Hydrol. Earth Syst. Sci., 17, 1893–1912, 2013
www.hydrol-earth-syst-sci.net/17/1893/2013/
doi:10.5194/hess-17-1893-2013
© Author(s) 2013. CC Attribution 3.0 License.

Hydrology and Earth System Sciences
Open Access

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 of Environment and Agro-Biotechnologies, 4422 Belvaux, Luxembourg
[4]UNESCO-IHE Institute for Water Education, P.O. Box 3015, 2601 DA Delft, the Netherlands

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

Received: 15 October 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 16 November 2012
Revised: 19 March 2013 – Accepted: 17 April 2013 – Published: 21 May 2013

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. We define performance as 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 is defined as 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 visualise the performance and consistency of a model structure for the study catchments. With this framework performance and consistency are evaluated to identify which model structure suits a catchment better compared to other model structures. Until now the framework has only been based on a qualitative analysis and not yet on a quantitative analysis.

1 Introduction

One of the main purposes of scientific hydrology is to develop better predictive models of rainfall-runoff processes. To improve these models it is crucial 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; Hrachowitz et al., 2013b). Each hydrological model concept can be seen as a hypothesis of catchment behaviour (Savenije, 2009), and it is therefore a suitable tool to gain more knowledge about catchment processes. However, for models to be a suitable tool, it is very important that the "right" model is selected for a certain catchment. Due to differences between catchments (cf. Beven, 2000), 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. Typically, every model (structure) consists of several components, representing different runoff processes. Fenicia et al. (2011) describe the SUPERFLEX framework, similar to the FUSE framework (Clark et al., 2008), which can be used to configure such different model structures. With these frameworks it is possible to conveniently compare different model structures and their underlying hypotheses and hence use them as a learning tool to improve our understanding of the behaviour of individual catchments (Dunn et al., 2008; Hrachowitz et al., 2013b). 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; Andréassian et al., 2009).

It is increasingly acknowledged that model evaluation based on single objective optimisation, often performed with standard least squares optimisation, 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; Wagener et al., 2003; Schaefli and Gupta, 2007; Winsemius et al., 2009; Hrachowitz et al., 2013a). The use of specific characteristics of the hydrograph, hereafter referred to as hydrological signatures (Jothityangkoon et al., 2001), for the (multi-objective) evaluation of the performance of hydrological models can give even more information about the hydrological behaviour of the modelled catchments (Hrachowitz et al., 2013b). 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; Wagener and Montanari, 2011). Using hydrological signatures for model evaluation has some advantages and disadvantages in relation to traditional hydrograph fitting. The main disadvantage is that a signature represents a certain aspect of the catchment response at the expense of others. It is therefore necessary to consider multiple signatures to fully characterise the system behaviour. The main advantage, however, is that signatures are better interpretable in terms of underlying processes than aggregate performance measures, as they are constructed to reflect specific aspects of the system behaviour.

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. Thus, here 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. In the following section the evaluation framework will be described, followed by an application of the framework in a case study (see Sects. 3, 4 and 5).

2 Description of the framework

FARM (Framework to Assess 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. derivation of a posterior parameter distribution for the selected model structures and catchments (calibration);

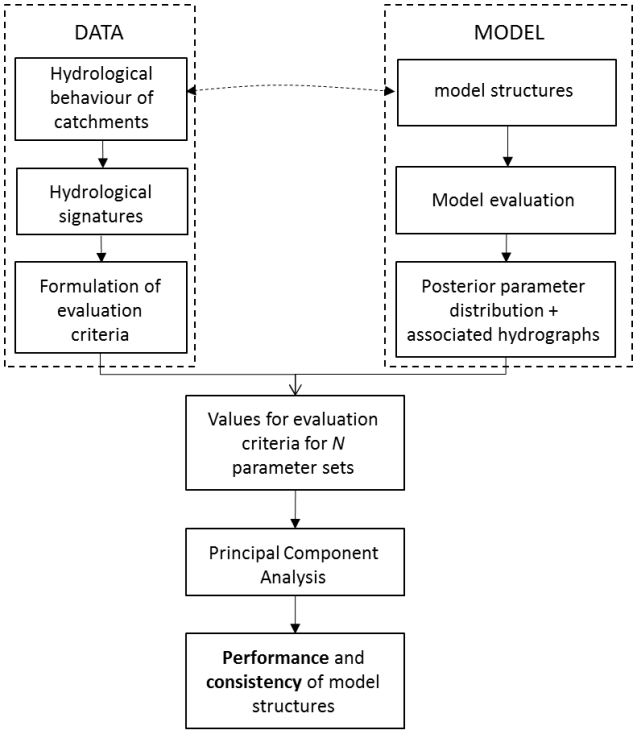


Fig. 1. Schematic overview of FARM to compare the performance and consistency of model structures with respect to hydrological signatures.

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 definitions in this paper; therefore, they are explained below.

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 indicative performance categories are defined: high, medium and low. A model structure is assumed to perform better when more evaluation criteria are in the highest performance category.

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 into the consistency of model structures for two reasons: first, a high consistency means that the model is capable of reproducing 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 (Wagener et al., 2003). 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. For a PCA the eigenvectors of a covariance matrix are determined. For many data sets most of the variance is described in the direction of a limited number of eigenvalues. By transforming the original axes towards the eigenvalues (principal components (PCs)), the original variable can be expressed in terms of the PCs (the variables have a certain loading on the PCs). More detailed descriptions on the principles of a PCA can be found in literature about multivariate analysis (e.g. Krzanowski, 2000; Härdle and Simar, 2003). In Appendix A an example can be found explaining the use of PCA for FARM. Note that here the vectors of the loadings are referred to as "vectors" thereafter.

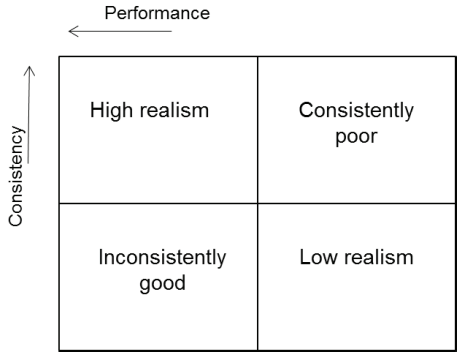


Fig. 2. Consequences for model structures for different combinations of performance and consistency, under the condition that the uncertainty of the input data is limited. The use of signatures for the evaluation of performance and consistency limits the influence of input uncertainty.

