Hydrol. Earth Syst. Sci. Discuss., 9, 12989–13036, 2012
www.hydrol-earth-syst-sci-discuss.net/9/12989/2012/
doi:10.5194/hessd-9-12989-2012

Hydrology and Earth System Sciences Discussions

A framework to assess the realism of model structures using hydrological signatures

T. Euser[1], **H. C. Winsemius**[2], **M. Hrachowitz**[1], **F. Fenicia**[1,3], **S. Uhlenbrook**[1,4], **and H. H. G. Savenije**[1]

[1]Delft University of Technology, Water Resources section, P.O. Box 5048, 2600 GA, Delft, The Netherlands
[2]Deltares, P.O. Box 177, 2600 MH Delft, The Netherlands
[3]Centre de Recherche Public Gabriel Lippmann, Department Environment and Agro-Biotechnologies, 4422 Belvaux, Grand-Duchy of Luxembourg
[4]UNESCO-IHE Institute for Water Education, P.O. Box 3015, 2601 DA Delft, The Netherlands

Received: 15 October 2012 – Accepted: 19 October 2012 – Published: 16 November 2012

Correspondence to: T. Euser (t.euser@tudelft.nl)

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

Abstract

The use of flexible hydrological model structures for hypothesis testing requires an objective and diagnostic method to identify whether a rainfall-runoff model structure is suitable for a certain catchment. To determine if a model structure is realistic, i.e. if it captures the relevant runoff processes, both performance and consistency are important. Performance describes the ability of a model structure to mimic a specific part of the hydrological behaviour in a specific catchment. This can be assessed based on evaluation criteria, such as the goodness of fit of specific hydrological signatures obtained from hydrological data. Consistency describes the ability of a model structure to adequately reproduce several hydrological signatures simultaneously, while using the same set of parameter values. In this paper we describe and demonstrate a new evaluation Framework for Assessing the Realism of Model structures (FARM). The evaluation framework tests for both performance and consistency using a principal component analysis on a range of evaluation criteria, all emphasizing different hydrological behaviour. The utility of this evaluation framework is demonstrated in a case study of two small headwater catchments (Maimai, New Zealand and Wollefsbach, Luxembourg). Eight different hydrological signatures and eleven model structures have been used for this study. The results suggest that some model structures may reveal the same degree of performance for selected evaluation criteria, while showing differences in consistency. The results also show that some model structures have a higher performance and consistency than others. The principal component analysis in combination with several hydrological signatures is shown to be useful to visualize the performance and consistency of a model structure for the study catchments. With this framework performance and consistency can be tested to identify which model structures suit a catchment better than other model structures.

1 Introduction

One of the main purposes of hydrological modelling is to develop better predictive models of rainfall-runoff processes. To really improve these models it is important to have a good understanding of the hydrological behaviour of catchments and to be able to explain the variability in catchment response and the factors influencing it (Kirchner, 2006; Fenicia et al., 2008b). Each hydrological model concept is a hypothesis of catchment behaviour (Savenije, 2009), and therefore a suitable tool to gain more knowledge about catchment response patterns. However, for models to be a suitable tool, it is very important that the "right" model is chosen for a certain catchment. Due to differences between catchments, different models can be "right" for different catchments (cf. McMillan et al., 2011).

Clark et al. (2011) argue that the use of multiple hypotheses (models) can help to develop a better understanding of the catchment behaviour. Every model structure consists of several components, representing different runoff processes. By using the ensemble of components that most adequately simulate the available data, the selected model structure can be assumed to be the one best representing real world processes. Fenicia et al. (2011) describe the SUPERFLEX framework which can be used to configure such different model structures. With this framework it is possible to conveniently compare different model structures and their underlying hypothesis and hence use them as a learning tool to improve our understanding of the behaviour of individual catchments. When different (flexible) model structures are used for hypothesis testing, the understanding of catchment behaviour can be increased by investigating whether a model is able to represent the dominant processes in the catchment (Fenicia et al., 2008a). When this is the case, it may be said that the hypothesis that a model structure "suits a catchment" cannot be rejected. To test if dominant processes are represented by a given model structure, it is important to have a sound method to evaluate which model structure suits better for a certain catchment and to understand the reasons behind it (Kirchner, 2006).

It is increasingly acknowledged that model evaluation based on single objective optimisation (often performed with standard least squares optimization) is insufficient to appropriately identify dominant processes. The use of a multi-objective optimisation offers more insight into the processes underlying the observed catchment response (e.g. Gupta et al., 1998; Seibert, 2000; Schaefli and Gupta, 2007; Wagener et al., 2003; Winsemius et al., 2009; Hrachowitz et al., 2012). The use of specific characteristics of the hydrograph, hereafter referred to as hydrological signatures, for the (multi-objective) evaluation of the performance of hydrological models can give even more information about the hydrological behaviour of the modelled catchments. The use of such hydrological signatures can therefore strengthen the link between the models and the underlying hydrological processes (e.g. Gupta et al., 2008; Yilmaz et al., 2008; Hingray et al., 2010). Using hydrological signatures for model evaluation has some advantages and disadvantages in relation to traditional hydrograph fitting. The main disadvantage is that for most signatures the phase information, i.e. the timing, is lost and the shape of the hydrograph is no longer taken into account. The main advantage, however, is due to not taking into account phase information, the heterogeneity and small measurement errors in the input data have less influence on the evaluation than with traditional hydrograph fitting.

In this paper a framework is proposed to evaluate the suitability of model structures for a given catchment (FARM – Framework for Assessing the Realism of Model structures). The realism, or suitability is defined as a function of both *performance* and *consistency* of different model structures. In this study, performance is defined as the ability of a model structure to reproduce several signatures, expressed as evaluation criteria; consistency is defined as the ability of a model structure to reproduce different signatures with the same set of parameters. So, for this study consistency implies satisfying different evaluation criteria simultaneously and does not explicitly relate to consistency in time or space. However, higher performance and better consistency result in higher confidence that a model represents the dominant processes of a given catchment, thereby to a certain level implying consistency in time and space. The novelty of this study is that in addition to performance also consistency based on different evaluation criteria is taken into account to identify the most suitable model structure for a given catchment.

A Principal Component Analysis (PCA) is a common statistical tool to decrease the dimensions of a problem. In hydrology it has been used for example in tracer studies to investigate the correlation between tracer response patterns (e.g. Brown et al., 1999; Worrall et al., 2006; Hrachowitz et al., 2011). In principle, a PCA can also be used to investigate the correlation between different evaluation criteria. Therefore, the objectives of this study are to test (1) whether an evaluation framework using a PCA together with hydrological signatures can help to determine the performance and consistency of model structures for a certain catchment and (2) if this framework can be used to identify whether certain model structures suit a catchment better than other model structures. Follows the evaluation framework will be described, followed by an application of the framework in a case study (Sects. 3, 4 and 5).

2 Description framework

FARM (Framework for Assessing the Realism of Model structures) makes use of three main elements: model structures, hydrological signatures and the Principal Component Analysis (PCA). Figure 1 describes how these elements interact in the general framework. The PCA is the general part of this framework; therefore, it will be described first. The model structures and hydrological signatures depend on the specific study this framework will be used for. Therefore, they are mainly described in the methodology part of the application.

The framework consists of the following steps (Fig. 1):

1. selection of a catchment and gathering of hydrological process knowledge;
2. definition of hydrological signatures;

3. definition of evaluation criteria to assess the models' ability to reproduce the hydrological signatures;
4. selection of a set of plausible model structures for hypothesis testing;
5. deriving a posterior parameter distribution for the selected model structures and catchments (calibration);
6. random sampling of N parameter sets from the derived posterior parameter distribution and calculation of the evaluation criteria for the modelled hydrographs;
7. Principal Component Analysis for each combination of catchment and model structure, and
8. assessment of relative performance and consistency for each combination of catchment and model structure.

2.1 Definitions

Performance and consistency are important concepts in this paper, therefore they are explained below.

- *Performance*: The performance of a model structure for a certain catchment is determined by its ability to reproduce a certain hydrological behaviour or signature. This can be measured with the maximum value for an evaluation criterion (belonging to the best parameter set), which describes this hydrological signature, and by the range of values covered by the evaluation criterion (belonging to all the parameter sets from the posterior distribution). Here, to assess the relative performance of a model structure three performance categories are defined: high, moderate and poor. A model structure is assumed to perform better when more evaluation criteria are in the highest performance category.

- *Consistency*: The consistency of a model structure for a certain catchment is determined by the number of evaluation criteria, describing different hydrological signatures, that have their best performance for a specific parameter set. The consistency of model structures can vary gradually between fully consistent and fully inconsistent. It is important to have insight in the consistency of model structures for two reasons: first, a high consistency means that the model is capable to reproduce several hydrological signatures with the same parameter set, implying a better representation of real world processes, i.e. the model can reproduce different, ideally contrasting, aspects of the hydrograph. Second, a highly consistent model is thus expected to behave comparably in the calibration and validation period (Kirchner, 2006; Fenicia et al., 2007) and would therefore have a reduced predictive uncertainty.

