Supporting M5 model trees with sensitivity information derived from conceptual hydrological models

Carolina Massmann

ABSTRACT

The main objective of this paper is assessing the usefulness of parameter sensitivity information from conceptual hydrological models for data-driven models, an approach which might allow us to take advantage of the strengths of both data-based and process-based models. This study uses the parameter sensitivity of three widely used conceptual hydrological models (GR4J, Hymod and SAC-SMA) and combines them with M5 model trees. The study was carried out for three case studies dealing with different problems to which model trees are applied: one using model trees as error correctors and two case studies in which model trees were used as rainfall–runoff models and which differ in how the sensitivity information is used. The results show that sensitivity time series can improve the predictions of M5 model trees, especially when they do not include the time series of previous discharge as predictor variables. The use of parameter sensitivity information for clustering the time series resulted in model trees that had a structure consistent with the hydrological processes that were taking place in the considered cluster, indicating that the use of sensitivity indices could be a viable way of introducing hydrological knowledge into data-based models.

Key words | eFAST, M5 model trees, modular model, rainfall–runoff, parameter sensitivity, time series clustering

Carolina Massmann
Department of Civil Engineering,
University of Bristol,
University Walk,
BS8 1TR,
Bristol,
UK
E-mail: *carolina.massmann@bristol.ac.uk*

INTRODUCTION

While process-based and data-driven models tend to be described as antagonistic ways of doing hydrological modelling, it is often heard from process-based modellers that they would like to make better use of the information contained in their data while data-based modellers are motivated by finding ways for including hydrological knowledge into their models (Jain & Kumar 2009). It seems therefore worthwhile to test and develop methodologies that combine physically and data-based models for taking advantage of the strengths of both approaches. Such a combination of models can be achieved in three ways: following a modular approach, using ensembles or by implementing hybrid solutions (Abrahart *et al.* 2012).

Modular models require decomposing the problem into sub-problems that are modelled independently in a first step. The models are combined afterwards using either competitive switching or a sequential combination (Baker & Ellison 2008). The main difference between these methods is that in a sequential combination the outputs of one module are used as inputs for the next, while in competitive switching the models run independently. A good example of this approach in hydrological modelling is the development of different models for simulating floods and low flow periods which are subsequently brought together.

Ensembles combine models predicting the same target, i.e., solving the same problem. They can be created by varying the initial or boundary conditions of a model, its structure, parameter values, calibration/training approaches or the data used (considered periods, scenarios) (Baker & Ellison 2008). A representative value for the whole ensemble can be obtained by averaging techniques, for instance, the statistical mean or median, using weighted averages,

doi: 10.2166/hydro.2015.111

Bayesian model averaging or belief-based methods (Baker & Ellison 2008).

Hybrid models combine models of different types, each of which is necessary for addressing a specific task (Baker & Ellison 2008; Abrahart et al. 2012). Some examples in hydrology are artificial neural networks (ANN) used for bias correction in operational flow forecasting or using the output of a conceptual or physical model as input to an ANN (Abrahart et al. 2012), as done by De Vos et al. (2010), who used the soil moisture estimated by a hydrological model as input to a neural network model.

This study shows the result of a hybrid combination of a data-based and a conceptual hydrological model, which allows incorporating part of the hydrological knowledge embedded in the conceptual model into the data-based model. By this, it might increase the acceptance of data-based models, which are often seen as being less trustworthy than process-based models because they do not 'understand' the processes they are modelling.

The source of information used here for providing hydrological information to data-based models are sensitivity indices derived from conceptual models. Sensitivity analyses give information about how changes in the model inputs are reflected in the model outputs. While model inputs are understood in a broad sense, encompassing the boundary and initial conditions, as well as the input data and parameter values, sensitivity analyses are carried out in most cases with respect to the parameters. When the parameter sensitivity is calculated for each time step, a procedure called time-varying sensitivity analysis (Herman et al. 2013a, b; Massmann et al. 2014), it is possible to identify the parameters (and processes) which are determining the model output at each point in time (Massmann & Holzmann 2015). In this way, sensitivity indices provide information about the hydrological processes taking place in the catchment, something that could be an interesting piece of hydrologic information for data-driven models.

