study. The SWAT model is spatially discretised into subbasins which are sub-divided into hydrological response units (HRUs) based on unique **underlying** information **on** landuse, soil and slope. **All water balance computations are conducted with daily temporal resolution at the scale of individual HRUs** as the central calculation unit. The change in soil water storage is **influenced** by inputs (e.g. precipitation) and outputs (e.g. evapotranspiration, runoff components).

In this study, the SWAT3S-version (Pfannerstill et al., 2014a) as a further development of the SWAT model was used. In SWAT3S, the groundwater modelling has been improved by subdividing the active aquifer contributing to river discharge into a fast and a slow responding one. **Different runoff components (surface runoff, lateral flow, groundwater flow) are separately computed for each HRU and summarised as water yield of a subbasin.** The SWAT model set-up for both catchments **is** described in Guse et al. (2016) and for a detailed description of the model set-up, **the readers are referred** to this study.

To analyse the relationship of performance criteria to model parameters, twelve SWAT model parameters from different hydrological components **which control different parts of the hydrograph** are selected in this study (Tab. 1). The final selection is based on studies with successful applications of the SWAT model within the studied catchments (Guse et al., 2016; Pfannerstill et al., 2015). **Thus, parameter ranges were constrained according to author's previous knowledge with the aim of reducing unrealistic parameter combinations which may lead to physically not plausible process description.**

Within the set of twelve parameters, two snow parameters regulate snowfall and snowmelt. Infiltration and surface runoff are captured by CN2 which is included in the curve number approach (SCS, 1972). The timing of different runoff components within the land phase is represented by lag time parameters (SURLAG, LATTIME, GDRAIN). Three soil parameters were included. For each soil layer, available water capacity (SOL_AWC) and saturated hydraulic conductivity (SOL_K) can be differentiated. The contribution of soil water from different soil depth for evaporation is regulated by a nonlinear function which is parameterised by ESCO. The groundwater module is parameterised

with a retention time from soil to groundwater (GW_DELAYfsh), a partitioning coefficient between the two aquifers (RCHRGssh) and the baseflow recession factor (ALPHA_BFssh).

[Table 1 about here.]

Model simulations **for the period from 2000 to 2010** were carried out based on 2000 different parameter sets that were generated with the Latin Hypercube sampling approach as it is implemented in the r-package FME (Soetaert and Petzoldt, 2010). In the Latin Hypercube sampling, all model parameters were changed simultaneously within the whole parameter space. For a more detailed description, **readers are referred to** Pfannerstill et al. (2014b).

2.3 Performance criteria

Ten performance criteria including five performance metrics and five signature measures were selected to capture different aspects of hydrological behaviour in models and as **it was** recommended in recent diagnostic model studies (Kling et al., 2012; Pechlivanidis et al., 2014; Pfannerstill et al., 2014b; Haas et al., 2016)

Nash-Sutcliffe Efficiency Criteria (NSE) (Eq. 1) is one of the most often used performance criteria in hydrology (Nash and Sutcliffe, 1970). NSE focuses on variability in measured discharge time series. It is known to give higher weights to high flows than to low flows (Schaefli and Gupta, 2007; Gupta et al., 2009; Pfannerstill et al., 2014b).

$$NSE = 1 - \frac{\sum\limits_{i=1}^{N} (Q_o - Q_s)^2}{\sum\limits_{i=1}^{N} (Q_o - \bar{Q_o})^2} \qquad (1)$$

where Q_o is measured discharge, Q_s modelled discharge and $\bar{Q_o}$ mean of measured discharge.

Kling-Gupta Efficiency (KGE) criteria (Gupta et al., 2009; Kling et al., 2012) is based on a decomposition of NSE into its three components (Eq. 2), which can be separately considered for each model run. Thus, model errors can be directly related to variability (KGE_alpha), bias (KGE_beta) and correlation (KGE_r) between measured and modelled discharge time series. KGE_alpha is the variability ratio between the standard deviation of **modelled (σ_s) and measured (σ_o)** discharge values. KGE_alpha larger than one shows that variability in modelled discharge time series is higher than in measured discharge time series, while **KGE_alpha lower** than one represents the opposite case. KGE_beta is the bias ratio between average values for **modelled (μ_s) and measured (μ_o)** discharge. KGE_beta larger than one represents an overestimation of discharge, i.e. a positive bias, while values lower than one illustrate an underestimation. KGE_beta and KGE_alpha represent the reproduction of the first and **the** second moments, respectively, as emphasised by Kling et al. (2012). KGE_r represents the correlation coefficient according to Pearson. **KGE_r is used to analyse** the agreement in temporal dynamics between measured and modelled discharge time series. To calculate KGE, the Euclidean distance to the ideal point in the three-dimensional criteria space which is created by its three components is calculated (Gupta et al., 2009). All three KGE components as well as KGE have an ideal value of one.

$$KGE = 1 - \sqrt{(KGE\_alpha - 1)^2 + (KGE\_beta - 1)^2 + (KGE\_r - 1)^2} \tag{2}$$

$$\begin{aligned} KGE\_alpha &= \sigma_s / \sigma_o \\ KGE\_beta &= \mu_s / \mu_o \\ KGE\_r &= \text{correlation coefficient} \end{aligned}$$

In addition to these five performance metrics, five signature measures are selected based on FDC. The FDC only considers the discharge magnitude without considering the temporal occurrence of discharge values (Vogel and Fennessey, 1996; Yilmaz et al., 2008; Westerberg et al., 2011). To evaluate the model performance, the FDC is subdivided into five FDC segments (very high (0-5% days of exceedance), high (5-20%), medium (20-70%), low (70-95%), very low (95-100%)) as proposed by Pfannerstill et al. (2014b) **and evaluated separately**. FDC signatures consider that different discharge magnitudes are controlled by different processes. Whilst high flow segment is mainly impacted by precipitation and fast runoff components, low flows are controlled by evapotranspiration and deep groundwater storages (Yilmaz et al., 2008; Cheng et al., 2012; Pokhrel et al., 2012; Yaeger et al., 2012; Guse et al., 2016)

For the evaluation of each FDC segment, the **RSR, the** ratio of the root mean square error to the standard deviation, was calculated for each FDC segment (Eq. 3) (Moriasi et al., 2007), **which** allows a fair comparison between different segments (Haas et al., 2016). The optimal value for RSR is zero. Using these five signature measures, it can be derived how model parameters are related to different discharge magnitudes (Pfannerstill et al., 2014b; Guse et al., 2016).

$$RSR = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_o - Q_s)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_o - \bar{Q_o})^2}} \tag{3}$$

These ten different performance criteria were calculated for all 2000 simulation runs. Both, parameter sets and calculated performance criteria from these simulations were then used for the following analyses.

To analyse the relationship among different performance criteria the correlation coefficients between all pairwise combinations were **computed**. This correlation analysis enables the detection of (dis)similarities between performance criteria. **(Dis)similarities** in performance criteria as indicated by a linear relationship in the dotty plots shows that these performance criteria **(do not)** capture a similar type of model error for this catchment. The intention **here is** to detect whether each performance criterion provides additional information of model error and whether **such hypothesis is valid for** both catchments **with different characteristics.**

2.4 Regression Trees

Regression Trees (RT) are a method to order the relationship between several explaining variables and a single target variable (Breiman et al., 1984). **It** is a binary algorithm **based** on logical expressions. In a sequence of regressions, **the explaining** variable is **subsequently** determined from a set of variables that has the highest predictive value for the target variable being analysed. **At each node of the regression trees, the (sub)set of model simulations** is subdivided into two subsets **based on a threshold value for one of the explaining variables** (Singh et al., 2014b, a). All simulations with a value in the explaining variable above the threshold belong to the one group, and those with a value below the threshold to the other group. For each node, this approach is repeated until no further subdivision of a variable at a certain node explains the target variable.

The sequence of decisions is visualised in a tree diagram **to detect** the importance of **different explaining** variables for the target variable. A regression tree consists of multiple branches. Either a different or an already chosen explaining variable is selected in the next branch of the tree. **The complexity of the tree reflects** the complexity of the relationship between explaining and target variables.

The earlier a variable is used in the construction of a RT, the higher is its importance. The variable used in the first split has thus the maximum importance. The gain in information is maximised by defining clearly separated subgroups of the whole simulation set (Singh et al., 2014b).

For our analyses, we used the R package rpart (Therneau and Atkinson, 2010). **In the rpart package**, the contribution of each explaining variable on changes in the target variable is calculated. **The variable importance describes how the prediction is reduced when removing this explaining variable. Thus, it summarises the contribution of the explaining variables in all nodes in explaining the changes in the target variable.**

Hereby, it is considered both that the explaining variable can be the most important one (and thus defines the subdivision of the subset) as well as that a variable has lower explanation power than the primary variable and is thus not shown in the trees. Due to that, also explaining variables that are not shown in the tree can have a relevant value of the variable importance. The percentage contribution of each explaining variable shows its importance for the target variable (Singh et al., 2014b).