2.2.1 Input for PCAs

For FARM PCAs are used to explore the correlation structure between different evaluation criteria. A PCA is performed for each model structure in each catchment for N parameter sets. Here N is the number of parameter sets needed to reach convergence (see Sect. 4.4.1). The parameter sets are randomly sampled from a derived posterior parameter distribution. For these N samples all the evaluation criteria for the selected signatures 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 are generated from a normal distribution (Johnson and Wichern, 1998). Normality is especially important for the marginal distributions. Multivariate normality is of less importance if the PCA is used for dimension reduction, and thus as a mere descriptive tool as is the case with FARM (Jolliffe, 1986). If the marginals are not normally distributed, 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).

2.2.2 Interpretation of PCAs

The PCA represents two model characteristics: the performance and the consistency. The three indicative performance categories (see Sect. 4.1) 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, leading to a high consistency. 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. It is assumed that the signatures used for FARM are constructed to reflect different aspects of the hydrograph and, therefore, are not correlated by construction. The diagram which is the result of the PCA can be characterised by five general 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 and thus are all directly correlated to a certain degree (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 medium 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 either directly correlated or uncorrelated ("L-shaped" diagram) (Fig. 3d). This configuration has a medium degree of consistency as well, as there are two sets of evaluation criteria. The criteria within the different sets are highly and directly correlated, but the sets themselves are uncorrelated.

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.

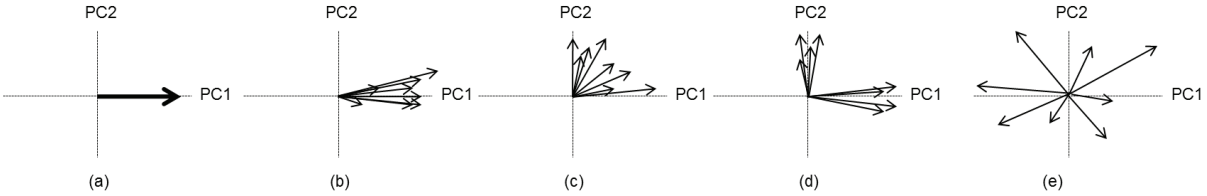


Fig. 3. Illustration of possible configurations for the PCA diagram: each vector represents an evaluation criterion (analysis is done per model structure). The axes are formed by the first two principal components (PCs). **(a)** represents a fully consistent model structure, **(e)** a fully inconsistent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

- Length of the vectors: the longer a vector is, the higher the loadings, and thus 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 are 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 autocorrelation coefficient. However, these signatures can in principle also be derived from other data sources, for example groundwater 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 ratio 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 is, 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

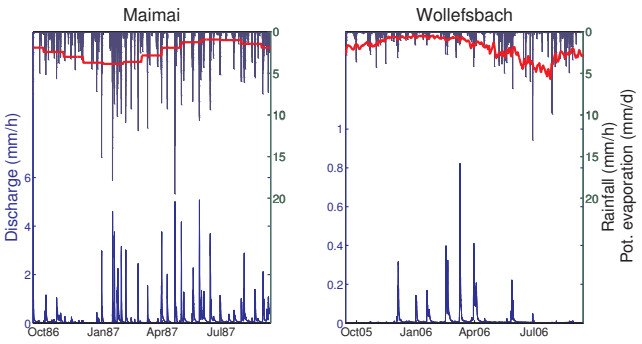


Fig. 4. Discharge, precipitation and potential evaporation data for Maimai and Wollefsbach catchments (discharge = bottom blue line, precipitation = top blue bars, potential evaporation = top red line). Note that the potential evaporation is presented in mm day^{-1} and the discharge and precipitation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

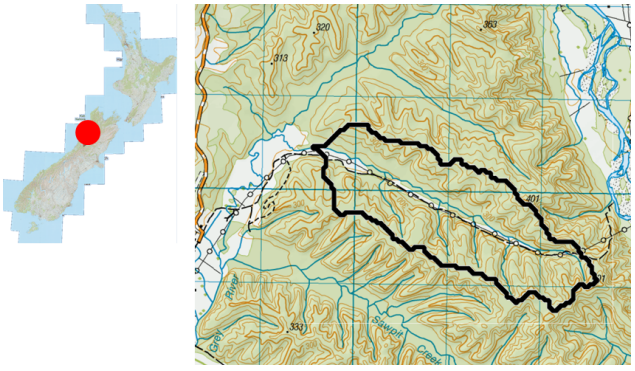


Fig. 5. Catchment area of the Maimai study area in New Zealand: the M8 catchment is one of the side branches of the main creek. Left: red dot indicates the location in New Zealand, right: topographic map of the Maimai study area with indicated the catchment boundary of the M8 catchment (source: http://www.topomap.co.nz/).

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 were used. The rainfall was measured with a recording rain gauge, which is located inside the catchment. The potential evaporation was estimated as described by Rowe et al. (1994). The first year of the data was used as a warm-up period; the last two years were 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 grassland and cropland. The discharge in the Wollefsbach is characterised 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 were used. The rainfall was measured with two tipping buckets, which are located inside the catchment, and the rainfall measurements were aggregated based on Thiessen polygons. The potential evaporation was estimated with the Penman equation (Penman, 1948). The first year of the data was used as a warm-up period, the following two years for calibration.

4 Methodology

In this section the specifics of FARM 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) and non-informative prior parameter distributions. A total of 1000 random samples were drawn from the posterior distributions, and all evaluation criteria were calculated for each random sample. The evaluation criteria distributions 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 three indicative performance categories for this case study are defined as follows:

- High (continuous and very bold vectors), when the maximum value for the evaluation criterion is higher than 0.8 and 90 % of the values for the evaluation criterion are higher than 0.65.
- Medium (dashed and bold vectors), when is the maximum value for the evaluation criterion is higher than 0.4 and 90 % of the values for the evaluation criterion are higher than 0.3.
- Low (dotted and thin vectors), for all other cases.