Two techniques for getting hybrid models and including hydrological information in data-based models are tested here (Corzo 2009):

(1) Including data about the hydrological processes (obtained with a process-based model) as predictor variables for the data-driven model. A good example of the usefulness of this is seen in rainfall–runoff modelling, where it is necessary to provide information about the catchment state to data-driven models because the response of a catchment to a given amount of precipitation depends on the antecedent moisture (Pathiraja et al. 2012). In most cases this is done using the previous discharge, but there are cases where also other variables have been used (e.g., moisture content in Casper et al. (2007) and de Vos et al. (2010)). In this case study the sensitivity indices are used as state indicators, so the performance of model trees without sensitivity information is compared to model trees using discharge as state variable and model trees using sensitivity indices as state variables.

(2) Using hydrological knowledge for separating the model space into areas of different hydrological characteristics and then developing a data-driven model for each part.

This paper uses the parameter sensitivity indices estimated for three widely used conceptual hydrological models and combines them with M5 model trees (MT), which are a type of data-based model. As data-based models are used in many different contexts (e.g., as error correctors, as rainfall–runoff models for simulation or forecasting), the study was carried out for three case studies:

- Error correction: where the data-based model considering parameter sensitivity time series as predictor variables was used for correcting the error of another model.
- Rainfall–runoff modelling: where the data-based model considering sensitivity time series was used for modelling the rainfall–runoff relationship.
- Modular model for rainfall–runoff modelling: where the input time series used by the data-driven model were separated into different clusters using sensitivity indices. A data-based model for rainfall–runoff modelling was then developed for each of these periods.

The main difference between the last two case studies is the way in which the sensitivity information was used. In the first case the indices were used directly as predictor variables, while in the second case the sensitivity indices were used for separating the input time series into groups with different hydrological characteristics each of which was modelled with a different model.

The main objective of this study was to find out whether parameter sensitivity information derived from conceptual hydrological models could be used for improving the performance of data-based models. Additionally, it was assessed whether there were differences in the performance of the data-based models when using sensitivity information derived from different conceptual models.

METHODOLOGY

Rainfall–runoff modelling with M5 model trees

Model trees can be easily understood and their structure can be visualized with simple tree plots, which is why they have been regarded as being more trustworthy than other data-based approaches (Elshorbagy et al. 2010a). They are also fast and can handle high-dimensional models with hundreds of attributes. Hydrological applications of model trees have been reported by Solomatine & Dulal (2003), Stravs et al. (2008), Pal & Deswal (2009), Elshorbagy et al. (2010b), Jung et al. (2010) and Sattari et al. (2013a, b). A comparison of their performance with other data-based modelling approaches was carried out by Elshorbagy et al. (2010b), who compared the results of artificial neural networks (ANNs), genetic programming (GP), evolutionary polynomial regression (EPR), support vector machines (SVM), M5 model trees (MT), K-nearest neighbours (K-nn) and multiple linear regression (MLR) techniques for different data sets including two rainfall–runoff case studies. In one of these case studies the only input to the system was the rainfall at the current and the three previous days, while the other also considered the preceding discharge. When rainfall was the only input, all data-based techniques performed similarly, suggesting that the amount of information was limiting the performance of the models and hiding possible differences between them. When also discharge was taken into account, it was found that GP, MT and EPR, followed by MLR techniques provided the best results. In general, the study showed that M5 model trees are among the techniques with the smallest deterioration when comparing the results for the training and testing data sets and that they had a good performance for many hydrological applications. This was confirmed by Solomatine & Dulal (2003) in a study looking at the capabilities of

model trees and ANNs for modelling the discharge for different lead times, where they found that both approaches gave comparable results.

Model trees were introduced by Quinlan (1992). They are similar to CART (Classification And Regression Trees) models in that both approaches separate the input space into different subareas and then assign a value to each of them. Their difference is that CART assigns a constant value to each subarea, while model trees can have a linear regression at the leaves (Quinlan 1992).