The consistency and performance of a model structure can be determined independently, but are both important for the evaluation of the model structures. Only a model with high performance and high consistency may be considered a suitable hypothesis for a certain catchment and therefore, points towards a high degree of realism. In reality all signatures occur simultaneously. Hence, a model that is able to reproduce all selected signatures to a high degree with the same parameter set has a higher degree of realism than a model structure that is not able to do that. However, it is possible that, for a certain model structure, the degree of performance is different from the degree of consistency. The consequences for different combinations of the degree of consistency and performance are shown in Fig. 2. For an inconsistently good model structure, signatures are reproduced well, but not with the same parameter set. For a consistently poor model structure, signatures are not represented correctly, although the model is consistent. So, a high degree of consistency only gives extra value in the evaluation process when it is combined with a high performance.

2.2 Principal Component Analysis (PCA)

A Principal Component Analysis (PCA) is a statistical tool which can be used to reduce the dimensions of a multivariate problem; the basic principles of a PCA can be found in literature about multivariate analysis (e.g. Krzanowski, 2000; Härdle and Simar, 2003). Note that here the vectors of the loadings are referred to as "vectors" thereafter.

2.2.1 Use of PCAs for this framework

Input for PCAs

For FARM PCAs are used to identify the correlation between different evaluation criteria. A PCA is performed for each model structure in each catchment for *N* parameter sets. Where *N* is the number of parameter sets needed to reach convergence (see Sect. 6.2). The parameter sets are randomly sampled from a derived posterior parameter distribution. For these *N* samples all the evaluation criteria are calculated (see Fig. 1), these values form the input to the PCA. Note that the model calibration strategy remains the choice of the modeller.

For a PCA it is assumed that the input data is generated from a multivariate normal distribution (Johnson and Wichern, 1998). If this is not the case, the values for the evaluation criteria have to be transformed to a normal distribution. This transformation could for example be done with a normal quantile transformation (Weerts et al., 2011; Montanari and Brath, 2004).

Interpretation of PCAs

The PCA represents two model characteristics: the performance and the consistency. The performance categories are presented by the thickness of the vectors in the PCA diagram (see for example the results of the Maimai in Fig. 8). Note that for each study specific values for the categories should be defined.

The degree of consistency is presented by the configuration of the vectors in the PCA. When a model structure is able to simulate different signatures well with the same set of parameter values, the corresponding evaluation criteria should be directly correlated. In other words, a better performance on one evaluation criterion also means a better performance on another evaluation criterion. For the PCA this results in the vectors representing the evaluation criteria pointing in the same direction. When evaluation criteria are inversely correlated, it means that a parameter set with a better performance for one criterion leads to a worse performance for another. The diagram which is the result of the PCA can be characterised by five different types of configurations (Fig. 3):

1. All evaluation criteria are completely and directly correlated ("line-shaped" diagram) (Fig. 3a). When this is the case, the model is fully consistent, which would be the case for a hypothetical "perfect" model.
2. All evaluation criteria have their highest loading in the same direction on one principal component (PC) and thus are all directly correlated (Fig. 3b). When this is the case, the model is consistent.
3. The evaluation criteria are all located in one quadrant of the diagram and are all partly directly correlated (Fig. 3c). An increase in performance for one criterion does not result in a decrease in performance for another criterion. Therefore this configuration has a moderate degree of consistency.
4. The evaluation criteria have their longest distance in the same direction on one of the two principal components and are therefore all directly correlated or uncorrelated ("L-shaped" diagram) (Fig. 3d). This configuration has a moderate degree of consistency as well, as some evaluation criteria are correlated, while others are uncorrelated.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀◀ ▶▶
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

5. The evaluation criteria show a “star shaped” diagram and some evaluation criteria are uncorrelated, while others are inversely correlated (Fig. 3e). In this case the model is inconsistent.

The configurations in Fig. 3 are basic configurations; in case of deviations from these basic configurations, three measures are important for interpretation of the PCA diagrams, these three are listed below. These measures can in principle be objectively determined, but in this study they are only determined visually.

- *Spreading on PC1 or PC2 (x or y axis)*: PC1 always represents a larger part of the explained variance in the data, so a spread or inversely correlated evaluation criteria on PC1 determine the consistency to a larger extent than inversely correlated evaluation criteria on PC2;
- *length of the vectors*: the longer a vector, thus the higher the loadings, the more influence the vector has on the total analysis. An inversely correlated vector which is relatively small, influences the consistency less than an inversely correlated vector which is relatively long;
- *inversely correlated thick vectors*: a thick vector means that there is a parameter set for which the signature can be modelled well; a thin vector indicates poorer model performance. So, inversely correlated thick vectors indicate that inconsistency is the main problem, while inversely correlated thin vectors indicate that performance is still the main problem.

Note that a PCA only shows the *relative* similarities and differences within the data used for the PCA, therefore the absolute values on PC1 and PC2 and the individual direction of the vectors are of no importance. When interpreting a PCA diagram only the relative directions of the vectors and the relative length differences of the vectors are important.

2.3 Hydrological signatures

The performance and consistency of the model structures is evaluated with evaluation criteria based on hydrological signatures. These signatures can be derived from the observed hydrograph, for example the flow duration curve or the auto correlation coefficient. However, these signatures can in principle also be derived from other data sources, for example ground water levels, tracer data or satellite data. Note that the "more independent" the selected signatures are, i.e. reflecting contrasting parts of the hydrograph, the higher the significance of their PCA interpretation.

Most signatures are represented by one value for the observed and one value for each modelled hydrograph. A possibility to formulate the evaluation criterion (F) is shown in Eq. (1). Only the value for the signature of the modelled hydrograph changes per parameter set, the value for the observed hydrograph is the same for each parameter set. By dividing the modelled value by the observed value the relative deviation of the modelled from the observed value can be obtained. The absolute value and "1-" the fraction are required to obtain the same result (F) for the same deviation of the modelled value above or below the observed value.

$$F = \left| 1 - \frac{S(Q_{\mathrm{mod}})}{S(Q_{\mathrm{obs}})} \right| \tag{1}$$

With $S(Q_{\mathrm{mod}})$ the value of the hydrological signature for the modelled hydrograph and $S(Q_{\mathrm{obs}})$the value of the hydrological signature for the observed hydrograph. With this formulation of the evaluation criterion, the lower the value for the evaluation criterion, the better the performance. For the PCA it is convenient to link a better performance to a higher value for the evaluation criterion. So, the formulation in Eq. (2) could be used for the PCA.

$$F_{\mathrm{PCA}} = 1 - F \tag{2}$$

3 Study areas

Two small headwater catchments have been selected for this case study: the Maimai M8 catchment in New Zealand (0.038 km^2) and the Wollefsbach catchment in Luxembourg (4.6 km^2). The catchments have been selected because of their small size and their data availability. Another advantage of these two catchments is their previous use in other research projects (e.g. McGlynn et al., 2002; Fenicia et al., 2008a; Kavetski and Fenicia, 2011). These previously obtained results can be used to check the new results for plausibility. Figure 4 shows the discharge, precipitation and potential evaporation for both catchments.

3.1 Maimai

The Maimai M8 catchment is located in the northern part of New Zealand's South Island (Fig. 5). It is small (0.038 km^2), but one of the most researched catchments worldwide (McGlynn et al., 2002). The Maimai has short, steep slopes and shallow soils, where saturation seldom decreases below 90 %. The subsoil is poorly permeable and the yearly deep percolation rate is approximately 100 mm yr^{-1}. The whole catchment is forested with a mixture of deciduous trees, which leads to an interception of about 26 % of the rainfall. The yearly rainfall and discharge are approximately 2600 mm yr^{-1} and 1550 mm yr^{-1}, respectively. More information about this catchment and previous research is described in a review by McGlynn et al. (2002). Due to the climate, the physical properties of the catchment and as a result of this, the fact that the catchment is most of the time saturated, the rainfall-runoff processes are relatively easy to model. The wet climate with little seasonality leads to a system with a limited number of hydrological regimes. The steep slopes together with the shallow, saturated soils and the impermeable subsurface lead to a quick response of the catchment (Vaché and McDonnell, 2006). For the Maimai catchment hourly data of discharge, precipitation and potential evaporation from 1 January 1985 till 31 December 1987 is used. The rainfall is measured with a recording raingauge which is located inside the catchment. The potential evaporation is estimated as described by Rowe et al. (1994). The first year of the data is used as a warm-up period; the last two years are used for calibration.

3.2 Wollefsbach

The Wollefsbach is located in the Attert catchment in Luxembourg (Fig. 6). The Wollefsbach is a small headwater catchment, like the Maimai; however, the catchment area is about 100 times larger (4.6 km^2). The Wollefsbach has shallow top soils, with a low permeable clay layer in the subsoil; therefore the deep percolation is minimal (Kavetski and Fenicia, 2011). The land use in the catchment consists mainly of grass and cropland. The discharge in the Wollefsbach is characterized by a quick response during the winter period and almost no discharge in the summer period (see also Fig. 4). For the Wollefsbach catchment hourly data of discharge, precipitation and potential evaporation from 1 September 2004 till 30 August 2007 is used. The rainfall is measured with two tipping buckets which are located inside the catchment. The potential evaporation is estimated with the Penman equation. The first year of the data is used as a warm-up period, the last two years for calibration.