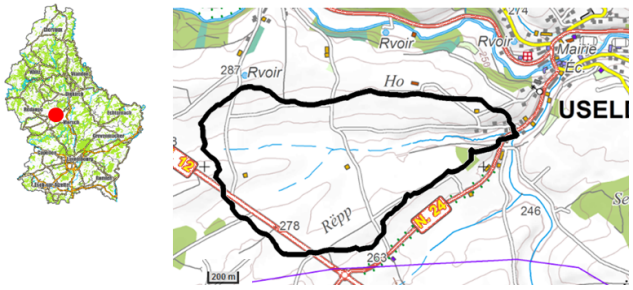


Fig. 6. Catchment area of the Wollefsbach catchment in Luxembourg. Left: red dot indicates the location in Luxembourg, right: topographic map of the Wollefsbach catchment with indicated the catchment boundary of the Wollefsbach catchment (source: http://eau.geoportail.lu/).

4.2 Hydrological signatures

The signatures 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 until September as low flow period and November until 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. Below the applied signatures are described in detail.

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 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) has been 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 considered 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 of the modelled and observed hydrograph are compared.

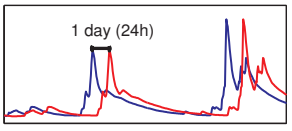
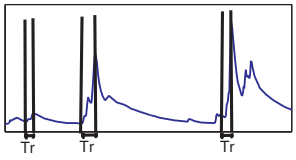
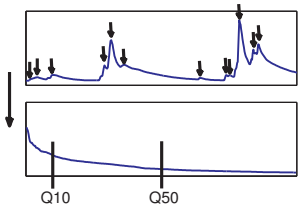
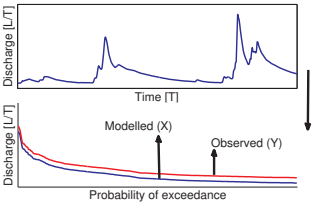
Low flow period (peaksLow)

The peak distribution during 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 uses 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 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

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, $\overline{Q}$ the average discharge, $X_{\mathrm{FDC},i}$ the value of the flow duration curve of the modelled discharge with i probability of exceedance, $Y_{\mathrm{FDC},i}$ the value of the flow duration curve of the observed discharge with i probability of exceedance, and $\overline{Y_{\mathrm{FDC}}}$ the average observed discharge).

Signature	Formula	Sketch
Autocorrelation	$\mathrm{AC} = \frac{\sum(Q_i-\overline{Q})(Q_{i+24}-\overline{Q})}{\sum(Q_i-\overline{Q})^2}$	
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	$\mathrm{FDC} = \frac{\sum(X_{\mathrm{FDC},i}-Y_{\mathrm{FDC},i})^2}{\sum(Y_{\mathrm{FDC},i}-\overline{Y_{\mathrm{FDC}}})^2}$	

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 taken 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 similar 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 either; it is expected that the result for the high flow period is similar 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 used because they (especially the Nash–Sutcliffe efficiency) are commonly used for the evaluation of hydrological models and are therefore suitable to use as a benchmark for this study.

4.3 Model structures

For this study nine flexible model structures are tested, and their performance and consistency are 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 (D. 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 Configuration 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 conditioning

The model conditioning 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 optimisation, 20 multi-starts are used. During the Markov chain Monte Carlo (MCMC) sampling, 5000 parameter sets were generated. The prior and posterior parameter ranges of the drawn samples are shown in Tables 2–4.

4.4 Plausibility checks

4.4.1 Sensitivity to number of parameter sets

In this case study 1000 parameter sets, randomly drawn from the posterior distribution, are used to construct the PCA. To investigate whether this number is sufficient for stable PCA patterns, 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 of the initial conditions (Descombes, 2012). To test whether the PCA is ergodic and to test if 1000 parameter sets are sufficient, a PCA was 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.

4.4.2 Independent test period

In addition to the sensitivity to the number of parameter sets, the obtained results can also be validated on a independent test period. It may be expected that a consistent model structure behaves similarly 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 additional year of data was available; for the Wollefsbach catchment two additional years of data were available. Both the performance and consistency are compared for the calibration and validation period.

5 Results

5.1 Maimai

The PCA results for the Maimai catchment of all model structures are shown in Fig. 8. The PCA results are based on the covariance matrix of the evaluation criteria. To illustrate what the PCA results are based on, the covariance matrix of model structure M8 in the Maimai is presented in Table 5.

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 (see Sects. 2.2.1 and 2.2.2), and only $E_{\log NS}$ is inversely correlated. Model structures M4 to M7 are much less consistent. Model structure M8 behaves differently 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, thus being of limited importance.