2.4.1 Regression trees using performance criteria as target variable (RTperf)

RTs are applied in this study in two approaches using 2000 model simulations with pre-selected model parameters and calculated performance criterion. In a first application the **selected** ten performance criteria are used consecutively as target variable to construct regression trees for each performance criterion (named RTperf). As explaining variables, model parameters are used to detect which **of them** lead to changes in a performance criterion. The relevance of each model parameter is derived from regression trees by calculating the percentage contribution of each model parameter in explaining variability in a performance criterion. **This leads to identification of** the most relevant model parameters for each performance criterion.

2.4.2 Regression trees using model parameters as target variable (RTpar)

It is not only interesting to detect the most relevant parameters for a certain performance criterion. **Thus, in the second application, the importance of performance criteria which are** most strongly impacted by changes in the value of a certain parameter **is identified**. The latter point **cannot** be derived from RTperf. **To achieve this, a bijective approach is required by looking from the point of model parameters.**

Thus, **this** step was initialised vice-versa to RTperf to analyse how changes in model parameter influence performance criteria. To achieve this, explaining and target variables in RT are permuted. Each model parameter is used as target variable in RT and all performance criteria as explaining variables (named RTpar). **In RTpar model parameters are analysed individually.** Similarly to RTperf, the percentage contribution of each performance criterion is calculated to explain the impact of changes in values of a certain model parameter.

2.4.3 Connective strength by comparing both regression tree approaches

A core advantage of RT is that subsets of simulation runs are constructed in a structured way. By subdividing the simulation set based on the major influencing variables at each branch, two distinct subsets occur which differ with respect to values of model parameters as well as with respect to performance criteria. With this subset construction, it can be detected which model parameter or performance criteria, respectively, has the highest explanatory power in a RT branch.

In addition to RTperf and RTpar, the percentage contributions as derived from both RT approaches are compared to analyse the connective strength between model parameters and performance criteria. **Herewith** four cases of connective strength for each pair of model parameter and performance **criterion can be differentiated** (Fig. 1).

1. High percentage contributions in both RTs (RTperf, RTpar):
   Similar results of high percentage contributions in both RTs indicate a high bijective relationship between model parameter and performance criterion. In this case, the model parameter is clearly identifiable by the selected performance criterion. This is the optimal case representing a high connective strength and occurs if a certain parameter influences one performance criterion to a large extent without influencing other performance criteria significantly.

2. High percentage contribution in RTperf, but low in RTpar:
   In this case, a certain model parameter controls the selected performance criterion. However, this model parameter also influences other performance criteria. This case occurs if the corresponding **hydrological component** is very dominant and influences multiple performance criteria. Here, the connective strength cannot be fully understood when using performance criteria as target variable. From the side of the model parameter, it has **to be investigated further** which performance criterion is most appropriate for parameter identification.

3. Low percentage contribution in RTperf, but high in RTpar:
   In this case, the model parameter is not the major controlling parameter on the selected performance criterion as detected

in RTperf. However, its impact on other performance criteria is even lower which results in a high value in RTpar. Thus, the selected performance criterion is appropriate to explain the impact of changes in this model parameter, but the performance criterion is even stronger impacted by other model parameters. This case occurs if the corresponding **hydrological component** is of minor relevance in describing the hydrological system of the catchment. Thus, due to **its** low relevance, the connective strength is also low and a parameter identification is not **precise**.

4. Low percentage contributions in both RTs:
   In this case, neither this model parameter impacts the performance criterion to a relevant extent, nor the performance criterion is impacted by changes in the parameter. Thus, the connective strength is low. This parameter is not identifiable due to low relevance of the corresponding **hydrological component** and **no distinct** relationship to one of the performance criteria.

[Figure 1 about here.]

By applying this approach in two catchments with different characteristics, it is analysed how strongly a certain performance criterion is connected to a specific model parameter and how this connective strength depends on the relevance of the corresponding **hydrological component**.

3 Results

3.1 Correlation between performance criteria

In order to understand similarities of performance criteria, pairwise correlation analysis **of all** performance criteria is carried out separately for each catchment. In the Treene catchment (Fig. 2, upper panel), NSE, KGE and RSR of very high and high segments of FDC are strongly correlated. Moreover, RSR of low and very low flows are highly correlated. KGE is mainly controlled by its variability component (KGE_alpha) **meaning that** good performance of KGE_alpha (optimum=1) also results in high performance in KGE. KGE_beta (bias component) is correlated with mid segment of FDC. Concerning values of the performance criteria, KGE_alpha and KGE_beta are mostly higher than one, indicating an overestimation and higher variability in modelled discharge than in measured one. Good performance in a certain segment of FDC occurs in the case of good performance in the adjacent segment(s). In the case of good performance for low flows also very low flows perform well. Similarly, also good performance for very high flows was detected in model runs with good performance in high flows. However, correlations between RSR of (very) high and (very) low flows are lower which indicates that there are less model runs with a good performance in both high and low flows.

[Figure 2 about here.]

In the Saale catchment (Fig. 2, lower panel), correlations are overall lower. The strongest correlation is observed between NSE and KGE_r. KGE is correlated to KGE_alpha and KGE_r. Thus, both variability and correlation in modelled discharge time series are relevant for good performance of KGE. KGE_beta in contrast, which is balanced between over- and underestimation, is of lower relevance. The correlation among signatures of FDC segments is lower compared to the Treene catchment, even between adjacent segments. Here, a good performance of low flows does not result in good performance of very low flows. Worse performance of KGE_alpha (higher or lower than one) leads to a decrease in KGE since it increases the Euclidean distance of the three KGE components. However, as shown in Fig. 2, a different result was obtained between KGE_alpha and NSE. Lower variability in modelled than in measured discharge time series (KGE_alpha<1) results in an improvement of NSE. In contrast, KGE_alpha larger than one indicates that variability is higher in modelled discharge time series which leads to a reduction of NSE. This corresponds with the calculation of NSE which strongly emphasises variability in measured time series.

3.2 Impact of model parameters on performance criteria (RTperf)

The connective strength between model parameters and performance criteria was investigated using regression trees (RT). At first, RTs were constructed using the ten performance criteria (RTperf) as target variables. The aim of this step was to detect which model parameter most strongly affects a certain performance criterion.

Fig. 3 (above) shows the regression tree for KGE for the Treene catchment exemplarily for RTperf. Looking from top to down, the most influencing model parameters for KGE are provided. The first branch is defined by groundwater retention time of the first aquifer (GW_DELAYfsh) and the second one on the one right side again by GW_DELAYfsh and on the left side by the aquifer partitioning coefficient (RCHRGssh). In total, only groundwater parameters affect KGE to a relevant extent. When going along the branch at the right side, parameter settings of the controlling model parameter at these nodes are identified which lead to the best KGE on average (0.83).

[Figure 3 about here.]

To assess the connective strength between model parameters and performance criteria, the percentage contribution of model parameters as explaining variables for each performance criterion is shown for both catchments (Fig. 4). The parameter contribution in the Treene catchment to explain variability in performance criteria can be classified into three groups (Fig. 4, above). At first, six performance criteria are mainly influenced by GW_DELAYfsh and to a lower extent by RCHRGssh which shows the strong dominance of groundwater processes (see also Guse et al. (2014)). However, since multiple performance criteria are influenced by GW_DELAYfsh and RCHRGssh, the most appropriate performance criterion to identify the impact of these parameter is not detectable. The second group consists of KGE_beta and RSR of mid flow FDC segment. Both are controlled strongly by soil evaporation (ESCO) and available soil water capacity (SOL_AWC). Thirdly, low and very low flows are controlled in addition to GW_DELAYfsh by the baseflow recession coefficient of the second aquifer (ALPHA_BFssh). Seven of twelve model parameters namely fast runoff (CN2, SURLAG, GDRAIN, LATTIME), soil (SOL_K) and snow parameters (SFTMP, SMTMP) have only a minor impact on all performance criteria and cannot be identified by the selected performance criterion.

[Figure 4 about here.]