4 Methodology

In this section the specifics of the framework are described for this case study.

4.1 PCA

Here, the model posterior parameter distributions were determined with Bayesian inference, using a heteroscedastic error model based on the Weighted Least Squares (WLS) scheme (Thyer et al., 2009). 1000 random samples are drawn from these posterior distributions and all the evaluation criteria are calculated for each sample. These evaluation criteria were then transformed to normal distributions with a normal quantile transformation (Weerts et al., 2011; Montanari and Brath, 2004). The transformed criteria were subsequently used as input for the PCAs.

The performance categories for this case study are defined as follows:

- *High* (continuous and very bold vectors): maximum value for the evaluation criterion higher than 0.8 with 90 % of the values for the evaluation criterion higher than 0.65.
- *Moderate* (dashed and bold vectors): maximum value for the evaluation criterion higher than 0.4 with 90 % of the values for the evaluation criterion higher than 0.3.
- *Poor* (dotted and thin vectors): all other cases.

4.2 Hydrological signatures

The signatures which have been used for this case study are described in the following. All the signatures are calculated for the total modelled period and in addition some are also calculated for specific periods. These periods are the periods in which the low flows (May–September) or high flows (November–April) occur in the Wollefsbach. In the Maimai the seasonality is minimal; therefore there are no clear periods of high and low flow; however, the same signatures and periods are used for both catchments: May till September as low flow period and November till April as high flow period. Most of the signatures are expressed as evaluation criterion as defined in Eq. (1), except for the flow duration curve, as this signature (the flow duration curve itself) is not represented by one value. The equations and a sketch of each signature are shown in Table 1.

4.2.1 Autocorrelation (AC)

The autocorrelation is a measure for the smoothness of a hydrograph: a high autocorrelation means a small difference between two consecutive points. For this signature the correlation coefficient of the autocorrelation with a lag of 1 day for a hydrograph is calculated (Winsemius et al., 2009). A lag of 1 day means that within a hydrograph a data point is compared with the data point 1 day earlier. For the total flow period this signature is used to represent the timing of the peaks.

Low flow period (AClow)

The low flow period is taken into account to investigate whether this signature can be used to evaluate a quick response of the catchment on rain events in the summer period. In the Maimai catchment there is no clear low flow period, so it is expected that for the Maimai the evaluation criterion for the low flow period is strongly directly correlated with the one for the total flow period.

4.2.2 Rising Limb Density (RLD)

Like the autocorrelation, this signature is an indication of the smoothness of the hydrograph, but the RLD is averaged over the total period and is completely independent of the flow volume (Shamir et al., 2005). This signature is calculated by dividing the number of peaks by the total time the hydrograph is rising. Therefore, the RLD is the inverse of the mean time to peak. Together with RLD also DLD (Declining Limb Density) was used before for supporting the calibration process (Shamir et al., 2005; Yadav et al., 2007) and for catchment classification (Sawicz et al., 2011).

4.2.3 Peak distribution (peaks)

This signature shows whether the peak discharges are of equal height, therefore only the peak discharges are taken into account. A peak discharge is the discharge at a time step of which both the previous and the following time step have a lower discharge. From these peak discharges a flow duration curve is constructed and the average slope between the 10th and 50th percentile is taken as the measure for this signature. By taking the 10th and 50th percentile, only the higher peaks (but not the extremes) are taken into account, which are the most interesting for this analysis (Sawicz et al., 2011). For the total flow period this signature is a measure for the differences in peak heights.

Due to measurement errors and heterogeneity the input rainfall for the modelled and observed discharge can be different, resulting in different peak heights. By using the slope of the flow duration curve, only the relative peak heights are compared, which should be comparable for the modelled and observed hydrograph.

Low flow period (peaksLow)

The low flow period is again taken into account to investigate whether this signature can identify the peaks in the discharge during the low flow period. For this reason the use of the 10th and 50th percentile are interesting, as identifying the small bumps is not useful for this analysis. In the Maimai catchment there is no clear low flow period, so it is expected that for the Maimai the evaluation criterion for the low flow period is strongly directly correlated with the one for the total flow period.

4.2.4 Flow Duration Curve (FDC)

For this signature a flow duration curve is constructed from all the discharge data. The Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 1970) between the observed and modelled flow duration curve is taken as the evaluation criterion. Flow duration curves are frequently used hydrological signatures to evaluate the overall behaviour of a catchment. Depending on the study, different parts of the FDC were previously investigated (Yadav et al., 2007; Yilmaz et al., 2008; Blazkova and Beven, 2009; Westerberg et al., 2011). The FDC for the total flow period represents the overall behaviour of a catchment. By taking the Nash-Sutcliffe Efficiency of the flow duration curve, instead of the Nash-Sutcliffe Efficiency of the flows, the magnitudes of flow are taking into account, without focusing on timing problems and missed or unrepresented rainfall events due to heterogeneity of rainfall.

Low flow period (FDClow)

When only using the total flow period, the low flows are not specifically taken into account. This signature for the low flow period represents the overall behaviour of a catchment during the low flow period. In the Maimai catchment there is no clear low flow period, so it is expected that the result for the low flow period is strongly correlated to the result of the total period.

High flow period (FDChigh)

When only using the total flow period, also the high flows are not specifically taken into account. This signature for the high flow period represents the overall behaviour of a catchment during the high flow period. As in the Maimai catchment there is no clear high flow period as well, it is expected that the result for the high flow period is strongly directly correlated to the result of the total and low flow period.

4.2.5 Reference evaluation criteria

In addition to the evaluation criteria based on a hydrological signature, also two reference evaluation criteria are used: Nash-Sutcliffe Efficiency (E_{NS}) and the Nash-Sutcliffe Efficiency of the log of the flows (E_{logNS}). These evaluation criteria are taken into account because they (especially the Nash-Sutcliffe Efficiency) are commonly used for the evaluation of hydrological models (Schaefli and Gupta, 2007) and are therefore suitable to use as benchmark for this study.

4.3 Model structures

For this study nine flexible model structures are tested, their performance and consistency is compared with 2 (fixed) benchmark models: GR4H (an hourly version of GR4J, Perrin et al., 2003) and a modified version of the HBV model (Lindström et al., 1997). The main adaptation on the HBV model is that river routing is not included (Dmitri

Kavetski, personal communication, 2012), because it is not considered as a crucial process due to the small size of the catchments. These benchmark models are mainly selected because they are widely used for hydrological modelling.

4.3.1 Configurations flexible model structures

The nine flexible model structures have been configured with the SUPERFLEX framework (Fenicia et al., 2011). Model structures built with the SUPERFLEX framework consist of reservoir elements, lag function elements and junction elements. The created model structures (M1 to M9, see also Fig. 7 and Table 2) differ in the number of reservoirs (1 to 5), the number of fluxes (3 to 10) and the number of parameters (1 to 9). The selection of the model structures is mainly based on the model structures used by Kavetski and Fenicia (2011) and on experiences of previous modelling exercises. A discussion of processes represented by the model structures can be found in Kavetski and Fenicia (2011).

4.3.2 Model evaluation

The model evaluation is done with Bayesian inference, as described by Kavetski and Fenicia (2011). The applied error model is based on weighted least squares. For the quasi-Newton parameter optimization 20 multi-starts are used. During the Markov-Chain Monte Carlo (MCMC) sampling 5000 parameter sets were generated. The prior and posterior parameter ranges are shown in Table 2–4.

5 Results

5.1 Maimai

The PCA results for the Maimai catchment of all model structures are shown in Fig. 8.

Performance vs. consistency

All the model structures developed with the flexible framework except M8 have a very small range in their maximum Nash-Sutcliffe Efficiency, M3 to M5 even have an equal maximum Nash-Sutcliffe Efficiency. However, the consistency (the configuration of the vectors in the diagrams) differs between the model structures. M1 and M3 show a comparatively high degree of consistency, i.e. a low spread of the vectors. For M1 the variance explained by PC2 is small compared to PC1, therefore the spreading on PC2 has a minor influence. The evaluation criteria for M3 almost show an L-shape, only E_{logNS} is inversely correlated. Model structures M4 to M7 are much less consistent. Model structure M8 behaves different from model structures M1 to M7: it has a relatively high maximum Nash-Sutcliffe Efficiency and a high performance for the other evaluation criteria, the diagram for M8 really shows an L-shaped configuration. Another interesting aspect is the high performance for most evaluation criteria for the HBV model, but a relatively low consistency. For the HBV model some evaluation criteria are inversely correlated on PC1 and the variance explained by PC2 is relatively high. GR4H has a high performance for most evaluation criteria, like the HBV model, but is more consistent than the HBV model, as the evaluation criteria are mainly inversely correlated on PC2.

5.2 Wollefsbach

The PCA results for the Wollefsbach catchment of all model structures are shown in Fig. 9. It can be seen that the results are less clear than for the Maimai: the consistency of the model structures is lower and it is more difficult to identify if a model structure has a higher degree of consistency than another.