Table 2. Prior and posterior parameter 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)	$S_{uz\,max}$ (mm)	β (-)	P_{max} (mm h^{-1})	Fr (-)	D (-)	T_f (h)	K_r (1/h)	K_f (1/h)	K_s (1/h)
Prior Maimai	1.0×10^{-2}–6.0	1.0×10^{-1}–1.0×10^{4}	1.0×10^{-3}–10	1.0×10^{-6}–1.0×10^{2}	0–2.0×10^{-1}	0–1.0	1.0–5.0×10^{1}	5.0×10^{-3}–4.0	1.0×10^{-9}–10 1.0×10^{-8}–4.0 1.0×10^{-4}–4.0	5.0×10^{-4}–1.0×10^{-2}
Posterior ranges Maimai										
M1	–	–	–	–	–	–	–	–	2.3×10^{-2}–2.5×10^{-2}	–
M2	–	4.2×10^{1}–4.6×10^{1}	3.61–4.38	–	–	–	–	–	–	–
M3	5.95–6.00	1.0×10^{-1}–1.3×10^{-1}	–	5.8×10^{-1}–6.4×10^{-1}	–	–	–	–	3.5×10^{-2}–3.7×10^{-2}	–
M4	1.0×10^{-2}–1.5	6.2×10^{1}–8.3×10^{1}	4.3×10^{-1}–5.1×10^{-1}	–	–	–	–	–	3.9×10^{-2}–4.3×10^{-2}	–
M5	3.5×10^{-1}–1.8	6.0×10^{1}–8.3×10^{1}	3.6×10^{-1}–5.1×10^{-1}	–	–	–	1.0–1.3	–	3.8×10^{-2}–4.2×10^{-2}	–
M6	1.99–2.51	3.5×10^{1}–3.8×10^{1}	–	–	–	2.5×10^{-1}–2.7×10^{-1}	1.0–1.3	–	6.0×10^{-2}–6.5×10^{-2}	1.2×10^{-3}–1.4×10^{-3}
M7	3.1–3.9	4.0×10^{1}–4.4×10^{1}	6.2×10^{-1}–7.2×10^{-1}	–	–	2.5×10^{-1}–2.7×10^{-1}	1.0–1.3	–	5.9×10^{-2}–6.0×10^{-2}	1.2×10^{-3}–1.4×10^{-3}
M8	1.0×10^{-2}–2.5×10^{-2}	1.5×10^{1}–1.6×10^{1}	1.6–1.8	–	1.99×10^{-1}–2.00×10^{-1}	–	1.0–1.4	5.0×10^{-3}–5.7×10^{-3}	9.7×10^{-2}–1.0×10^{-1}	–
M9	4.7–6.0	5.0×10^{1}–6.0×10^{1}	6.3×10^{-1}–7.7×10^{-1}	–	1.98×10^{-1}–2.00×10^{-1}	2.9×10^{-1}–3.1×10^{-1}	1.0–2.3	2.0×10^{-1}–2.4×10^{-1}	2.8×10^{-2}–3.0×10^{-2}	9.2×10^{-4}–1.1×10^{-3}
Prior Wolfelsbach	1.0×10^{-2}–6.0	1.0×10^{-1}–1.0×10^{4}	1.0×10^{-3}–10	1.0×10^{-6}–1.0×10^{2}	0–0.2	0–1.0	1.0–5.0×10^{1}	5.0×10^{-2}–4.0	1.0×10^{-9}–10 1.0×10^{-8}–4.0 1.0×10^{-4}–4.0	5.0×10^{-5}–1.0×10^{-3}
Posterior ranges Wolfelsbach										
M1	–	–	–	–	–	–	–	–	1.5×10^{-4}–1.7×10^{-4}	–
M2	–	2.0×10^{1}–2.2×10^{1}	4.1–5.4	–	–	–	–	–	–	–
M3	1.0×10^{-2}–1.2×10^{-2}	4.4×10^{1}–4.6×10^{1}	–	1.0×10^{-1}–1.1×10^{-1}	–	–	–	–	4.2×10^{-2}–4.6×10^{-2}	–
M4	1.0×10^{-2}–1.1×10^{-2}	8.8×10^{1}–9.8×10^{1}	1.7–1.8	–	–	–	–	–	3.0×10^{-2}–3.4×10^{-2}	–
M5	1.0×10^{-2}–1.1×10^{-2}	9.0×10^{1}–1.0×10^{2}	1.7–1.8	–	–	–	2.4–3.4	–	3.2×10^{-2}–3.4×10^{-2}	–
M6	1.0×10^{-2}–1.2×10^{-2}	1.2×10^{2}–1.3×10^{2}	–	–	–	1.4×10^{-1}–1.6×10^{-1}	3.6–4.5	–	4.2×10^{-2}–4.7×10^{-2}	9.9×10^{-4}–1.0×10^{-3}
M7	1.0×10^{-2}–1.1×10^{-2}	8.5×10^{1}–9.3×10^{1}	1.3–1.4	–	–	1.5×10^{-1}–1.6×10^{-1}	3.5–4.4	–	4.5×10^{-2}–4.8×10^{-2}	9.9×10^{-4}–1.0×10^{-3}
M8	1.0×10^{-2}–1.1×10^{-2}	7.8×10^{1}–8.2×10^{1}	1.7–1.8	1.6×10^{-2}–1.9×10^{-2}	–	3.4–4.3	5.00×10^{-3}–5.03×10^{-3}	4.8 $\times 10^{-2}$–5.3×10^{-2}	–	
M9	1.0×10^{-2}–1.1×10^{-2}	8.3×10^{1}–9.1×10^{1}	1.3–1.4	6.9×10^{-7}–6.0×10^{-5}	1.5×10^{-1}–1.6×10^{-1}	3.6–4.4	5.4×10^{-2}–3.9	4.5×10^{-2}–4.9×10^{-2}	9.96×10^{-4}–1.00×10^{-3}	

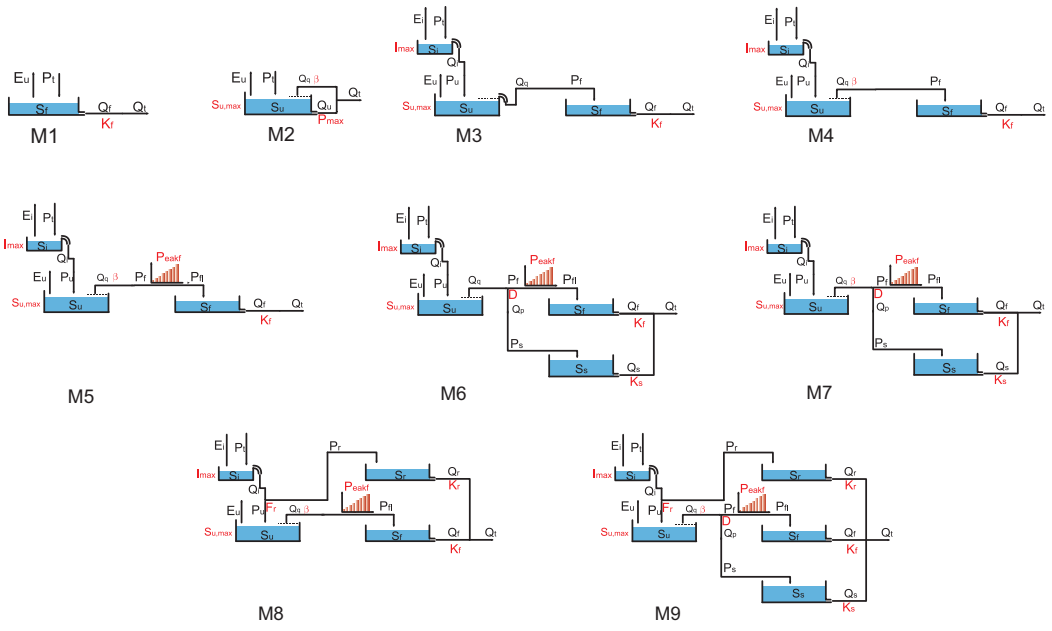


Fig. 7. Conceptual configurations of the flexible model structures used for this study.