Fig. 4 (below) shows that the relationship between model parameters and performance criteria is more complex in Saale catchment compared to Treene. A clear classification into groups of performance criteria which are controlled by certain model parameters is more difficult. Four performance criteria (KGE_alpha, KGE_r, NSE, very high flow segment of FDC) are controlled by lateral flow lag time (LATTIME). But these performance criteria are also influenced by groundwater parameters (GW_DELAYfsh, RCHRGssh). Furthermore, RSR for high flows is not controlled by LATTIME but by these two groundwater parameters and hydraulic conductivity in soil (SOL_K). The KGE is controlled by a parameter (GW_DELAYfsh) which has not the largest percentage contribution for one of its three components. Water balance (KGE_beta) is controlled by ESCO and SOL_AWC, while mid flows are mainly influenced by SOL_AWC. Low flows are controlled by GW_DELAYfsh and very low flows by ALPHA_BFssh. Snow and fast runoff parameters except of LATTIME do not influence one of the performance criteria to a high extent. Thus, also in the Saale catchment, parameters exist without significant impact, while LATTIME controls multiple performance **criteria**.

By detecting the controlling model parameters for each performance criterion, in both catchments no appropriate performance criteria are found for several model parameters (e.g. CN2) which shows the low connective strength between these model parameters and performance criteria. **This leads to more challenging** identification of parameter values. It is of importance to detect whether the low relevance of a model parameter is related to a minor relevance of the corresponding process or whether the selected performance criterion is inappropriate to identify this model parameter. Moreover, some model parameters **highly influence** multiple performance criteria (e.g. GW_DELAYfsh and RCHRGssh in the Treene, LATTIME in the Saale), which leads to unclear results in the connective strength between model parameters and performance criteria. This suggests that these parameters govern the overall hydrological system in the model.

3.3 Impact of changes in model parameters on performance criteria (RTpar)

In the second RT **application**, the relationship between model parameters and performance criteria is analysed using twelve model parameters consecutively as target variables. It is investigated which performance criteria are impacted by changes in model parameters (RTpar, Fig. 5).

Fig. 3 (below) shows the regression tree examplarily for the model parameter GW_DELAYfsh for the RTpar approach in the Treene catchment. Here, KGE_alpha separates the data set at the first node and occurs once at the two following branches. Moreover contrasting performance criteria are included (KGE_r, RSR for very high flows and very low flows).

In the Treene catchment (Fig. 5, above), curve number (CN2) is most significantly related to RSR for very high flows and furthermore to NSE, KGE and RSR for high flows. This shows that CN2 and thus surface runoff controls **high** flow conditions. Snow parameters (SFTMP, SMTMP), the timing parameters for surface runoff (SURLAG) and tile flow (GDRAIN) as well as soil hydraulic conductivity (SOL_K) **highly influence** KGE_r (correlation). Thus, variations in these parameters lead to changes in correlation between measured and modelled discharge time series. Concerning the soil model component, SOL_AWC and ESCO are strongly related to water balance (KGE_beta) and to a lower extent to RSR for mid flows. ALPHA_BFssh is related to RSR of low and very low flows as well as to NSE. For the two groundwater parameters (GW_DELAYfsh, RCHRGssh), four performance criteria (NSE, KGE, KGE_alpha, RSR for high flow) have a similar percentage contribution. This point shows that both groundwater parameters control different aspects of hydrological model behaviour without having a clear relationship to a certain part of the hydrograph.

[Figure 5 about here.]

In the Saale catchment (Fig. 5, below), snow parameters (SFTMP, SMTMP) as well as LATTIME affect both KGE_r and NSE. Changes in curve number (CN2) mainly influence KGE_alpha and RSR for very high flows. Thus, variability between measured and modelled discharge time series and in particular high flows are influenced by CN2. All three soil parameters (SOL_AWC, SOL_K, ESCO) influence water balance (KGE_beta) and mid flow segment of FDC. However, evaporation (ESCO) is more related to KGE_beta while SOL_AWC has the largest impact on mid flows. In the case of GW_DELAYfsh several performance criteria are affected to a similar extent, but none of them has a high percentage contribution. While RCHRGssh affects KGE and high flows, ALPHA_BFssh strongly controls very low flows.

3.4 Comparing RTperf and RTpar

Subsequently, both RT approaches are compared by relating the percentage contribution from RTperf to RTpar and analysing these patterns for each performance criterion for both catchments (Fig. 6).

[Figure 6 about here.]

For mid and low flow conditions, both RTperf and RTpar provide a strong connective strength with high percentage contribution in RTperf and RTpar for the same pair of model parameter and performance criterion in both catchments. The strong relationship of evaporation (ESCO) and available soil water capacity (SOL_AWC) to RSR of mid flows and KGE_beta is derived in both RT approaches (Fig. 6). Water balance (KGE_beta) is hereby more controlled by ESCO, whilst SOL_AWC is the dominant parameter for mid flows especially in the Saale catchment.

Similarly, the connection between RSR for the very low segment of FDC and the baseflow recession coefficient (ALPHA_BFssh) is strong in particular in the Saale catchment. In both catchments, also the retention time of recharge into groundwater (GW_DELAYfsh) is of relevance.

KGE is dominated by GW_DELAYfsh and the aquifer partitioning coefficient (RCHRGssh) in a similar way in both catchments despite of contrasting catchment characteristics. In RTperf, KGE is most strongly impacted by GW_DELAYfsh. However, in RTpar, KGE has a higher percentage contribution in explaining changes in RCHRGssh than in GW_DELAYfsh.

In contrast, performance criteria related to high flows (NSE, very high segment) are controlled in the Treene catchment by groundwater (GW_DELAYfsh, RCHRGssh) and in the Saale catchment by lateral flow (LATTIME). This pattern shows that NSE focuses on model errors at high flows. A lower connective strength between model parameters and performance criteria was detected for high flow conditions. A bijective relationship between high flow related performance criteria and certain model parameters is more difficult to detect. The five performance criteria representing high flow conditions in the Treene catchment are related to the same two groundwater parameters (GWDELAYfsh and RCHRGssh). However, whilst GWDELAYfsh and RCHRGssh are the most dominant model parameters in RTperf, the percentage contribution in RTpar is lower. These two parameters dominate five performance criteria, but it remains unclear which is the best performance criterion in terms of parameter identification. Thus, while model errors in mid and low flows are identified **in both catchments** by the same performance criteria (Case 1, see Chapter 3.3.3), it is more complex to find appropriate performance criteria for errors in high flows. Here, a more complex hydrological behaviour is in particular detected in the Saale catchment as indicated by different controlling parameters on the performance criterion (Case 2 and 3 in Chapter 3.3.3). Moreover, the most dominant parameters in both catchments (GW_DELAYfsh in the Treene, LATTIME in the Saale) have a high percentage contribution in particular in RTperf both for high and low flows.

A very specific pattern is detected for KGE_r in the Treene catchment. In RTpar, a high percentage contribution of KGE_r for model parameters of lower relevance is detected. High values for RTpar and low values for RTperf in Fig. 6 show that KGE_r is the most appropriate performance criterion to assess changes in these model parameters. However, due to low relevance of snow or surface runoff, KGE_r is controlled by groundwater parameters. Here, we see a large difference in the interrelationship between model parameters and performance criteria by comparing RTperf and RTpar.

4 Discussion

The aim of this study was to improve the understanding of the relationship between model parameters and performance criteria. For this, the connective strength between model parameters and performance criteria was **introduced, which is based on two approaches of regression trees first using performance criteria and then model parameters as target variables**. **Based on this, it is discussed how the connective strength to the performance criteria varies between different model parameters and how the use of different performance criteria help in identifying model parameters.**

4.1 Benefit of analysing bijectively the relationship between model parameters and performance criteria

By analysing the connective strength, the performance criteria are identified which are appropriate to constrain at best the model parameters. The novelty of this study lies in the assessment of the relationship between model parameters and performance criteria bijectively (RTperf, RTpar).

In RTperf, it is detected for ten performance criteria whether the same model parameters affect different performance criteria. It is shown that not all model parameters influence one of the selected performance criteria (see Fig. 4). **This indicates that not all model parameters can be identified with this approach since either the model parameters are not relevant or that performance criteria are still missing to describe the changes in this model parameter.**

The impact of performance criteria depends on the relevance of the corresponding process. In the case that the relevance of the associated process is very low, parameters from other more dominant processes control performance criteria. This is for example shown for curve number (CN2). Its impact on performance criteria in Treene catchment is low due to higher contribution of groundwater flow compared to surface runoff.

In RTpar with model parameters as target variables, each model parameter is individually assessed. By comparing RTpar **for** different model parameters, **the influence of model parameters on** different performance criteria **is identified**. It can be derived whether parameters and their associated processes are of low relevance or whether an appropriate performance criterion for a model parameter is missing. The RTpar approach shows for the majority of the model parameters that changes in their values are detectable at least by one of selected performance criteria. **This indicates that** the impact of parameters from processes of minor relevance on performance criteria can be derived **with** RTpar.