Performance vs. consistency

The performance of all model structures is relatively low: only GR4H and HBV have four thick vectors, M1 to M5 only have one thick vector. It can be seen that M5 to M7 have a low consistency, i.e. a high degree of spreading, but their performance is better than for M1 to M4. The consistency of HBV and M8 is higher and their performance is higher than most of the other model structures. Although the consistency of M1 and M2 is also relatively good (the evaluation criteria are mainly spread on PC2), their performance is poor, so these model structures are consistently poor.

5.3 Comparison of catchments

The two catchments show large differences in performance and consistency. Both are much higher in the Maimai than in the Wollefsbach. The main similarity between the two catchments is the low consistency for the model structures with a groundwater reservoir (M6, M7 and M9). The performance and consistency for the model structures in both catchments are compared in Fig. 10. The classification for this figure is purely indicative with the purpose of showing the performance and consistency of model structures *relative* to those of other model structures. In this figure it can be seen that for both catchments M1 and M2 are consistently poor. Another thing is the difference between the catchments for M8 and M3. Both performance and consistency are much better for the Maimai, most likely because it is a very small and homogeneous catchment.

5.4 Independent test period

It may be expected that a consistent model structure behaves similar in the calibration and validation period as it is assumed to capture the dominant processes better than an inconsistent model (cf. Seibert, 2000). Therefore, the model structures are run for an independent test period with the parameter sets derived during the calibration. For the Maimai catchment one extra year of data was available, for the Wollefsbach catchment two extra years of data were available. Both the performance and consistency are compared for the calibration and validation period. In Fig. 11 an example is given to show the differences between a more (M8) and a less (M7) comparable behaviour between the calibration and validation period.

A summary of the results of both catchments is presented in Table 5 and 6. The model structures in these tables are ordered by consistency for the calibration period. For the Maimai it can be seen that both the performance and consistency changed between the calibration and validation period. Model structures with a low consistency in the calibration period, have bit larger changes for the validation period. For the Wollefsbach it can be seen that there are mainly changes in consistency between the calibration and validation period. For most model structures with a low consistency the configuration in the validation period changed much more than for the model structures with a higher consistency.

6 Discussion

6.1 Applicability

Comparing model structures based on both performance and consistency has some advantages with respect to a comparison based on either performance or consistency. This can especially be seen for M8, M3, GR4H and HBV in the Maimai catchment. Their performance is more or less equal, but their consistency is not. Another example is M1 and M2 for the Wollefsbach. Their performance is poor, while their consistency is relatively good for the hydrological signatures used for this study. This also shows that consistency on itself does not give useful information about a model structure. Rather, for model structures with a high performance, the degree of consistency gives useful information about the suitability for a certain catchment.

The results for the Wollefsbach are not as clear as for the Maimai, but for both catchments it is possible to point out model structures that better simulate the selected signatures than other model structures. Sometimes the differences between PCA diagrams are small; when comparing diagrams with small differences, it is important to keep in mind the three measures described in the section Description framework:

1. spreading on PC1 or PC2;
2. length of the vectors and
3. inversely correlated thick lines.

A model structure that suits a certain catchment is more likely to represent the dominant processes that actually occur in the catchment than model structures that are less suited for the catchment. Therefore, the model structure is an indication for dominant processes in a catchment. However, when the hydrograph does not contain information about certain processes, these processes will not be taken into account for the analysis. In that case, auxiliary data sources are required to reveal these processes (e.g. Vaché and McDonnell, 2006; Son and Sivapalan, 2007; Fenicia et al., 2010; Hrachowitz et al., 2012; Birkel et al., 2010). When extra data sources give extra information, it is expected that the evaluation criteria belonging to the extra hydrological signatures are uncorrelated with the evaluation criteria from the streamflow data.

In addition, poor performance and poor consistency of a certain model structure can be an indicator for the absence of a certain runoff processes in the catchment. This can be seen in the Maimai and the Wollefsbach: the consistency and performance (especially the Nash-Sutcliffe Efficiency) of M6, M7 and M9 are relatively low. These are the only flexible model structures with a groundwater reservoir, so probably a groundwater reservoir is not important or incorrectly represented for both catchments. This is also in accordance to the site description of both catchments: both have shallow soils and (almost) impermeable subsurface layers. The performance and consistency of M8 in the Maimai are very good; M8 has a riparian zone reservoir, which probably fits well with the almost year round saturated soils of the Maimai catchment.

The use of a PCA can also help to identify the relation between the dominant processes and the response behaviour of the catchment (the hydrograph). For example, from the PCA diagram of model structure M6 in the Wollefsbach catchment, it can be seen that FDClow has a low performance and is inversely correlated with FDC and FDChigh, for the calibration period. It can also be seen that peaksLow has a low performance and is inversely correlated with AClow. So, no parameter set can be selected with a good performance for signatures focusing on the high and low flow period, but also, no parameter set can be selected with a good performance for different signatures focusing on the low flow period. Therefore, it is likely that the representation of dominant processes for the low flow period should be adapted. In this case the existence of a groundwater reservoir in the model structure can have a high influence on the modelled discharge in the low flow period.

6.2 Sensitivity to number of parameter sets

In this case study 1000 parameter sets are used to construct the PCA. To investigate whether this number is sufficient, the sensitivity to the number of parameter sets was tested. To test the sensitivity of the PCA it is important to know if the PCA is ergodic. When this is the case there is a convergence to a stationary measure when enough samples are taken into account, this convergence is independent from the initial conditions (Descombes, 2012). To test whether the PCA is ergodic and to test if 1000 parameter sets are sufficient, a PCA is also performed with 500 and 200 parameter sets. When the differences between the diagrams with 200 and 500 parameter sets are larger than between the diagrams for 500 and 1000 parameter sets, it is an indication of convergence and ergodicity can be assumed. Figure 12 shows the PCA diagrams for M8 in both catchments for 200, 500 and 1000 parameter sets. In the figure it can be seen that the difference between selecting 1000 and 500 parameter sets is smaller than the difference between selecting 500 and 200 parameter sets. This sensitivity analysis is performed for all the model structures, the results are compared with a visual inspection. Convergence is present to varying degree for all model structures. Model structures with a higher performance and consistency and the model structures with less complexity exhibit larger convergence. However, these are not always the model structures with a more constrained posterior parameter distribution. In general, the convergence for all model structures shows that ergodicity can be assumed and that the use of 1000 parameter sets is sufficient to have an indication of consistency of the evaluated model structures in this study.

6.3 Validity of the framework

The use of PCAs for model evaluation also has limitations. The main limitation is the low variance explained by the first two principal components. For most model structures the variance explained is below 80 %. More reliable diagrams would therefore also incorporate the third principal component; however, a 3-D graph is more difficult to visualise and interpret than a 2-D graph. There are two situations related to a low explained variance, which are good to keep in mind when interpreting the PCA diagrams.

- *Consistent configuration with low variance explained*: the higher principal components (PC3 and higher) explain a smaller amount of variance, this variance can decrease the high consistency, but will not make the model really inconsistent;
- *inconsistent configuration with low variance explained*: the first two principal components already show inconsistency. The variance explained by the higher principal components is lower, so they are unlikely to change a diagram from inconsistent to consistent.

The diagrams presented in Figs. 8 and 9 are suitable to get some information about the consistency of a model structure in a catchment. When the results from the PCA are evaluated in a more quantitative way, more principal components should be taken into account.

Next to this limitation, also three other aspects influence the validity of the framework. These include the selection of hydrological signatures, the sometimes different PCA results for calibration and validation periods and the application of the framework in larger catchments. First the hydrological signatures: selecting different signatures from different data sources, results in testing different aspects, which leads to different results. The selection of the signatures is highly subjective and influences the results. For this framework a good approach would be to start with many signatures for a catchment and test which signatures are directly correlated. The signatures that are strongly directly correlated with another signature for each model structure can be omitted.

Second, the different PCA results for the calibration and validation period for some model structures. In Sect. 5.4 it is shown that generally the model structures with a higher consistency behave more similar in the calibration and validation period. However, this does not hold for all model structures. Therefore, before selecting a model structure which seems to have a very high consistency and performance it may be beneficial to test the performance and consistency on a different time period.

Finally, the scale of the catchment: for this study the framework has only been tested for two small headwater catchments. When applying the framework in larger scale catchments, additional questions will arise. The main question will be whether the model structures still function on this larger scales. Large catchments are more heterogeneous and the effect of the heterogeneity of the rainfall is larger. Therefore, the signal detected in the PCA will likely to be weaker, as the signatures in the hydrograph are a mixture of different processes in different parts of the catchment. Due to this, it will be more difficult to relate them to specific dominant runoff processes. For larger scale catchments it might also be required to use auxiliary data sources and formulate additional signatures and evaluation criteria from these data sources in order to also take into account the processes which are not presented by the hydrograph.