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

x_1 (mm)	x_2 (mm)	x_3 (mm)	x_4 (h)
Prior			
1.0 to 2.0×10^3	-1.0×10^2 to 1.0×10^2	1.0 to 5.0×10^2	5.1×10^{-1} to 2.0×10^1
Posterior ranges Maimai			
1.2×10^2 to 1.3×10^2	-9.7×10^{-1} to -8.5×10^{-1}	1.7×10^1 to 1.9×10^1	6.1 to 6.7
Posterior ranges Wollefsbach			
9.2×10^1 to 1.2×10^2	-5.1×10^{-1} to -4.0×10^{-1}	5.5×10^1 to 5.9×10^1	1.9 to 2.0

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 for the Maimai than for 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 in both catchments M1 and M2 are consistently poor. Another observation is the difference between

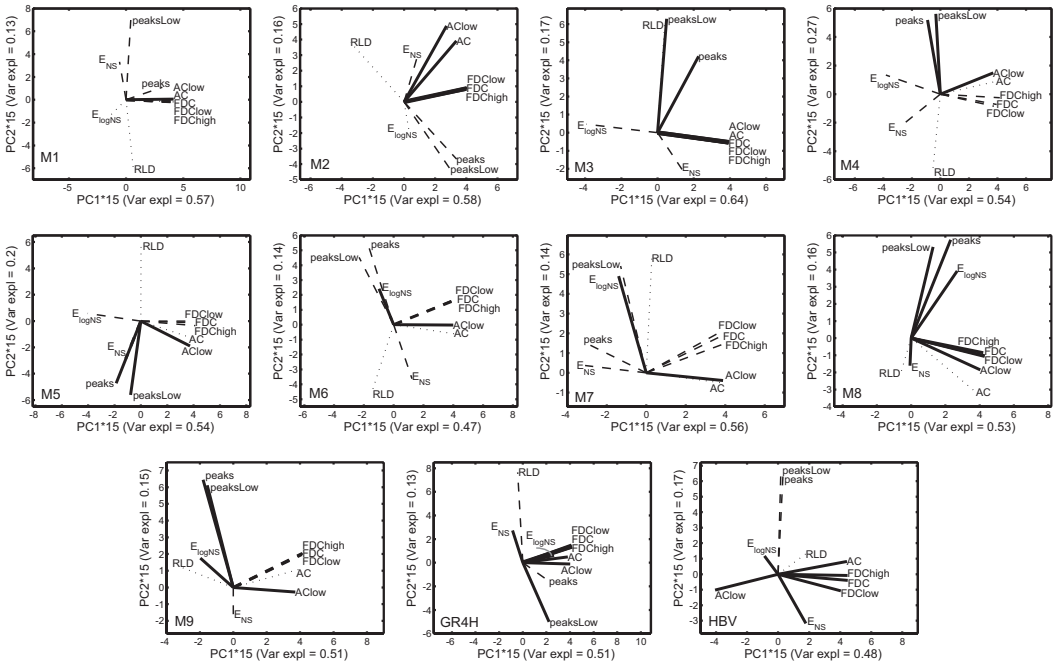


Fig. 8. Results for PCA for the Maimai catchment. Each figure represents one of the model structures. The figures are based on 1000 parameter sets. The principal components are dimensionless, because the ratios of specific signatures of the modelled and observed hydrographs are used to construct the evaluation criteria and these ratios are dimensionless. The total variance explained by these figures is the sum of the explained variance per PC.

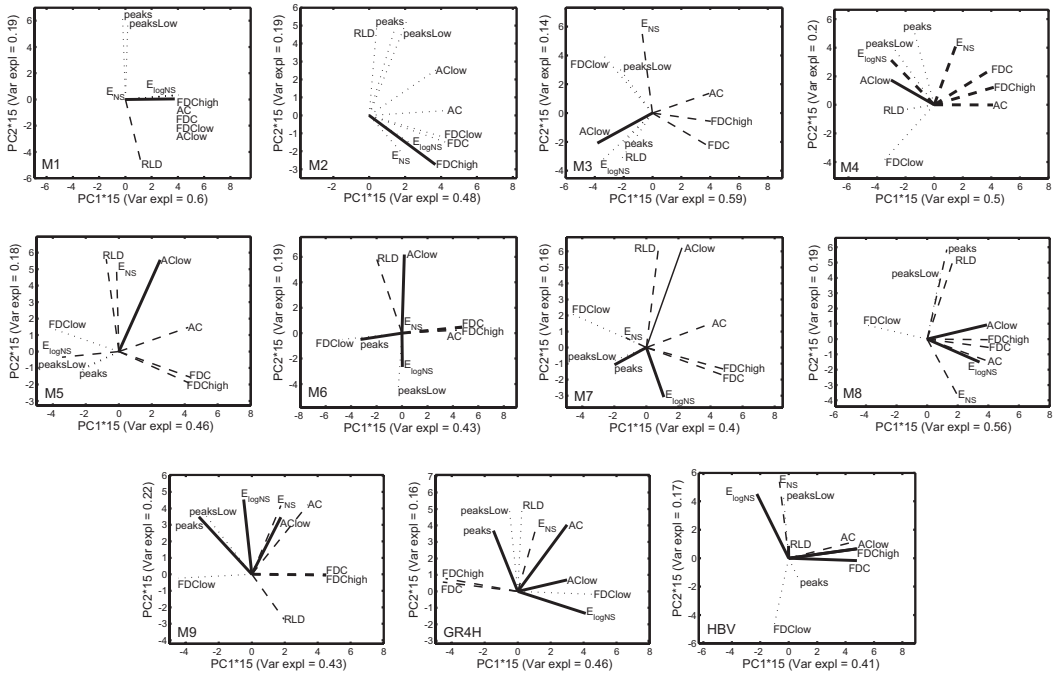


Fig. 9. Results for PCA for the Wollefsbach catchment. Each figure represents one of the model structures. The figures are based on 1000 parameter sets. The principal components are dimensionless, because the ratios of specific signatures of the modelled and observed hydrographs are used to construct the evaluation criteria and these ratios are dimensionless. The total variance explained by these figures is the sum of the explained variance per PC.