Differences between RTperf and RTpar are obtained for parameters related to the most dominant process(es). The groundwater parameters (mainly GW_DELAYfsh) control most of performance criteria for the Treene catchment (Fig. 4). A similar result is obtained for the Saale catchment with a dominance of lateral flow lag time (LATTIME). Comparing results of two RT approaches, a higher similarity between both catchments is detected in RTpar (Fig. 5).

The interpretation of the relationship between model parameters and performance criteria from both sites by means of the suggested connective strength extends the classical one-sided analysis of the impact of model parameters on performance criteria as it is done e.g. in sensitivity analysis (van Werkhoven et al., 2008; Herman et al., 2013). In our approach both, performance criteria and model parameters, were analysed separately as target variables. In comparison to the established one-sided approaches, this yields additional information on which performance criteria are appropriate for a certain model parameter.

Thus, we do not only investigate how variations are propagated in the model up to the output but also which outputs (i.e. performance criteria) are impacted by a certain model parameter. The comparison of parameter relevances with former studies on temporally resolved parameter sensitivity analyses (Guse et al., 2014, 2016) shows that the overall ranking of model parameters is similar.

4.2 Benefit of using different performance criteria

Furthermore, it was analysed how the use of different performance criteria help in identifying model parameters. The differences in the relevance of model parameters on ten performance criteria emphasised the benefit of **using** this set of performance criteria. The separate consideration of KGE components demonstrates that different parameters are related to these three performance metrics of KGE. While relevant parameters on KGE and KGE_alpha are similar in the Treene catchment, the most relevant parameter on KGE in the Saale catchment (GW_DELAY) is not the relevant one for the three KGE components. Since each KGE component can be clearly related to a specific part of hydrological behaviour (Kling et al., 2012), the **RT** shows whether a model parameter is more relevant in representing variability, correlation or bias in modelled discharge time series. By comparing **the performance in KGE with** its components, the most important aspect **in evaluating the model performance as captured in the KGE components (variability, bias, correlation)** becomes apparent such as variability as detected by KGE_alpha in the Treene catchment.

The differences in pairwise correlations of performance criteria between both catchments also results in differences in controlling model parameters on a certain performance criterion. Similar results in KGE and NSE are calculated in the case of high correlation between KGE and KGE_alpha and thus the most relevant model parameters on these two performance criteria are similar. The opposite result is obtained in the Saale catchment. Due to low values of KGE_alpha, different parameters control KGE and NSE. **This pattern is reasonable since both, KGE_alpha and NSE, focus on assessment of variability in discharge time series, while the three components (variability, bias, correlation) are equally weighted in KGE.**

Concerning signature measures, this study shows that different parameters are related to FDC segments. This is in line with studies stating that each FDC segment can be related to certain catchment processes (Yilmaz et al., 2008; Yaeger et al., 2012; Pfannerstill et al., 2015). The strong connective strength of model parameters regulating water balance to mid flows as well as of parameters from slow reacting aquifer storages to very low flows is derived in this study. However, a typical sequence of a high connective strength of high flows to surface runoff parameters is not identified. Moreover, we **observe** that dominant processes, i.e. groundwater flow in the Treene and lateral flow in the Saale catchment, influence both high and low flows. This leads to a trade-off in parameter identification since the same model parameters control high and low flows. **This highlights that performance criteria should be specific for different model parameters, in this case specific for either high or low flow, to avoid a high parameter uncertainty and equifinality in the estimation of behavioural parameter sets.**

4.3 Number of required performance criteria

The analysis of the bijective connective strength between pairs of model parameter and performance criteria in the two catchments shows that the most appropriate performance criterion varies depending on different model parameters. Pairs with a high connective strength were detected and grouped. This results in a minimum number of required performance criteria of at least three performance criteria related to high, mid and low flow conditions since the most relevant parameters between these three types of performance criteria vary. This is in line with other studies on performance criteria stating that 3-4 performance criteria capturing different parts of the hydrological system are a minimum number (Madsen, 2000; Boyle et al., 2001; van Werkhoven et al., 2008, 2009).

In addition, an individual performance criterion for a single model parameter might be needed, e.g. to assess the importance of mid flows (van Werkhoven et al., 2008; Wagener et al., 2009; Herman et al., 2013). **This is shown in** the RT analysis, **where the** controlling parameters for KGE_beta and RSR for mid flows are different compared to other performance criteria and these dominant model parameters are from soil components and related to water balance. In these cases, the connective strength is very high.

Also for an assessment of low flow conditions an individual performance criterion is needed which was in this case RSR for low and very low flows. A high connective strength between model parameter and performance criterion was detected for very low flows. Moreover, the requirement for a segmentation of FDC into very low and low flows as introduced by Pfannerstill et al. (2014b) is emphasised by identifying different relevant model parameters. The high correlation between RSR for very low and low flows in the Treene catchment also results in similar dominant parameters, while different parameters control these signature measures in the Saale catchment (see Fig. 4). Thus, similar influencing parameters in RTperf for different performance criteria are detected if both are highly correlated.

High flows are driven by interacting and overlaying processes from different hydrological components. Here, the most influential parameters and the most appropriate performance criterion vary depending on the type of errors which are dominant in the modeling process. The complexity in the representation of high flows depends on involved processes. In the groundwater dominated Treene catchment, the RT analysis for five performance criteria related to high flows provides very similar results. The RT analysis using model parameters as explaining variables (RTpar) however highlights differences in the relationship of model parameters and these five performance criteria. In the Saale catchment, relevant model parameters largely vary between all performance criteria related to high flows. Here, all selected performance criteria capture different types of errors in modelled discharge time series. The analysis of deviations between measured and modelled discharge in this catchment is more complex so that more than one performance criterion for high flows is required. With higher heterogeneity in dominant processes and strong interaction of different processes in controlling hydrological behaviour, a more distinct selection of a larger set of performance measures is required.

Thus, it can be recommended to include several performance criteria to capture all types of potential errors both in dynamic and magnitude of modelled discharge. In addition, it is relevant to consider which model parameters dominate performance criteria. This can help to understand why a certain model error might occur and to which processes this model error is related.

As demonstrated in this study, the results vary between different catchments. Further studies in other catchments might improve additionally the understanding of the connective strength between model parameters and performance **criteria using this methodology. A separate approach for specific time periods, e.g. for winter time to capture the impact of snow parameters might be an upcoming issue.** Due to the generality of the suggested approach, an applicability with other models is expected.

5 Conclusion

For achieving a precise parameter identification, the connective strength between model parameters and different performance criteria is analysed. For this, two regression tree (RT) approaches are applied using consecutively performance criteria and model parameters as target variables. This method derives which model parameters affect a performance criterion (RTperf) and which performance criteria are impacted by changes in model parameters (RTpar). **By detecting the connective strength between model parameters and performance criteria, appropriate performance criteria can be derived for different model parameters. Based on a precise parameter identification and a better understanding of model parameters, parameter uncertainty and thus equifinality among different model runs can be reduced.**

Thus, the main outcomes **of this study** are:

1. The pairwise correlation between performance criteria varies between the two catchments depending on the model error. Thus, different performance criteria are required to disentangle the impact of different hydrological behaviour on modelled discharge. The number of required performance criteria is higher for catchments with a higher process complexity.

2. In RTperf, it becomes apparent how largely the relevance of model parameters varies between different performance criteria. Our study emphasises the importance of a separate consideration of KGE components and of a signature-

based analysis of different FDC segments for a precise parameter identification. Differences in dominant parameters are detected between performance criteria related to high, mid or low flow conditions.

3. RTpar, **which uses** model parameters as target variables, shows which performance criterion is appropriate to identify a model parameter. Similar results in RTperf and RTpar demonstrate high capability of a performance criterion to consider the impact of a model parameter accurately. Contrasting results are in particular derived for model parameters which are related to processes of minor relevance. A bijective connective strength between model parameters and performance criteria is detected for low and mid flows, whilst modelling of high flows is more complex both in terms of relevant model parameters and appropriate performance criteria.

Overall, this study shows that multiple performance criteria are required for an accurate parameter identification for reliable hydrological modelling. **However, no general conclusion regarding universal performance criteria can be drawn, since** the connective strength between model parameters and performance criteria varies between catchments depending on hydrological complexity of the catchments with respect to processes and their relevance in controlling hydrological behaviour in models. Using **the presented** approach, it can be derived how precisely model parameters can be identified by a set of performance criteria.