7 Conclusions

In this study we present a framework to jointly evaluate the performance and consistency of different model structures. The framework can be used to compare different candidate model structures for a certain catchment. The framework consists of a PCA in combination with several hydrological signatures. The configuration of the

PCA is a good measure to evaluate the consistency of model structures and different line widths in the PCA are a good addition to evaluate the performance of a model structure for a certain catchment as well. The framework is tested on two headwater catchments. Comparison of the model structures for these catchments showed clear differences between the model structures and the catchments. Therefore, this framework can help to test multiple hypotheses for a certain catchment. The comparison also showed that a high performance is not always related to a high consistency. Even if some evaluation criteria show a high performance, others might show a very low performance. Thus, it is important to take both aspects into account when evaluating whether a model structure suits a catchment.

Acknowledgements. We thank the Centre de Recherche Public – Gabriel Lippmann for providing us the data of the Wollefsbach catchment and John Payne and Lindsay Rowe from Landcare NZ and professor Jeff McDonnell (University of Saskatchewan, Canada) for providing us the data of the Maimai catchment.

References

Birkel, C., Dunn, S. M., Tetzlaff, D., and Soulsby, C.: Assessing the value of high-resolution isotope tracer data in the stepwise development of a lumped conceptual rainfallrunoff model, Hydrol. Process., 24, 2335–2348, doi:10.1002/hyp.7763, 2010. 13010

Blazkova, S. and Beven, K.: A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: skalka catchment, Czech Republic, Water Resour. Res., 45, W00B16, doi:10.1029/2007WR006726, 2009. 13004

Brown, V. A., McDonnell, J. J., Burns, D. A., and Kendall, C.: The role of event water, a rapid shallow flow component, and catchment size in summer stormflow, J. Hydrol., 217, 171–190, 1999. 12993

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resour. Res., 47, W09301, doi:10.1029/2010WR009827, 2011. 12991

Descombes, X.: Stochastic Geometry for Image Analysis, ISTE Ltd, London, 2012. 13011

Fenicia, F., Savenije, H. H. G., and Pfister, P. M. L.: A comparison of alternative multiobjective calibration strategies for hydrological modeling, Water Resour. Res., 43, W03434, doi:10.1029/2006WR005098, 2007. 12995

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: on the contribution of complementary data to process understanding, Water Resour. Res., 44, W06419, doi:10.1029/2007WR006386, 2008a. 12991, 13000

Fenicia, F., Savenije, H. H. G., and Pfister, P. M. L.: Understanding catchment behavior through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi:10.1029/2006WR005563, 2008b. 12991

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Assessing the impact of mixing assumptions on the estimation of streamwater mean residence time, Hydrol. Process., 24, 1730–1741, doi:10.1002/hyp.7595, 2010. 13010

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010WR010174, 2011. 12991, 13006

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, 1998. 12992

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, doi:10.1002/hyp.6989, 2008. 12992

Härdle, W. and Simar, L.: Applied Multivariate Statistical Analysis, Springer-Verlag, Berlin Heidelberg, 2003. 12996

Hingray, B., Schaefli, B., Mezghani, A., and Hamdi, Y.: Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments, Hydrological Sci. J., 55, 1002–1016, doi:10.1080/02626667.2010.505572, 2010. 12992

Hrachowitz, M., Bohte, R., Mul, M. L., Bogaard, T. A., Savenije, H. H. G., and Uhlenbrook, S.: On the value of combined event runoff and tracer analysis to improve understanding of catchment functioning in a data-scarce semi-arid area, Hydrol. Earth Syst. Sci., 15, 2007–2024, doi:10.5194/hess-15-2007-2011, 2011. 12993

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., and Soulsby, C.: What can flux tracking teach us about water age distributions and their temporal dynamics?, Hydrol. Earth Syst. Sci. Discuss., 9, 11363–11435, doi:10.5194/hessd-9-11363-2012, 2012. 12992, 13010

Johnson, R. A. and Wichern, D. W.: Applied Multivariate Statistical Analysis, Prentice-Hall, Inc., Upper Saddle River, 1998. 12996

Kavetski, D. and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resour. Res., 47, W11511, doi:10.1029/2011WR010748, 2011. 13000, 13001, 13006

Kirchner, J. W.: Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi:10.1029/2005WR004362, 2006. 12991, 12995

Krzanowski, W. J.: Principles of Multivariate Analysis, a User's Perspective, Oxford University Press Inc, New York, 2000. 12996

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, 1997. 13005

McGlynn, B. L., McDonnel, J. J., and Brammer, D. D.: A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand, J. Hydrol., 257, 1–26, 2002. 13000

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure, Hydrol. Process., 25, 511–522, doi:10.1002/hyp.7841, 2011. 12991

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004. 12996, 13002

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – a discussion of principles, J. Hydrol., 10, 282–290, 1970. 13004

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003. 13005

Rowe, L. K., Pearce, A. J., and O'Loughlin, C. L.: Hydrology and related changes after harvesting native forest catchments and establishing pinus radiata plantations, Part 1: Introduction to study, Hydrol. Process., 8, 263–279, 1994. 13001

Savenije, H. H. G.: HESS Opinions "The art of hydrology", Hydrol. Earth Syst. Sci., 13, 157–161, doi:10.5194/hess-13-157-2009, 2009. 12991

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15, 2895–2911, doi:10.5194/hess-15-2895-2011, 2011. 13003

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, doi:10.1002/hyp.6825, 2007. 12992, 13005

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215–224, doi:10.5194/hess-4-215-2000, 2000. 12992, 13008

Shamir, E., Imam, B., Morin, E., Gupta, H. V., and Sorooshian, S.: The role of hydrograph indices in parameter estimation of rainfall-runoff models, Hydrol. Process., 19, 2187–2207, doi:10.1002/hyp.5676, 2005. 13003

Son, K. and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resour. Res., 43, W01415, doi:10.1029/2006WR005032, 2007. 13010

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis, Water Resour. Res., 45, W00B14, doi:10.1029/2008WR006825, 2009. 13001

Vaché, K. B. and McDonnell, J. J.: A process-based rejectionist framework for evaluating catchment runoff model structure, Water Resour. Res., 42, W02409, doi:10.1029/2005WR004247, 2006. 13000, 13010

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrol. Process., 17, 455–476, doi:10.1002/hyp.1135, 2003. 12992

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrol. Earth Syst. Sci., 15, 255–265, doi:10.5194/hess-15-255-2011, 2011. 12996, 13002

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15, 2205–2227, doi:10.5194/hess-15-2205-2011, 2011. 13004

Winsemius, H. C., Schaefli, B., Montanari, A., and Savenije, H. H. G.: On the calibration of hydrological models in ungauged basins: a framework for integrating hard and soft hydrological information, Water Resour. Res., 45, W12422, doi:10.1029/2009WR007706, 2009. 12992, 13002

Worrall, F., Burt, T., and Adamson, J.: Long-term changes in hydrological pathways in an upland peat catchment-recovery from severe drought?, J. Hydrol., 321, 5–20, doi:10.1016/j.jhydrol.2005.06.043, 2006. 12993

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774, doi:10.1016/j.advwatres.2007.01.005, 2007. 13003, 13004

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi:10.1029/2007WR006716, 2008. 12992, 13004

Table 1. Explanation of the different hydrological signatures used for this study. The formula for FDC directly gives the evaluation criterion. The formulas for AC, RLD and peaks only give the signature, the evaluation criterion can be derived with Eq. (1) (Q_i is the discharge at time step i, $\overline{Q}$ is the average discharge, $X_{\mathrm{FDC},i}$ is the value of the flow duration curve of the modelled discharge with i probability of exceedance, $Y_{\mathrm{FDC},i}$ is the value of the flow duration curve of the observed discharge with i probability of exceedance, $\overline{Y_{\mathrm{FDC}}}$ is the average observed discharge).

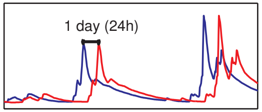
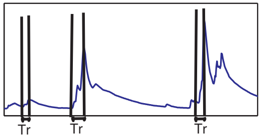
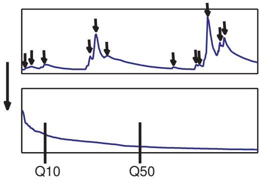
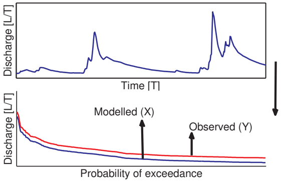
Signature	Formula	Sketch
Autocorrelation	$\mathrm{AC} = \frac{\sum\left(Q_i-\overline{Q}\right)\left(Q_{i+24}-\overline{Q}\right)}{\sum\left(Q_i-\overline{Q}\right)^2}$	
Rising Limb Density	$\mathrm{RLD} = \frac{T_r}{N_{\mathrm{peaks}}}$	
Peak distribution	$\mathrm{peaks} = \frac{Q_{10}-Q_{50}}{0.9-0.5}$	
Flow Duration Curve	$\mathrm{FDC} = \frac{\sum\left(X_{\mathrm{FDC},i}-Y_{\mathrm{FDC},i}\right)^2}{\sum\left(Y_{\mathrm{FDC},i}-\overline{Y_{\mathrm{FDC}}}\right)^2}$	

Table 2. Prior and posterior parameters ranges for both catchments and all flexible model structures. The first prior value of K_f is for M1, the second for M3-M5 and M8, the last for M6,M7 and M9.