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

FC (mm)	β (–)	PWP* (mm)	L (mm)	k_0 (1/h)	k_1 (1/h)	k_{Perc} (1/h)	k_2 (1/h)	$I_{\max}$ (mm)
Prior								
1.0–5.0 × 10^2	1.0–10	1.0–5.0 × 10^2	5.0 × 10^{-2}–5.0 × 10^1	1.0 × 10^{-3}–3.0 × 10^1	1.0 × 10^{-4}–3.0 × 10^1	1.0 × 10^{-3}–3.0 × 10^1	1.0 × 10^{-3}–3.0 × 10^1	1.0 × 10^{-7}–10
Posterior ranges Maimai								
9.4 × 10^1–1.0 × 10^2	5.8–6.4	5.9 × 10^1–6.7 × 10^1	7.0 × 10^{-1}–8.0 × 10^{-1}	3.7 × 10^{-2}–4.0 × 10^{-2}	8.7 × 10^{-3}–1.1 × 10^{-2}	6.7 × 10^{-3}–7.4 × 10^{-3}	1.5 × 10^{-3}–1.7 × 10^{-3}	5.3–5.6
Posterior ranges Wollefsbach								
4.5 × 10^1–5.3 × 10^1	2.9–3.5	3.4 × 10^1–4.4 × 10^1	1.16 × 10^1–1.22 × 10^1	1.8 × 10^{-1}–2.1 × 10^{-1}	3.5 × 10^{-2}–3.8 × 10^{-2}	1.4 × 10^{-2}–1.5 × 10^{-2}	2.1 × 10^{-3}–2.7 × 10^{-3}	6.7–7.5

* PWP = Perm wilting point [mm m^{-1}] · soil thickness [m].

Table 5. Covariance matrix of the evaluation criteria for M8 in the Maimai catchment. The data are normally transformed, and therefore the variances within a evaluation criterion are very high. However, a small experiment shows that this does not influence the PCA results a lot.

	E_{NS}	E_{logNS}	AC	AClow	RLD	peaks	peaksLow	FDC	FDClow	FDChigh
E_{NS}	0.989	−0.187	0.026	−0.014	−0.068	−0.039	−0.017	0.014	−0.005	0.053
E_{logNS}	−0.187	0.989	0.225	0.410	−0.030	0.671	0.249	0.549	0.557	0.522
AC	0.026	0.225	0.989	0.921	−0.058	0.127	0.113	0.886	0.900	0.846
AClow	−0.014	0.410	0.921	0.989	−0.226	0.281	0.129	0.904	0.919	0.867
RLD	−0.068	−0.030	−0.058	−0.226	0.989	−0.146	−0.229	−0.054	−0.062	−0.057
peaks	−0.039	0.671	0.127	0.281	−0.146	0.989	0.531	0.433	0.413	0.458
peaksLow	−0.017	0.249	0.113	0.129	−0.229	0.531	0.989	0.187	0.188	0.198
FDC	0.014	0.549	0.886	0.904	−0.054	0.433	0.187	0.989	0.985	0.975
FDClow	−0.005	0.557	0.900	0.919	−0.062	0.413	0.188	0.985	0.989	0.961
FDChigh	0.053	0.522	0.846	0.867	−0.057	0.458	0.198	0.975	0.961	0.989

the catchments for M8 and M3. Both performance and consistency are much better for the Maimai, most likely because the catchment is small and homogeneous, and the climate is very humid/wet.

5.4 Plausibility of results

5.4.1 Sensitivity to number of parameter sets

Figure 11 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, and the results are compared with a visual inspection. Convergence is present to a 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.

5.4.2 Independent test period

In Fig. 12 an example is given to show the differences between two model structures with a more (M8) and a less (M7) comparable behaviour between the calibration and validation period for the Maimai catchment. A summary of the results of both catchments is presented in Tables 6 and 7. 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 slightly 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

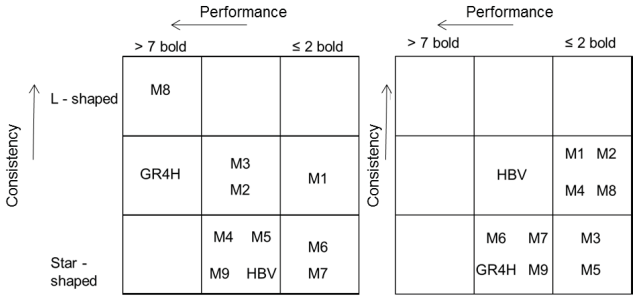


Fig. 10. Overview of the performance (columns) and consistency (rows) of the Maimai (left) and Wollefsbach (right). The middle row and column indicate a medium consistency and performance. The location of the model structures in this overview is determined based on visual inspection. There is only a difference between the squares: the exact position of a model structure within 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.

Table 6. 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 are 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.

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 Sect. 2:

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 of 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., 2013a; 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 certain runoff processes in the catchment. This can be seen in the Maimai and the Wollefsbach: the consistency and performance of M6, M7 and M9 are relatively low. These are the only flexible model structures with a groundwater reservoir, so possibly a groundwater reservoir is not important or incorrectly represented for both catchments. This is also in accordance with 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