Acknowledgements. We thank the DFG for financial support of the first author (project GU 1466/1-1 Hydrological consistency in modelling). Furthermore we thank the CAWa (Central Asian Water) project (www.cawa-project.net, Contract No. AA7090002), by the German Federal Foreign Office as part of the German Water Initiative for Central Asia ("Berlin Process") for the funding of the third author (AG). The fourth author (JK) was funded through the "GLANCE" project (Global change effects in river ecosystems; 01LN1320A) supported by the German Federal Ministry of Education and Research (BMBF). The fifth author (CL) received funding by the Leibniz Association (SAW-2012-IGB-4167) within the international Leibniz graduate school: Aquatic boundaries and linkages in a changing environment (Aqualink) (http://www.igb-berlin.de/aqualink.html).
We thank Schleswig-Holstein Agency for Coastal Defence, National Park and Marine Conservation of Schleswig-Holstein (LKN-SH) and the State Institute for Environment and Geology of Thuringia (TLUG) for the discharge data. We thank Martin Volk and Michael Strauch (UFZ) for contributing to the SWAT modelling in the Saale catchment. Furthermore, we thank the community of the open source software R, which we used for this study.

References

N.A. Abebe, F.L. Ogden, and N.R. Pradhan. Sensitivity and uncertainty analysis of the conceptual HBV rainfall-runoff model: Implications for parameter estimation. *J. Hydrol.*, 389:301–310, 2010.

J.G. Arnold, R. Srinivasan, R.S. Muttiah, and J.R. Williams. Large area hydrologic modeling and assessment part I: model development. *J. Am. Water Res. A.*, 34:73–89, 1998.

S. E. Atkinson, R. A. Woods, and M. Sivapalan. Climate and landscape controls on water balance model complexity over changing timescales. *Water Resour. Res.*, 38(12):doi:10.1029/2002WR001487, 2002.

D. P. Boyle, H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith. Toward improved streamflow forecasts: Value of semidistributed modeling. *Water Resour. Res.*, 37(11):2749–2759, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. CRC Press, Wadsworth, Belmont, CA., 1984.

L. Cheng, M. Yaeger, A. Viglione, E. Coopersmith, S. Ye, and M. Sivapalan. Exploring the physical controls of regional patterns of flow duration curves - part 1: Insights from statistical analyses. *Hydrol. Earth Syst. Sci.*, 16:4435–4446, 2012.

T. Euser, H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. Savenije. A framework to assess the realism of model structures using hydrological signatures. *Hydrol. Earth Syst. Sci.*, 17:1893–1912, 2013.

F. Fenicia, H. H. Savenije, P. Matgen, and L. Pfister. A comparison of alternative multiobjective calibration strategies for hydrological modeling. *Water Resour. Res.*, 43:W03434, doi: 10.1029/2006WR005098, 2007.

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

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

H.V. Gupta, S. Sorooshian, and P.O. Yapo. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resour. Res.*, 34(4):751–763, 1998.

H. V. Gupta, C. Perrin, G. Bloeschl, A. Montanari, R. Kumar, M. Clark, and V. Andreassian. Large-sample hydrology: a need to balance depth with breadth. *Hydrol. Earth Syst. Sci.*, 18:463–477, 2014.

B. Guse, D. E. Reusser, and N. Fohrer. How to improve the representation of hydrological processes in SWAT for a lowland catchment - temporal analysis of parameter sensitivity and model performance. *Hydrol. Process.*, 28:2651–2670. doi: 10.1002/hyp.977, 2014.

B. Guse, M. Pfannerstill, and N. Fohrer. Dynamic modelling of land use change impacts on nitrate loads in rivers. *Environ. Process.*, 2(4): 575–592. doi:10.1007/s40710-015-0099-x, 2015.

B. Guse, M. Pfannerstill, M. Strauch, D. E. Reusser, M. Volk, H. V. Gupta, and N. Fohrer. On characterizing the temporal dominance patterns of model parameters and processes. *Hydrol. Process.*, 30(13):2255–2270, doi:10.1002/hyp.10764, 2016b.

M.B. Haas, B. Guse, M. Pfannerstill, and N. Fohrer. A joined multi-metric calibration of river discharge and nitrate loads with different performance measures. *J. Hydrol.*, 536:534–545, doi:10.1016/j.jhydrol.2016.03.001, 2016.

J. D. Herman, P. M. Reed, and T. Wagener. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resour. Res.*, 49:doi:10.1002/wrcr.20124, 2013.

C. Jothityangkoon and M. Sivapalan. Framework for exploration of climatic and landscape controls on catchment water balance, with emphasis on inter-annual variability. *J. Hydrol.*, 371:154–168, 2009.

J. Kiesel, N. Fohrer, B. Schmalz, and M. J. White. Incorporating landscape depressions and tile drainages of lowland catchments into spatially distributed hydrologic modeling. *Hydrol. Process.*, 24:1472–1486, 2010.

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

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

H. Madsen. Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *J. Hydrol.*, 235:276–288, 2000.

H. Madsen, G. Wilson, and H. C. Ammentorp. Comparison of different automated strategies for calibration of rainfall-runoff models. *J. Hydrol.*, 261:48–59, 2002.

G. F. Martinez and H. V. Gupta. Toward improved identification of hydrological models: A diagnostic evaluation of the "abcd" monthly water balance model for the conterminous united states. *Water Resour. Res.*, 46:W08507, doi:10.1029WR008294, 2010.

G. F. Martinez and H. V. Gupta. Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental united states. *Water Resour. Res.*, 47:W12540, doi:10.1029/2011WR011229, 2011.

R. Merz and G. Blöschl. Regionalisation of catchment model parameters. *Journal of Hydrology*, 287(1-4):95–123, 2004.

D. N. Moriasi, J.R. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900, 2007.

J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models: part i - a discussion of principles. *J. Hydrol.*, 10:282–290, 1970.

S.L. Neitsch, J.G. Arnold, J.R. Kiniry, and J.R. Williams. Soil and water assessment tool - Theoretical documentation version 2009. *Texas Water Resources Institute Technical Report*, 406, 2011.

I. G. Pechlivanidis and B. Arheimer. Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case. *Hydrology and Earth System Sciences*, 19:4559–4579, 2015. doi:10.5194/hess-19-4559-2015.

I. G. Pechlivanidis, B. Jackson, H. McMillan, and H. Gupta. Use of entropy-based metric in multiobjective calibration to improve model performance. *Water Resour. Res.*, 50:8066–8083, 2014. doi:10.1002/2013WR014537.

M. Pfannerstill, B. Guse, and N. Fohrer. A multi-storage groundwater concept for the swat model to emphasize nonlinear groundwater dynamics in lowland catchments. *Hydrol. Process.*, 28:5599–5612, doi:10.1002/hyp.10062, 2014a.

M. Pfannerstill, B. Guse, and N. Fohrer. Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *J. Hydrol.*, 510:447–458, doi:10.1016/j.jhydrol.2013.12.044., 2014b.

M. Pfannerstill, B. Guse, D. Reusser, and N. Fohrer. Process verification of a hydrological model using a temporal parameter sensitivity analysis. *Hydrol. Earth Syst. Sci.*, 19:4365–4376, doi:10.5194/hess-19-4365-2015, 2015.

P. Pokhrel, K. K. Yilmaz, and H. V. Gupta. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *J. Hydrol.*, 418-419:49–60, 2012.

D.E. Reusser, T. Blume, B. Schaefli, and E. Zehe. Analysing the temporal dynamics of model performance for hydrological models. *Hydrol. Earth Syst. Sci.*, 13:999–1018, 2009.

Reusser, D. E., Buytaert, W., Zehe, E., 2011. Temporal dynamics of model parameter sensitivity for computationally expensive models with FAST (Fourier Amplitude Sensitivity Test). Water Resour. Res. 47(7), doi:10.1029/2010WR009947.

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

SCS, 1972. Hydrology. Soil Conservation Service, Ch. Section 4.

R. Singh, T. Wagener, K. van Werkhoven, M. E. Mann, and R. Crane. A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate accounting for changing watershed behavior. *Hydrology and Earth System Sciences*, 15(11): 3591–3603, 2011.

R. Singh, S. A. Archfield, and T. Wagener. Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments - a comparative hydrology approach. *J. Hydrol.*, 517:985–996, 2014a.

R. Singh, T. Wagener, R. Crane, M. E. Mann, and L. Ning. A vulnerability driven approach to identify adverse climate and land use change combination for critical hydrologic indicator thresholds: Application to a watershed in Pennsylvania, USA. *Water Resour. Res.*, 1–19, 2014b. doi:10.1002/2013WR014988.

K. Soetaert and T. Petzoldt. Inverse modelling, senstivity and monte carlo analysis in r using package FME. *J. Stat. Softw.*, 33(3):1–28, 2010. doi:10.1.1.160.5179.