	$I_{\max}$ (mm)	$S_{u,\max}$ (mm)	β (–)	$P_{\max}$ (mm h^{-1})	Fr (–)	D (–)	T_f (h)	K_r (1 h^{-1})	K_f (1 h^{-1})	K_s (1 h^{-1})
Prior Maimai	0.01–6.0	0.1–10000	0.001–10.0	10^{-6}–100.0	0.0–0.2	0.0–1.0	1.0–50.0	0.005–4.0	10^{-9}–10.0 10^{-8}–4.0 0.0001–4.0	0.0005–0.01
Posterior ranges Maimai										
M1	–	–	–	–	–	–	–	–	0.023–0.025	–
M2	–	42.4–46.1	3.61–4.38	0.58–0.64	–	–	–	–	–	–
M3	5.95–6.00	0.10–0.13	–	–	–	–	–	–	0.035–0.037	–
M4	0.010–1.51	61.5–83.1	0.43–0.51	–	–	–	–	–	0.039–0.043	–
M5	0.35–1.80	59.9–82.9	0.36–0.51	–	–	–	1.00–1.30	–	0.038–0.042	–
M6	1.99–2.51	34.6–38.0	–	–	–	0.25–0.27	1.00–1.30	–	0.06–0.065	0.0012–0.0014
M7	3.11–3.87	39.8–44.2	0.62–0.72	–	–	0.25–0.27	1.00–1.31	–	0.059–0.06	0.0012–0.0014
M8	0.010–0.025	14.8–16.2	1.60–1.82	–	0.199–0.200	–	1.00–1.36	0.0050–0.0057	0.097–0.104	–
M9	4.73–5.99	50.2–60.1	0.63–0.77	–	0.198–0.200	0.29–0.31	1.02–2.34	0.20–0.24	0.028–0.030	0.00092–0.00109
Prior Wollefs-bach	0.01–6.0	0.1–10000	0.001–10.0	10^{-6}–100.0	0.0–0.2	0.0–1.0	1.0–50.0	0.005–4.0	10^{-9}–10.0 10^{-8}–4.0 0.0001–4.0	0.00005 – 0.001
Posterior ranges Wollefsbach										
M1	–	–	–	–	–	–	–	–	0.00015–0.00017	–
M2	–	19.63–21.77	4.14–5.37	0.10–0.11	–	–	–	–	–	–
M3	0.010–0.012	44.08–45.89	–	–	–	–	–	–	0.042–0.046	–
M4	0.010–0.011	88.08–98.50	1.69–1.81	–	–	–	–	–	0.030–0.034	–
M5	0.010–0.011	90.26–99.96	1.69–1.79	–	–	–	2.43–3.36	–	0.032–0.034	–
M6	0.010–0.012	122.45–134.61	–	–	–	0.14–0.16	3.63–4.50	–	0.042–0.047	0.00099–0.00100
M7	0.010–0.011	84.68–92.53	1.32–1.41	–	–	0.15–0.16	3.49–4.355	–	0.045–0.048	0.00099–0.00100
M8	0.010–0.011	78.19–82.38	1.73–1.81	–	0.016–0.019	–	3.42–4.30	0.00500–0.00503	0.048–0.053	–
M9	0.010–0.011	82.86–90.61	1.32–1.40	–	0.0000069–0.000060	0.15–0.16	3.56–4.36	0.054–3.85	0.045–0.049	0.000996–0.001000

Table 3. Prior and posterior parameters ranges for both catchments for GR4H.

x_1 (mm)	x_2 (mm)	x_3 (mm)	x_4 (h)
	Prior		
1.0–2000.0	−100.0–100.0	1.0–500.0	0.51–20.0
	Posterior ranges Maimai		
118.0–129.4	−0.97–0.85	16.9–18.6	6.09–6.74
	Posterior ranges Wollefsbach		
92.4–117.1	−0.51–0.40	55.4–58.5	1.85–2.02

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc

Printer-friendly Version
Interactive Discussion

Table 4. Prior and posterior parameters ranges for both catchments for HBV.

FC (mm)	β (–)	PWP (mm)	L (mm)	k_0 (1 h^{-1})	k_1 (1 h^{-1})	k_{Perc} (1 h^{-1})	k_2 (1 h^{-1})	$I_{\max}$ (mm)
Prior								
1.0–500.0	1.0–10.0	1.0–500.0	0.05–50.0	0.001–30.0	0.0001–30.0	0.001–30.0	1.e-3–30.0	10^{-7}–10.0
Posterior ranges Maimai								
94.3–99.5	5.82–6.42	58.9–66.9	0.70–0.80	0.037–0.040	0.0087–0.011	0.0067–0.0074	0.0015–0.0017	5.26–5.60
Posterior ranges Wollefsbach								
44.5–52.5	2.90–3.54	34.3–43.8	11.6–12.2	0.18–0.21	0.035–0.038	0.014–0.015	0.0021–0.0027	6.67–7.50

Table 5. Summary of differences between the PCA graphs for the calibration and the independent test period for the Maimai catchment (EC = evaluation criterion). The model structures are ordered by consistency in the calibration period. “1, 2 or 3 EC changed” in the last column means the configuration of the PCA diagram of the calibration and validation is equal, but 1, 2 or 3 vectors have a different direction and/or length. “conf. changed” means that the relative direction of almost all vectors changed.

	Performance original[a]	Consistency original[b]	Performance validation[a]	Consistency change
HBV	7	low	7	config. changed
M7	2	low	4 (+2)	3 EC changed
M6	2	low	5 (+3)	2 EC changed
M9	4	low	5 (+1)	2 EC changed
M4	3	low	2 (−1)	small differences
M5	3	low	2 (−1)	small differences
M1	1	middle	2 (+1)	2 EC changed
M3	6	middle	5 (−1)	1 EC changed
GR4H	8	middle	9 (+1)	1 EC changed
M2	5	middle	5	small differences
M8	8	high	5 (−3)	1 EC changed

[a] The number of signatures in performance category high (thick vectors) is taken as a measure.

[b] According to Fig. 10.

Table 6. Summary of differences between the PCA graphs for the calibration and the independent test period for the Wollefsbach catchment (EC = evaluation criterion). The model structures are ordered by consistency in the calibration period. "1, 2 or 3 EC changed" in the last column means the configuration of the PCA diagram of the calibration and validation is equal, but 1, 2 or 3 vectors have a different direction and/or length. "conf. changed" means that the relative direction of almost all vectors changed.

	Performance original[a]	Consistency original[b]	Performance validation[a]	Consistency change	Performance validation[a]	Consistency change
M3	3	low	2 (−1)	config. changed	3	config. changed
M6	3	low	3	config. changed	3	config. changed
M7	3	low	3	config. changed	3	config. changed
M9	3	low	3	config. changed	2 (−1)	config. changed
GR4H	3	low	3	config. changed	2 (−1)	config. changed
M5	3	low	3	1 EC changed	3	1 EC changed
M1	2	middle	3 (+1)	config. changed	1 (−1)	config. changed
M8	2	middle	2	3 EC changed	2	3 EC changed
M4	2	middle	2	2 EC changed	3 (+1)	2 EC changed
HBV	2	middle	2	2 EC changed	1 (−1)	2 EC changed
M2	2	middle	1 (−1)	1 EC changed	2	1 EC changed

[a] The number of signatures in performance category high (thick vectors) is taken as a measure.

[b] According to Fig. 10.

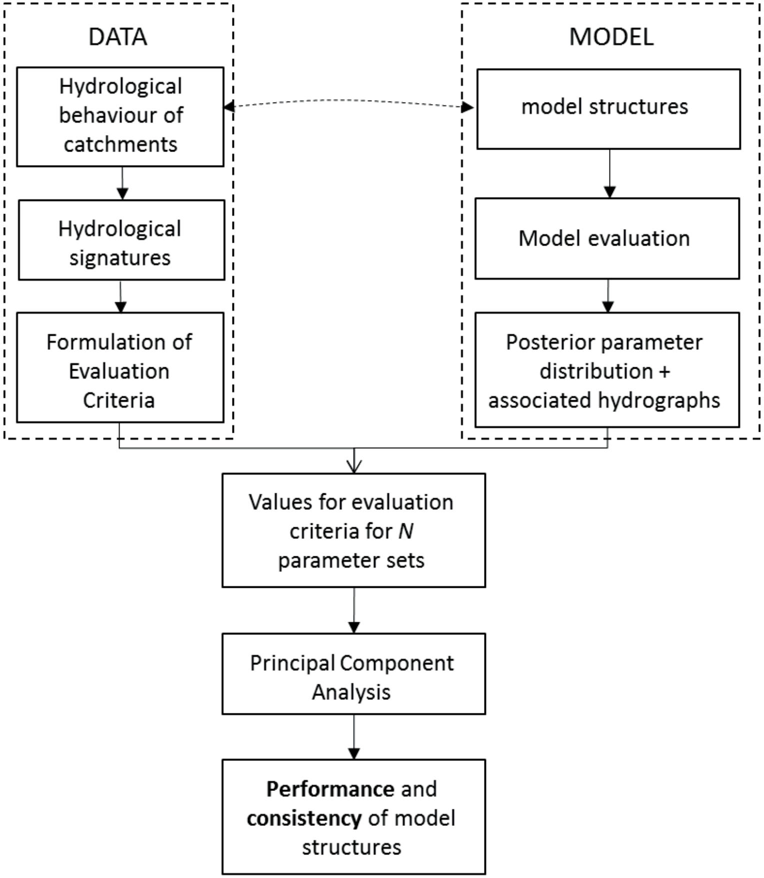


Fig. 1. Schematic overview of FARM to compare the performance and consistency of model structures with respect to hydrological signatures.