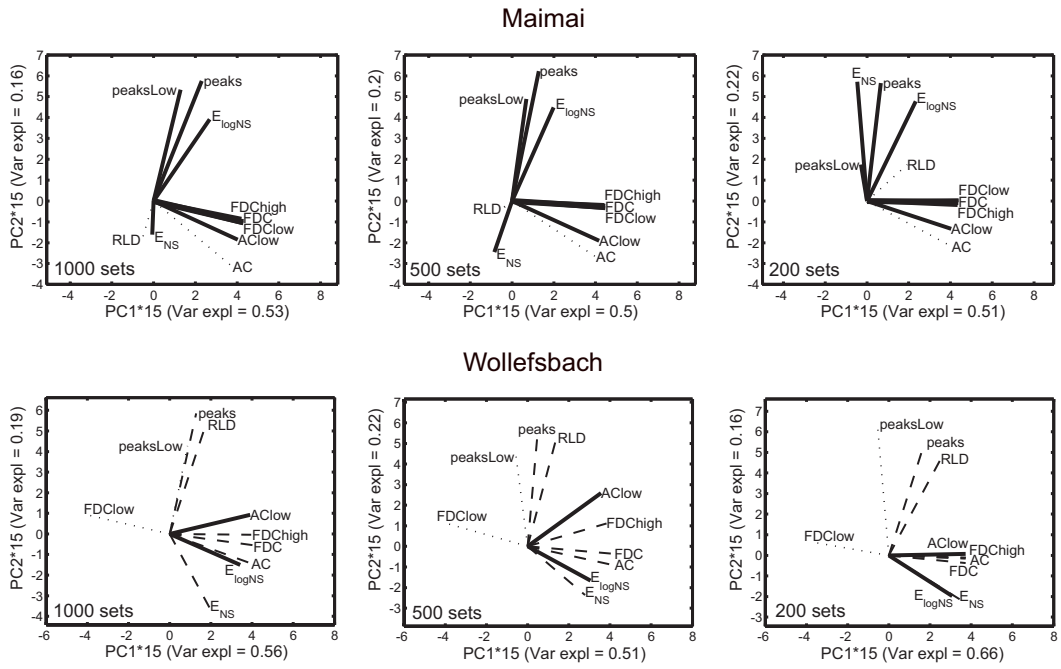


Fig. 11. Result PCA for M8 in Maimai (top) and Wollefsbach (bottom) for different number of parameter sets: 200(left)/500(middle)/1000(right). The difference between the diagrams with 1000 and 500 parameter sets is smaller than the difference between the diagrams with 500 and 200 parameter sets. The principal components are dimensionless.

Table 7. 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 are 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.

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.

It should be noted that FARM is meant to indicate which model structures have a higher performance and consistency than others. However, data errors can influence the performance and consistency of a model, and this possible influence is not explicitly included in FARM (Bárdossy and Singh, 2008). The influence of these errors will be different for different signatures. On the other hand, by using signatures, mainly the dynamics of the measured and observed hydrograph are taken into account. These dynamics are more likely to represent catchment behaviour and to be less sensitive to small measurement errors than evaluation criteria that compare each point of the hydrograph individually.

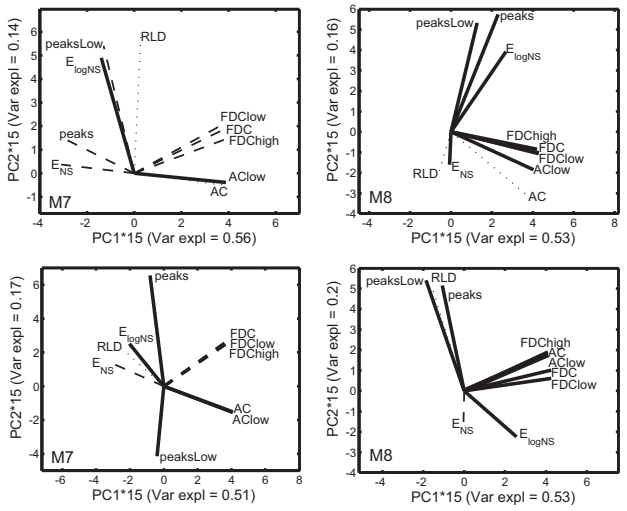


Fig. 12. PCA diagrams for M7 (left) and M8 (right) for both the calibration (top) and validation (bottom) period. The principal components are dimensionless. M8 shows a higher consistency for the calibration period and a more consistent behaviour between the calibration and validation period. Presented results are for the Maimai catchment.

6.2 Using the framework

The use of PCAs for model evaluation also has limitations. The main limitation may be the low variance explained by the first two principal components as obtained in this study. 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 reduce the high consistency, but will not make the model fully 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 reveal 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.

In addition to this limitation, also three other aspects influence the usefulness 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 – result 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.

The second is the different PCA results for the calibration and validation period for some model structures. In Sect. 5.4.2 it is shown that generally the model structures with a higher consistency behave more similarly in the calibration and validation period. However, this does not hold for all model structures, and the similarity between the calibration and validation period can be influenced by the length of the used time series as well. 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 influences the framework: 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 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 be weaker, as the signatures in the hydrograph are a mixture of different processes in different parts of the catchment. Therefore, it will be more difficult to relate them to specific dominant runoff processes. For larger scale catchments it may 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 and thus the realism 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 for different performance categories in the PCA diagrams 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, using eleven model structures. Comparison of the model structures for these catchments showed clear

Table A1. Covariance matrix, left: case 1; right: case 2.

Directly correlated ECs		Inversely correlated ECs	
0.124	0.125	0.070	−0.070
0.125	0.128	−0.070	0.095

Table A2. Eigenvalues and eigenvectors, left: case 1; right: case 2.

Directly correlated ECs		Inversely correlated ECs	
eigenvalues			
0.0004	0.251	0.011	0.154
eigenvectors			
−0.714	0.700	−0.768	−0.641
0.700	0.714	−0.641	0.768

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 may show a very low performance. Thus, it is important to take both aspects into account when evaluating whether a model structure suits a catchment.

Appendix A

Example PCA

A1 Introduction

This appendix gives a synthetic example of the use of principal component analysis for FARM. For FARM multiple evaluation criteria are used; however, for this example only two evaluation criteria are used, to be able to visualise the results. In this example two cases will be discussed:

1. two directly correlated evaluation criteria and
2. two indirectly correlated evaluation criteria.

A2 Basic principles of PCA

The PCA applied for FARM consists of several steps, which are listed below.

– The original data with values for evaluation criterion 1 (EC_1) and evaluation criterion 2 (EC_2) (first row of Fig. A1) are obtained.

– The covariance matrix of the evaluation criteria is calculated (Table A1).