T.M. Therneau and B. Atkinson. Rpart: Recursive partitioning. *R package*, 2010. URL http://CRAN.R-project.org/package=rpart.

K. van Werkhoven, T. Wagener, P. Reed, and Y. Tang. Rainfall characteristics define the value of streamflow observations for distributed watershed model identification. *Geophys. Res. Lett.*, 35:L11403, doi:10.1029/2008GL034162., 2008.

K. van Werkhoven, T. Wagener, P. Reed, and Y. Tang. Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. *Adv. Water Resour.*, 32(8):1154–1169, 2009. doi:10.1016/j.advwatres.2009.03.002.

R. M. Vogel and N. M. Fennessey. Flow-duration curves II: A review of applications in water resources planning. *Water Resources Bulletin*, 31(6):1029–1039, 1996.

J. Vrugt, H.V. Gupta, L.A. Bastidas, W. Bouten, and S. Sorooshian. Effective and efficient algorithm for multiobjective optimization of hydrological models. *Water Resour. Res.*, 39:1214–1232, 2003.

T. Wagener and A. Montanari. Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resour. Res.*, 47:W06301, doi:10.1029/2010WR009469, 2011.

T. Wagener, D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian. A framework for development and application of hydrological models. *Hydrol. Earth Syst. Sci.*, 5(1):13–26, 2001.

T. Wagener, N. McIntyre, M.J. Lees, H.S. Wheater, and H.V. Gupta. Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrol. Process.*, 17:455–476, 2003.

T. Wagener, K. van Werkhoven, P. Reed, and Y. Tang. Multiobjective sensitivity analysis to understand the information content in streamflow observations for distributed watershed modeling. *Water Resour. Res.*, 45:W02501, doi:10.1029/2008WR007347, 2009.

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

M. Yaeger, E. Coopersmith, S. Ye, L. Cheng, A. Viglione, and M. Sivapalan. Exploring the physical controls of regional patterns of flow duration curves - part 4: A synthesis of empirical analysis, process modeling and catchment classification. *Hydrol. Earth Syst. Sci.*, 16: 4483–4498, 2012.

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

K. K. Yilmaz, H.V. Gupta, and T. Wagener. 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.

List of Figures

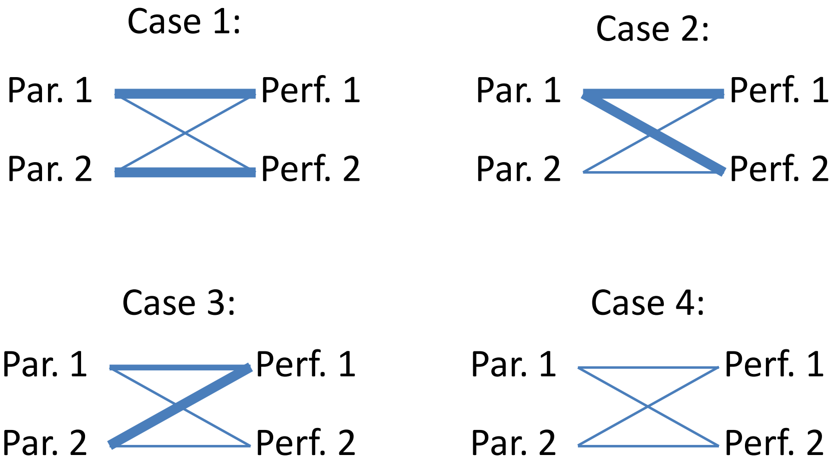


Figure 1. Flowchart of four cases of connective strength between model parameters and performance measures. A **thicker** blue line shows a higher impact of the model parameter on the performance criteria

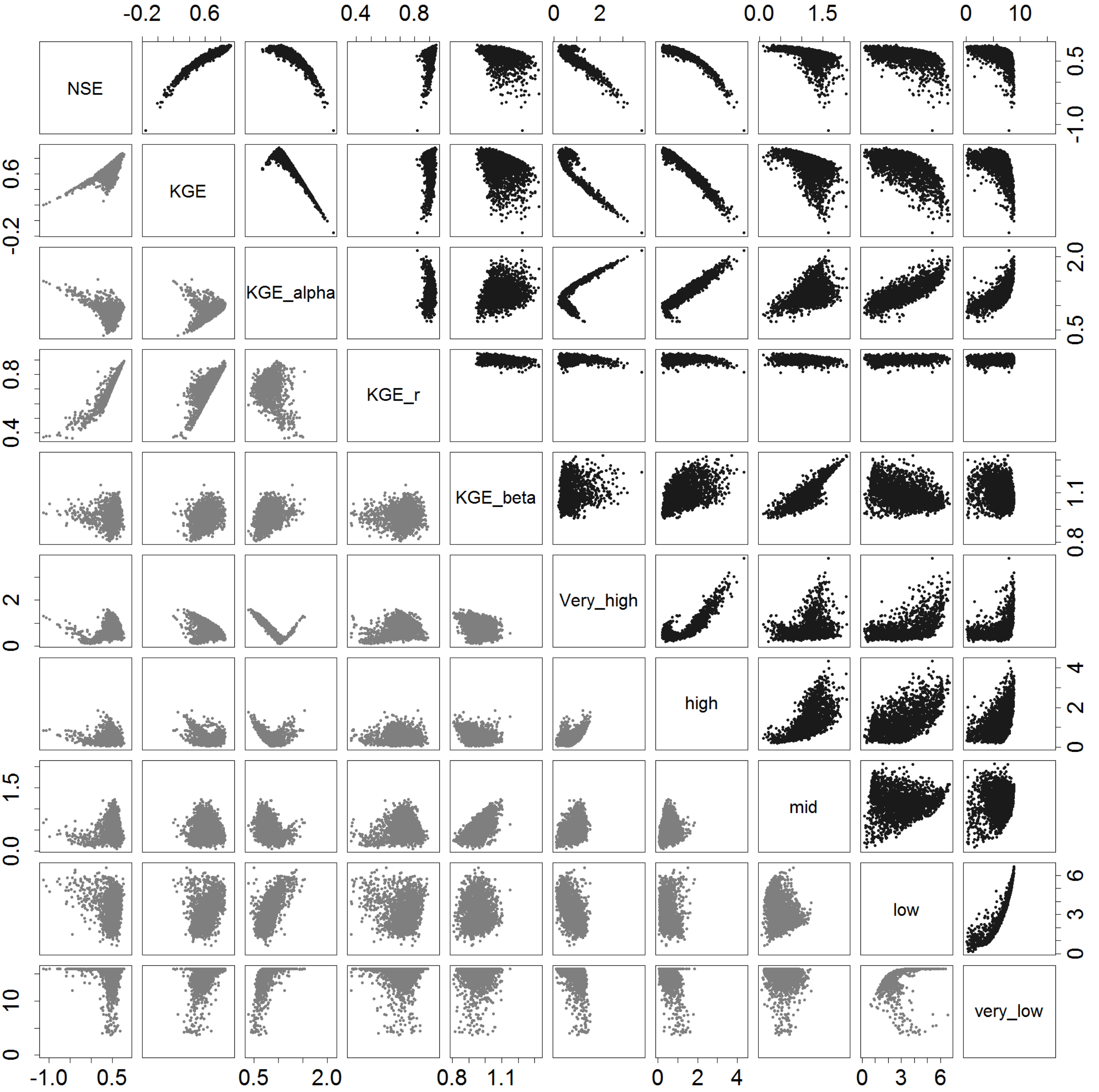


Figure 2. Scatterplot matrix of performance criteria of Treene (in black, upper panel) and Saale (in grey, lower panel) catchment showing pairwise performance criteria plots for 2000 model simulations. The scales on the sides show values of the respective performance criterion.