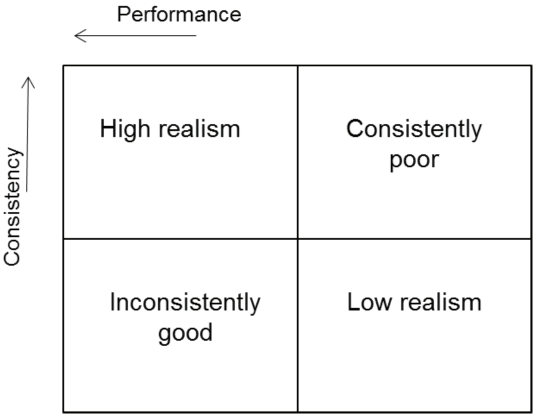


Fig. 2. Consequences for model structures for different combinations of performance and consistency, under the condition that the uncertainty of the input data is limited. The use of signatures for the evaluation of performance and consistency limits the influence of input uncertainty.

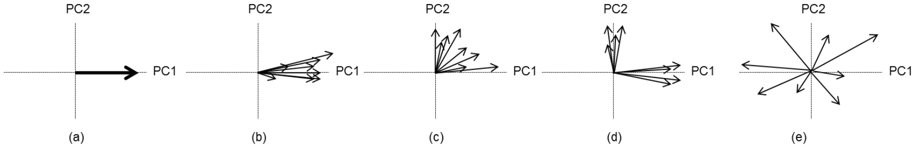


Fig. 3. Possible configurations for the PCA diagram: each vector represents an evaluation criterion (analysis is done per model structure). The axes are formed by the first two principal components (PC). **(a)** represents a fully consistent model structure, **(e)** a fully inconsistent model structure.

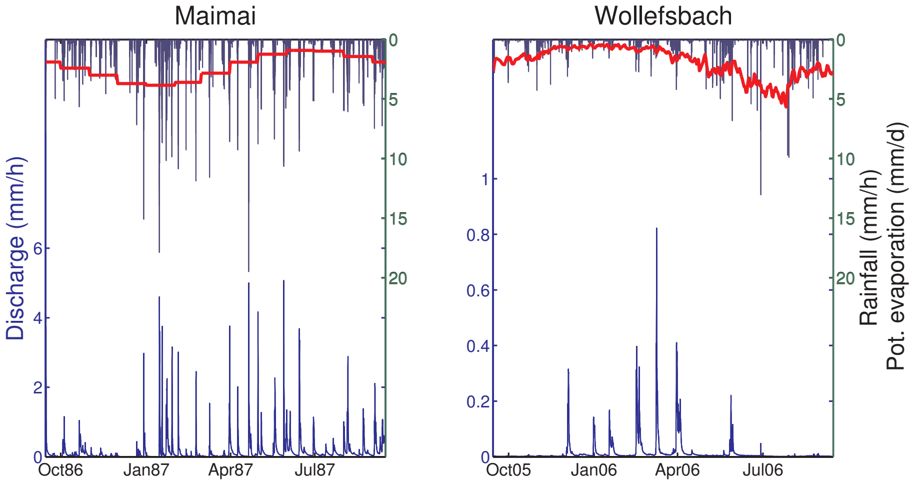


Fig. 4. Discharge (bottom), precipitation and potential evaporation (top) data for Maimai and Wollefsbach catchments (discharge = blue line, precipitation = blue bars, potential evaporation = red line). Note that the potential evaporation is presented in mm day^{-1} and the discharge and precipitation in mm h^{-1}. The discharge scale for both catchments differs: the discharge in the Wollefsbach is much lower.

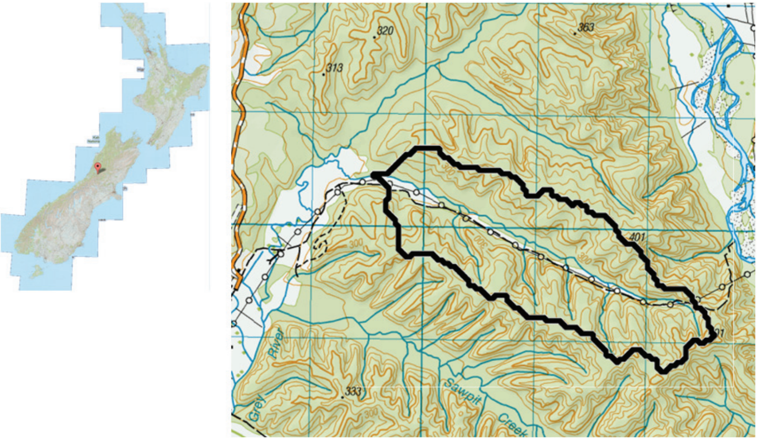


Fig. 5. Catchment area of the Maimai study area in New Zealand, the M8 catchment is one of the side branches of the main creek . Left: red dot indicates the location in New Zealand, right: topographic map of the Maimai study area with indicated the catchment boundary of the M8 catchment (available at: http://www.topomap.co.nz/).

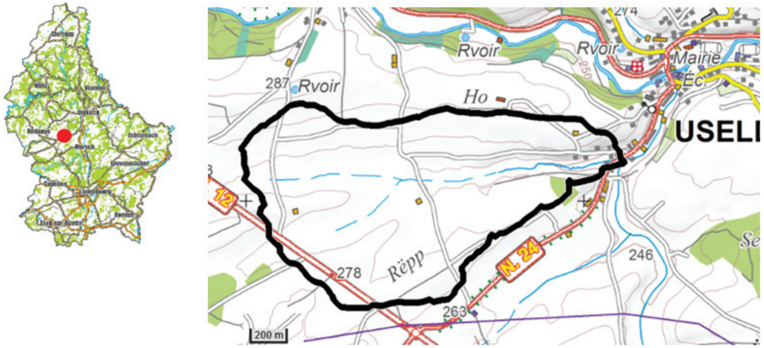


Fig. 6. Catchment area of the Wollefsbach catchment in Luxembourg. Left: red dot indicates the location in Luxembourg, right: topographic map of the Wollefsbach catchment with indicated the catchment boundary of the Wollefsbach catchment (available at: http://eau.geoportail.lu/).

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀◀ | ▶▶
◀ | ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

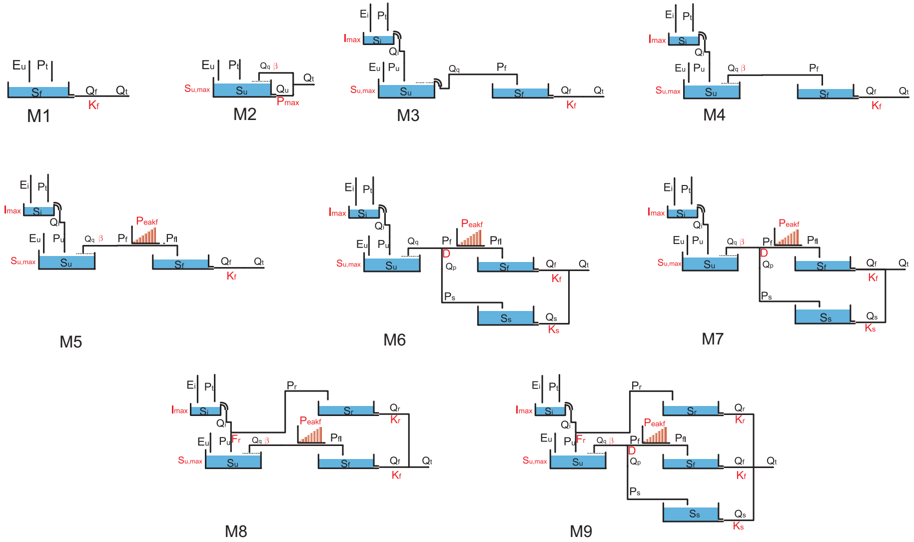


Fig. 7. Conceptual configurations of the flexible model structures used for this study.