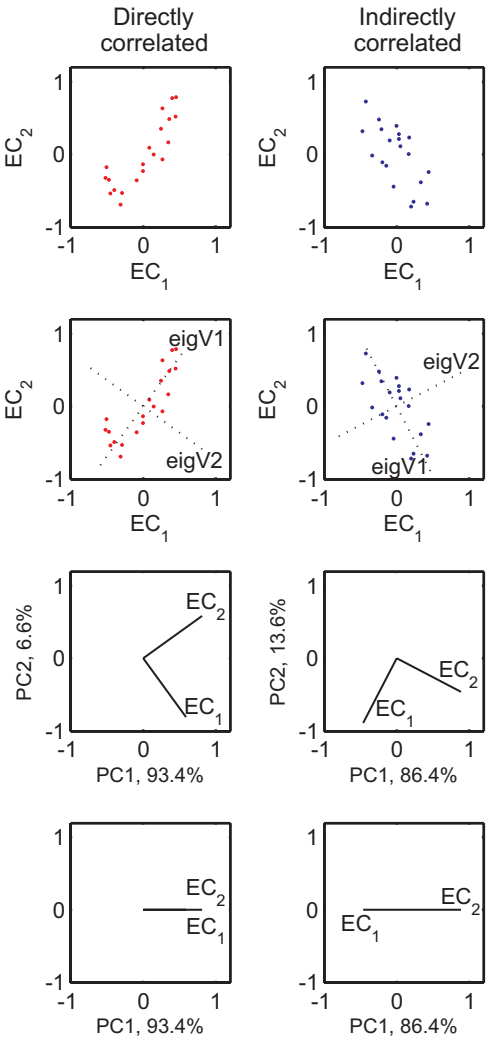


Fig. A1. Example showing the basic principles of principal component analysis and the application of PCA for FARM. left column: case 1; right column: case 2; first row: original data for EC_1 and EC_2; second row: original data with eigenvectors ("eigV", dotted); third row: ECs expressed in terms of PC1 and PC2; fourth row: ECs expressed in terms of PC1.

– Calculation of the eigenvalues and eigenvectors of the covariance matrix (Table A2) results in as many eigenvectors as evaluation criteria. The eigenvectors are orthogonal and the eigenvector with the largest eigenvalue describes the largest amount of variance in the data. The eigenvectors can be expressed in terms of EC_1 and EC_2 (second row of Fig. A1).

– Selection of the amount of principal components (PCs) (the eigenvalues) that are taken into account is done based on the variance explained by each PC. The explained variance per PC is the eigenvalue of that PC divided by the sum of all eigenvalues. In case of two

evaluation criteria, all PCs can be presented in a 2-D graph, as there are only 2 PCs.

– Expression of evaluation criteria is in terms of the principal components (in this case: PC1 and PC2). Therefore, the third row in Fig. A1 shows the loadings of both ECs on PC1 and PC2.

– The relative direction of the vectors can be used to identify the consistency. In Fig. A1 the relative direction of the vectors in both cases seems similar. However, for case 1 the vectors have an opposite loading on PC2, which represents a very small amount of the variance. For case 2 the vectors have an opposite loading on PC1, which describes the largest amount of the variance. Therefore, case 1 has a much higher consistency than case 2 (of course, for two ECs this can be easily deduced from the original data as well).

A3 Reduction of dimensions

In the step-wise approach described above, both PCs are kept. However, as can be seen in Fig. A1 PC1 describes a much larger part of the variance than PC2; thus, PC2 can be disregarded. The result of disregarding PC2 is shown in the last row of Fig. A1. A reduction of the dimensions leads for case 1 to two vectors in the same direction, while for case 2 it leads to two vectors with exactly opposite directions. This is because for case 1 the vectors had an opposite loading on PC2 and for case 2 an opposite loading on PC1.

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. We also thank Dmitri Kavetski for his help with the model structure configuration and conditioning. We thank professor András Bárdossy and an anonymous referee for their constructive comments to help clarify the use of PCA for FARM.

Edited by: R. Merz

References

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A.: HESS Opinions "Crash tests for a standardized evaluation of hydrological models", Hydrol. Earth Syst. Sci., 13, 1757–1764, doi:10.5194/hess-13-1757-2009, 2009.

Bárdossy, A. and Singh, S. K.: Robust estimation of hydrological model parameters, Hydrol. Earth Syst. Sci., 12, 1273–1283, doi:10.5194/hess-12-1273-2008, 2008.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrol. Earth Syst. Sci., 4, 203–213, doi:10.5194/hess-4-203-2000, 2000.

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 rainfall-runoff model, Hydrol. Process., 24, 2335–2348, doi:10.1002/hyp.7763, 2010.

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.

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.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008.

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.

Descombes, X.: Stochastic geometry for image analysis, ISTE Ltd, London, 2012.

Dunn, S. M., Freer, J., Weiler, M., Kirkby, M. J., Seibert, J., Quinn, P. F., Lischeid, G., Tetzlaff, D., and Soulsby, C.: Conceptualization in catchment modelling: simply learning?, Hydrol. Process., 22, 2389–2393, doi:10.1002/hyp.7070, 2008.

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

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.

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

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.

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.

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

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

Härdle, W. and Simar, L.: Applied multivariate statistical analysis, Springer-Verlag Berlin Heidelberg, 2003.

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

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.

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., and Soulsby, C.: What can flux tracking teach us about water age distribution patterns and their temporal dynamics?, Hydrol. Earth Syst. Sci., 17, 533–564, doi:10.5194/hess-17-533-2013, 2013a.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins, Hydrolog. Sci. J., online first, doi:10.1080/02626667.2013.803183, 2013b.

Johnson, R. A. and Wichern, D. W.: Applied multivariate statistical analysis, Prentice-Hall, Inc., Upper Saddle River, 1998.

Jolliffe, I. T.: Principal Component Analysis, Springer-Verlag, New York Inc., 1986.

Jothityangkoon, C., Sivapalan, M., and Farmer, D. L.: Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development, J. Hydrol., 254, 174–198, doi:10.1016/S0022-1694(01)00496-6, 2001.

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.

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, doi:10.1029/2005WR004362, 2006.

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

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.

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.

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.

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.

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.

Penman, H. L.: Natural evaporation from open water, bare soil and grass, P. Roy. Soc. London, 193, 120–146, 1948.

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

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.

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.

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.

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

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.