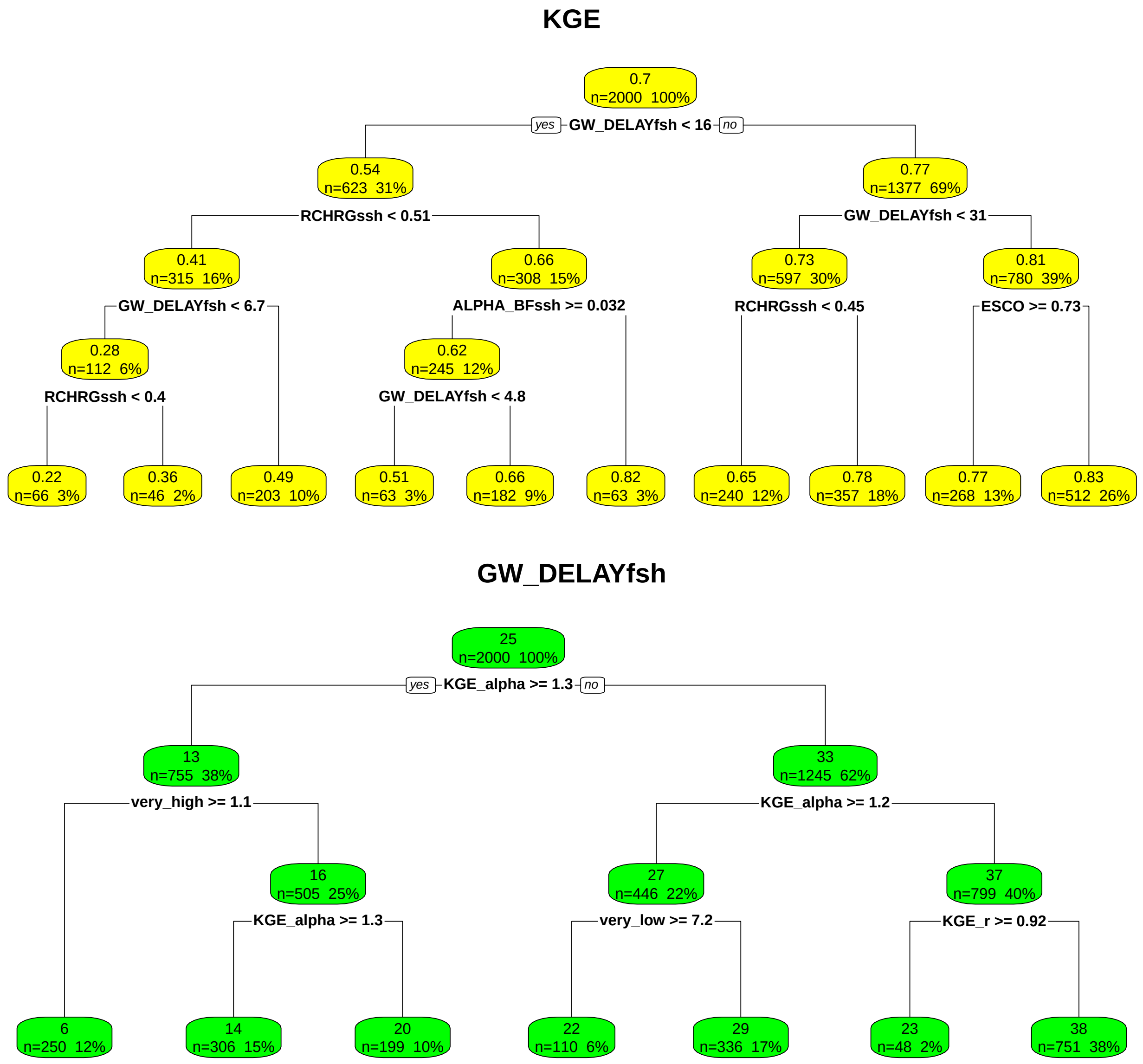


Figure 3. Example of a regression tree (RT) using (above) KGE as target variable and model parameters as explaining variables and (below) the model parameter GW_DELAYfsh as target variable and performance measures as explaining variables for the Treene catchment.

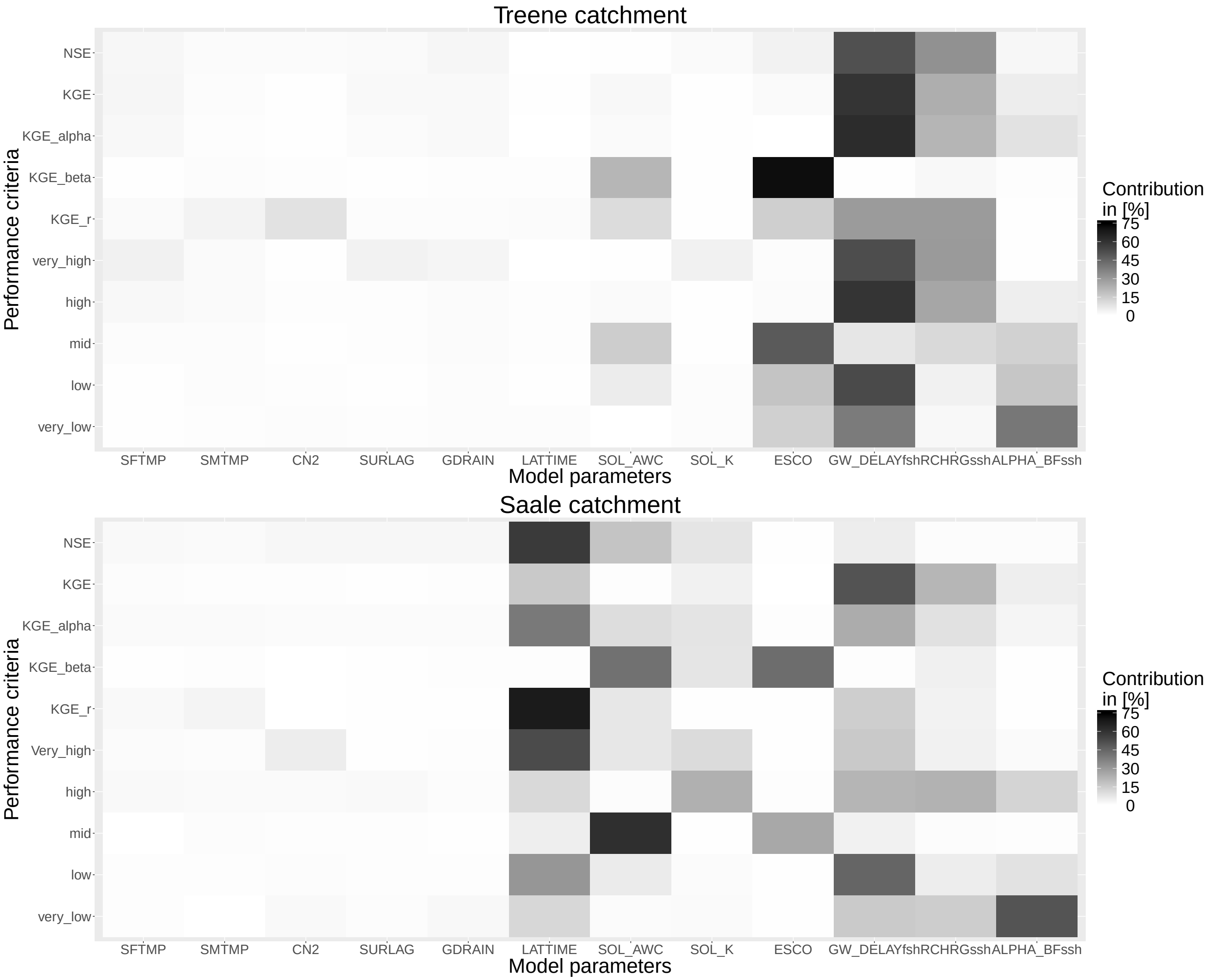
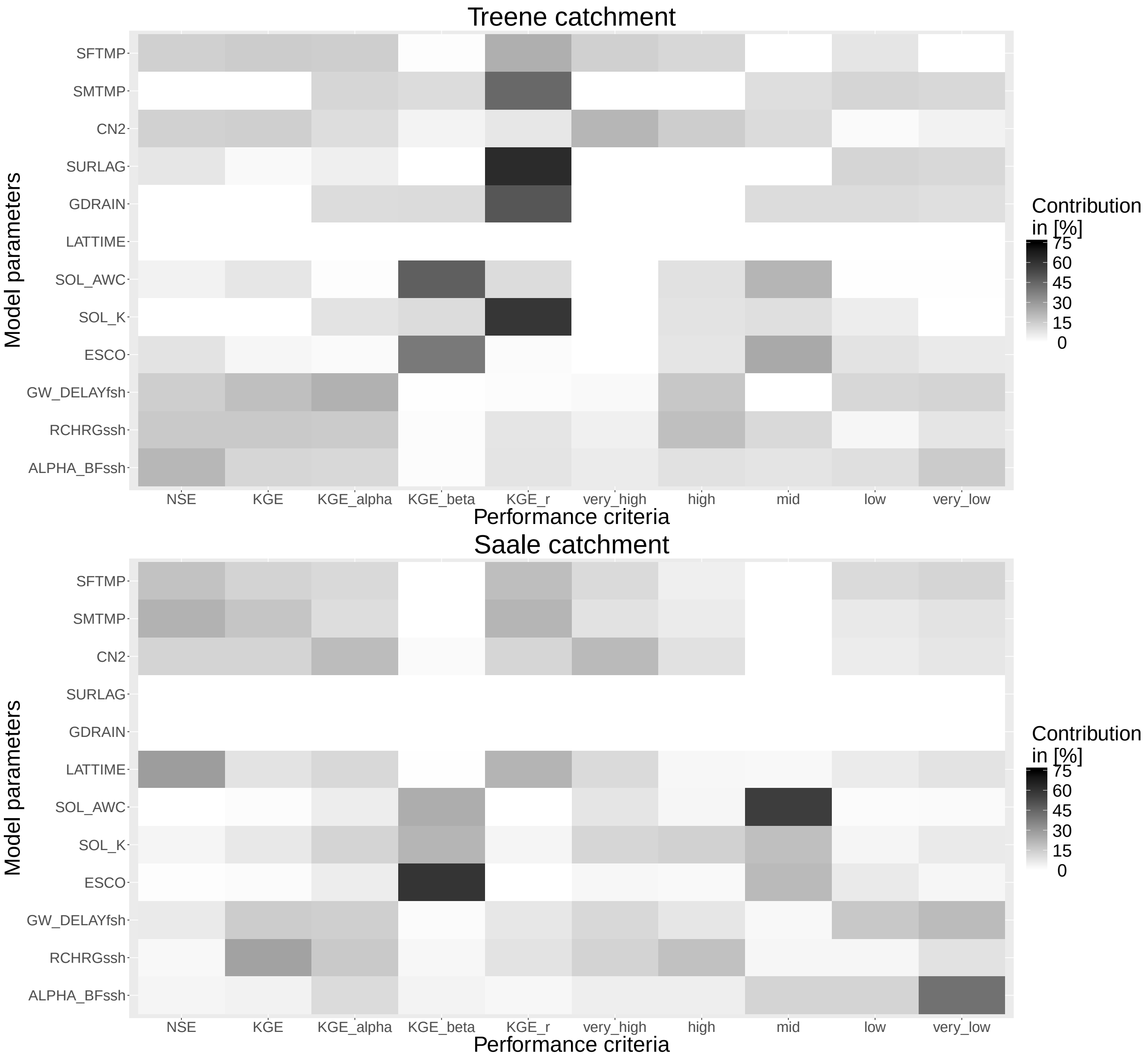