HESSD

9, 12989–13036, 2012

Assessment of model structure realism

T. Euser et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀◀ | ▶▶
◀ | ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

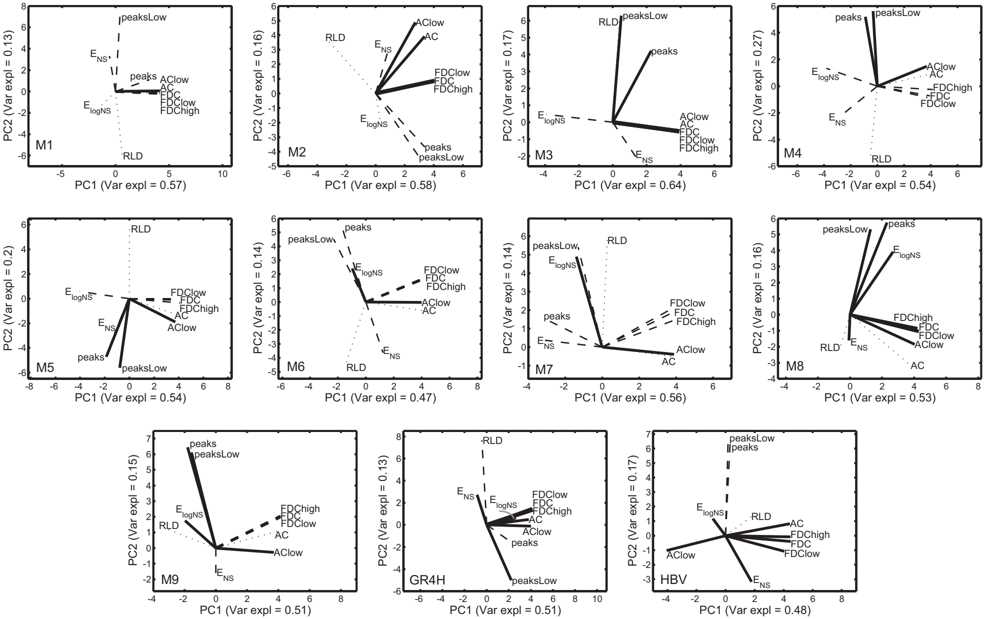


Fig. 8. Results for PCA for the *Maimai* catchment. Each figure represent one of the model structures. The figures are based on 1000 parameter sets. The total variance explained by these figures is the sum of the explained variance per PC.

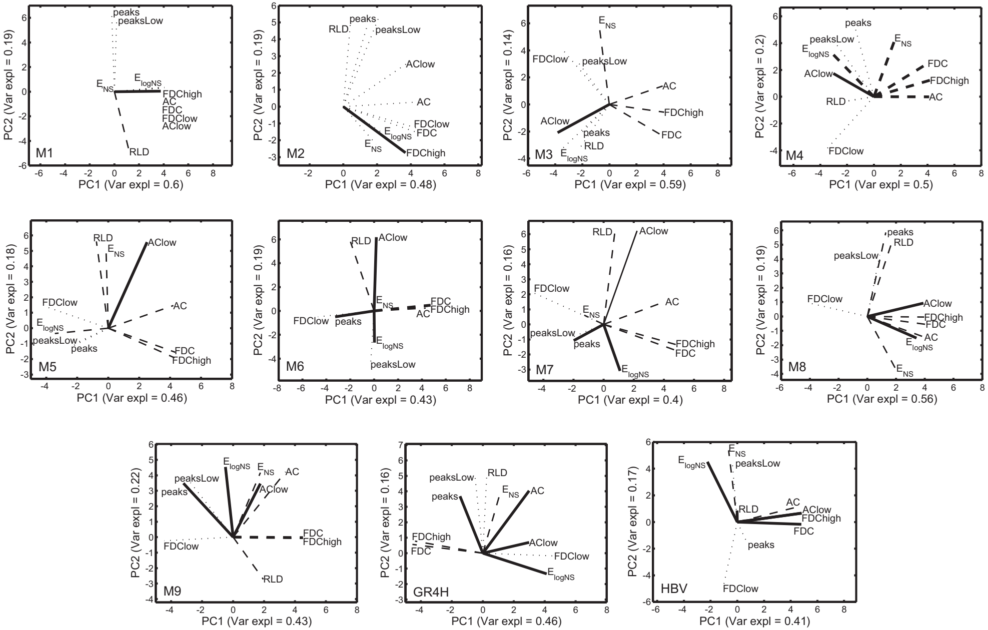
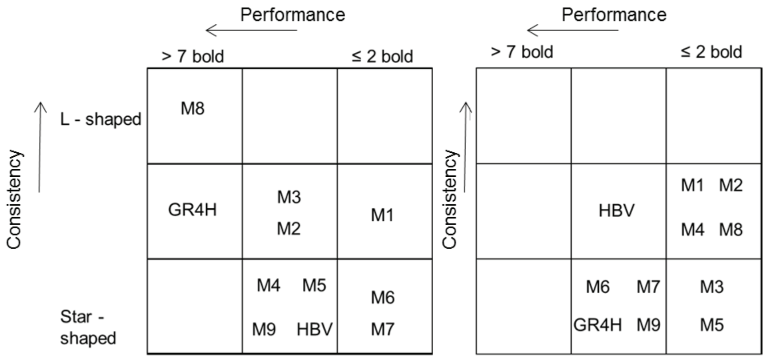


Fig. 9. Results for PCA for the *Wollefsbach* catchment. Each figure represent one of the model structures. The figures are based on 1000 parameter sets. The total variance explained by these figures is the sum of the explained variance per PC.



	> 7 bold		≤ 2 bold
L - shaped	M8		
	GR4H	M3 M2	M1
Star - shaped		M4 M5 M9 HBV	M6 M7

	> 7 bold		≤ 2 bold
		HBV	M1 M2 M4 M8
		M6 M7 GR4H M9	M3 M5

Fig. 10. Overview of the performance (columns) and consistency (rows) of the Maimai (left) and Wollefsbach (right). The middle row and column indicate a moderate consistency and performance. There is only a difference between the squares: the position of a model structure in a square is arbitrary. The PCA configurations for a high consistency (line-shaped) are not presented in this figure, as those configurations did not occur among the results.

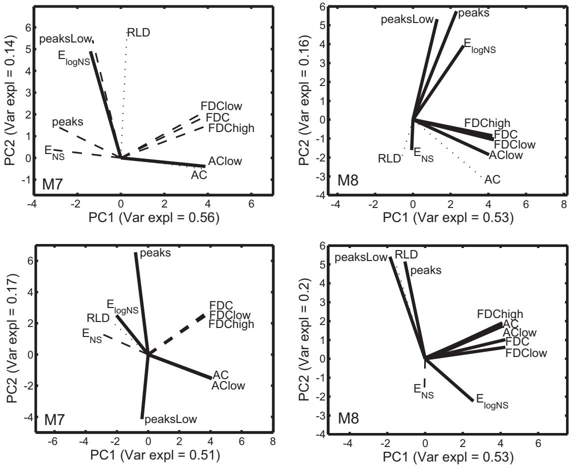
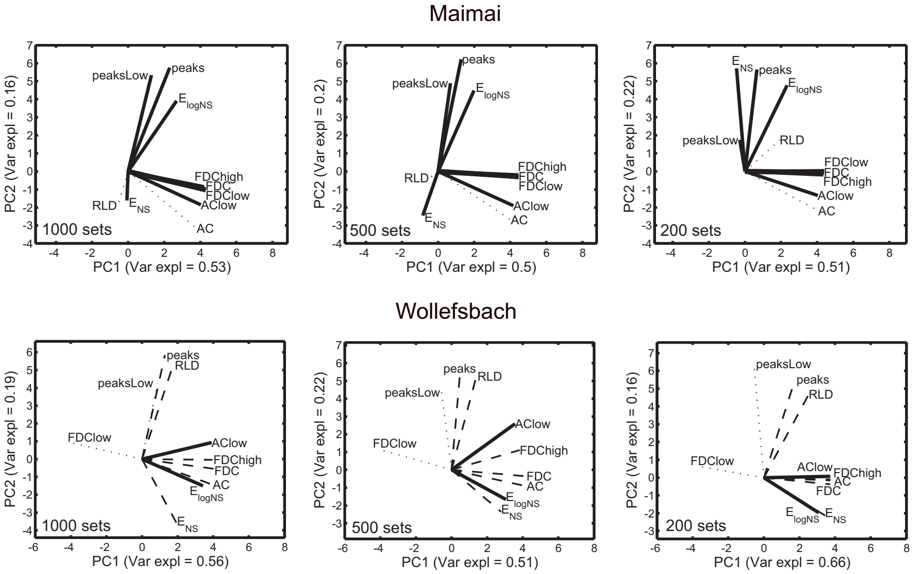


Fig. 11. PCA diagrams for M7 (left) and M8 (right) for both the calibration (top) and validation (bottom) period. M8 shows a higher consistency for the calibration period and a more consistent behaviour between the calibration and validation period.


Maimai

Wollefsbach

Fig. 12. Results PCA for M8 in Maimai (top) and Wollefsbach (bottom) for different number of parameter sets: 200(left)/500(middle)/1000(right). The difference between the diagrams with 1000 and 500 parameter sets is smaller than the difference between the diagrams with 500 and 200 parameter sets.