Figure 4. Regression trees (RT) using performance criteria as target variables. The percentage contribution of model parameters in explaining performance criteria is shown for Treene (above) and Saale (below) catchment. In every row the percentage contributions sum up to 100%.



Treene catchment

Saale catchment

Figure 5. Regression trees (RT) using model parameters as target variables (RTpar) for Treene (above) and Saale (below) catchment and performance criteria as explaining variables. All values of a parameter are in white in the case that the resulting variation among performance criteria for this parameter was too low to construct a regression tree. In every row the percentage contributions sum up to 100%.

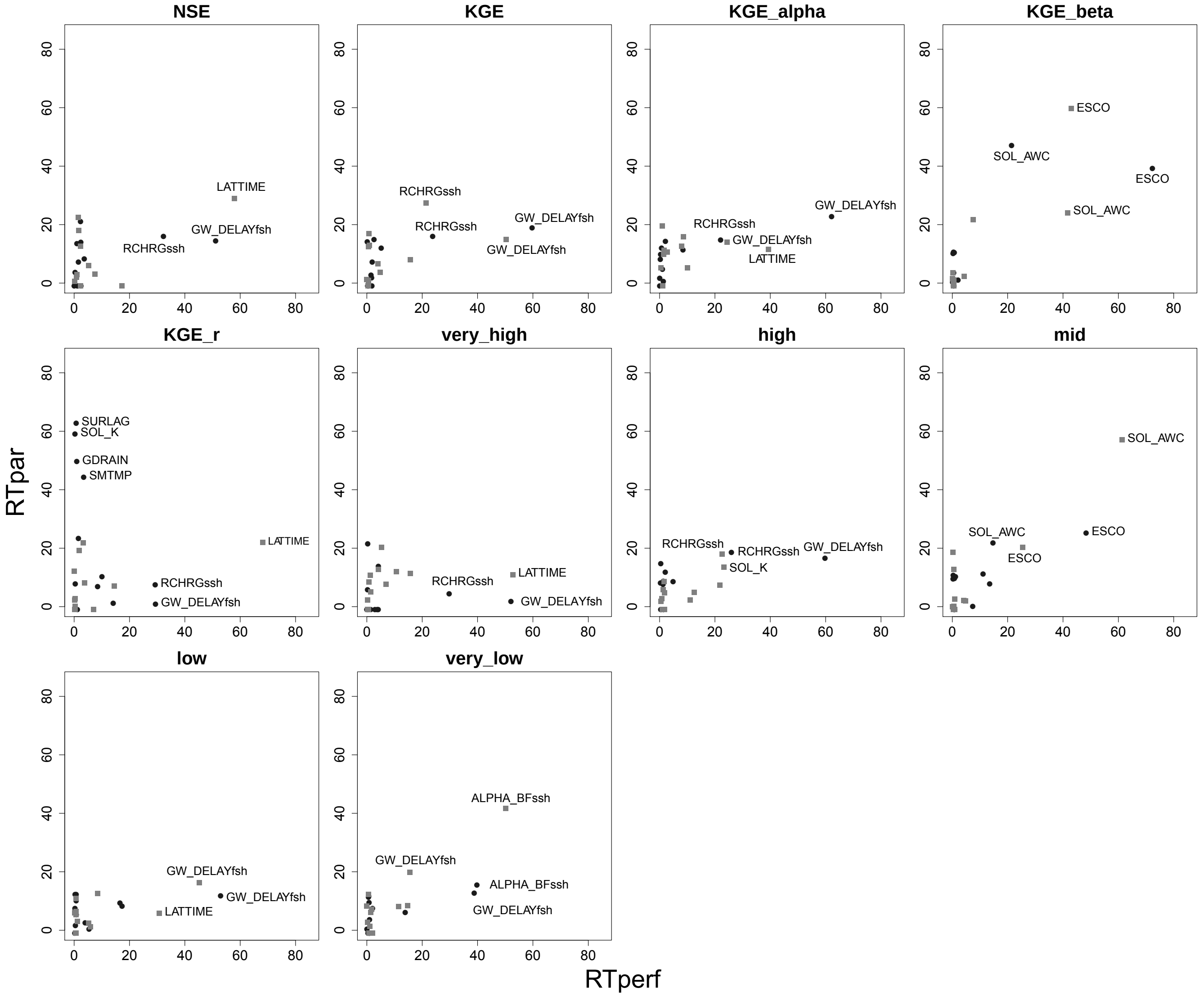


Figure 6. Connective strength between performance criteria and model parameters. The percent contribution of pairs of model parameter and performance criterion are shown as derived from RTperf (x-axis) and RTpar (y-axis). A high value along the x-axis shows a high contribution of a model parameter in explaining variability in the performance criterion as detected by RTperf. A high value along the y-axis (RTpar) shows that this performance criterion among all performance criteria is most strongly impacted by changes in the model parameter. A strong connective strength is detected if both values are high. The pairs with at least one high percent contribution are labeled. The results from Treene catchment are shown as black circles and from Saale catchment in grey squares. Please note that percentage contributions on the x-axis sum up to 100%, while this is not the case for the y-axis.

List of Tables

Table 1. List of SWAT models parameters. Lower and upper ranges are given as absolute range (r), additive (a) or multiplicative (m) value. Further information can be found in the theoretical documentation of the SWAT model (Neitsch et al., 2011).

Parameter name	Abbreviation	Process	Units	Range	Lower	Upper
			type	range	range	
Snow fall temperature	SFTMP	Snow	C	r	-2.5	2.5
Snow melt temperature	SMTMP	Snow	C	r	-2.5	2.5
Curve number	CN2	Surface runoff		a	-10	10
Surface runoff lag time	SURLAG	Surface runoff		r	0.8	4
Lateral flow lag time	LATTIME	Lateral flow	days	r	0.2	8
Tile flow lag time	GDRAIN	Tile flow	hours	m	0.5	1.5
Available water capacity of a soil layer	SOL_AWC	Soil water	mm H20/mm soil	a	-0.02	0.1
Saturated hydraulic conductivity of a soil layer	SOL_K	Soil water	mm/h	m	0.5	3
Soil evaporation compensation factor	ESCO	Evapotranspiration		r	0.2	1
Groundwater delay time (fast aquifer)	GW_DELAYfsh	Groundwater	days	r	1	50
Aquifer fraction coefficient (slow aquifer)	RCHRGssh	Groundwater		r	0.2	0.8
Baseflow alpha factor (slow aquifer)	ALPHA_BFssh	Groundwater	1/days	r	0.001	0.2