The algorithm determines first which input variable is best suited for splitting the data. This is done by finding the splitting criterion (i.e., the variable and its threshold value) that maximizes the standard deviation reduction (SDR):

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \tag{1}$$

where T represents all elements reaching a given node and sd(T) equals their standard deviation. T_1, T_2,… are subsets of T resulting from a given split arrangement. This splitting is done recursively for the subsets and terminates when the difference between the values reaching a node is small or when nodes are reached by only a few elements. This often results in large structures that need to be processed further for avoiding overfitting (Elshorbagy et al. 2010a). These steps are explained in Quinlan (1992) and are based on the estimation error (EE):

$$EE = \sum (|Oobs - Opred|)/n*(n+v)/(n-v) \tag{2}$$

with n being the number of training cases, v the number of considered inputs and $Oobs$ and $Opred$ the observed and predicted outputs, respectively. The correction factor in the EE formula (last term in Equation (2)) influences mostly the error of models having few training cases and a large number of inputs.

The first processing step is the construction of linear multivariate models for all nodes. These multivariate models consider the same inputs used by the model trees for allowing a comparison between them. The linear models are then simplified by looking at the estimation error and removing the variables that do not contribute to minimizing this error, a procedure that might leave only a

constant if all variables are removed. This elimination of variables usually increases the residuals, but since it can also reduce the value of the correction factor in Equation (2), it might result in overall smaller estimation errors.

Finally, it is possible to conduct a smoothing for adjusting the predicted values. This is important when the model is based on only few training cases and when different branches of the model predict very different values. If S is a sub-tree with a branch S_i, n_i equals the number of training cases for the branch S_i, $PV(S_i)$ is the predicted model value at S_i and $M(S)$ is the modelled value at S, then the predicted value $PV(S)$ is changed to:

$$PV(S) = (n_i * PV(S_i) + k * M(S))/(n_i + k) \qquad (3)$$

with k being the smoothing factor.

This study used the model tree implementation for R by Kuhn *et al.* (2014).

Hydrological models

Sacramento Soil Moisture Accounting Model (SAC-SMA)

This conceptual model, described in Burnash *et al.* (1973), conceptualizes the catchment as three vertical layers: the surface, the upper soil zone and the lower zone soil layers (Figure 1). The surface can be pervious, allowing the infiltration of water, permanently impervious (*PCTIM*) or only temporally impervious (*ADIMP*) (e.g., small reservoirs which are filled with water and act as impervious areas when the catchment gets wetter).

The soil layers consist of two storage types: the tension and the free water storages. They differ in the depletion mechanism, with water being removed from tension storages through evapotranspiration, while free water storages release water by lateral and downward movements. The upper soil zone has a tension water storage (*UZTWM*) representing the maximal amount of water that can be stored before drainage takes place. Only when this storage is filled is there water being routed to the upper free water storage (*UZFWM*). Interflow is proportional to the proportion of free water in this storage and is modelled using the *UZK* coefficient.

The lower soil zone consists also of a tension storage (*LZTWM*) and two free water storages, which determine

baseflow. The primary storage (*LZFPM*) is a slow draining storage with a depletion coefficient *LZPK*, while the supplementary storage (*LZFSM*) releases the water faster as a function of the coefficient *LZSK*.

Percolation depends on the amount of water in the upper zone as well as on the moisture deficit in the lower zone. It is described by the parameters *REXP* (characterizing the relationship between the moisture in the upper and lower zones) and *ZPERC* (maximum percolation rate). *PFREE* describes the proportion of water reaching the lower zone free water storage when the tension water storage is not yet filled, an effect observed due to heterogeneities in the rainfall and soil properties.

The parameter *RIVA* gives information about the percentage of the catchment covered by vegetation, a characteristic affecting evapotranspiration.

GR4J model

This model was developed in France for rainfall–runoff modelling at daily scales (Perrin *et al.* 2003). The parameter *X1* describes the size of the production storage receiving a variable proportion of the precipitation input (Figure 1). This storage water is the source for evapotranspiration and, together with the precipitation bypassing the production storage, forms the percolation outflow. This water is then split into two parts: 90% is processed by the unit hydrograph *UH1* followed by a non-linear routing storage, while the remaining 10% is processed by the unit hydrograph function *UH2*. The reason for splitting percolation into two parts is that this allows modelling the time lag by splitting the water output to different days. Both functions depend on the parameter *X4*, which determines the time base for the UH functions (*X4* and *2*X4* for *UH1* and *UH2*, respectively). The water processed by *UH1* serves as input to the routing storage of size *X3*. The model further considers water transfers from and to groundwater which are described by the parameter *X2*.

Hymod model

This model is based on Moore's probability distributed principle (Moore 1985) which is briefly explained in Bos & Vreng (2006). The model has five parameters that need to be
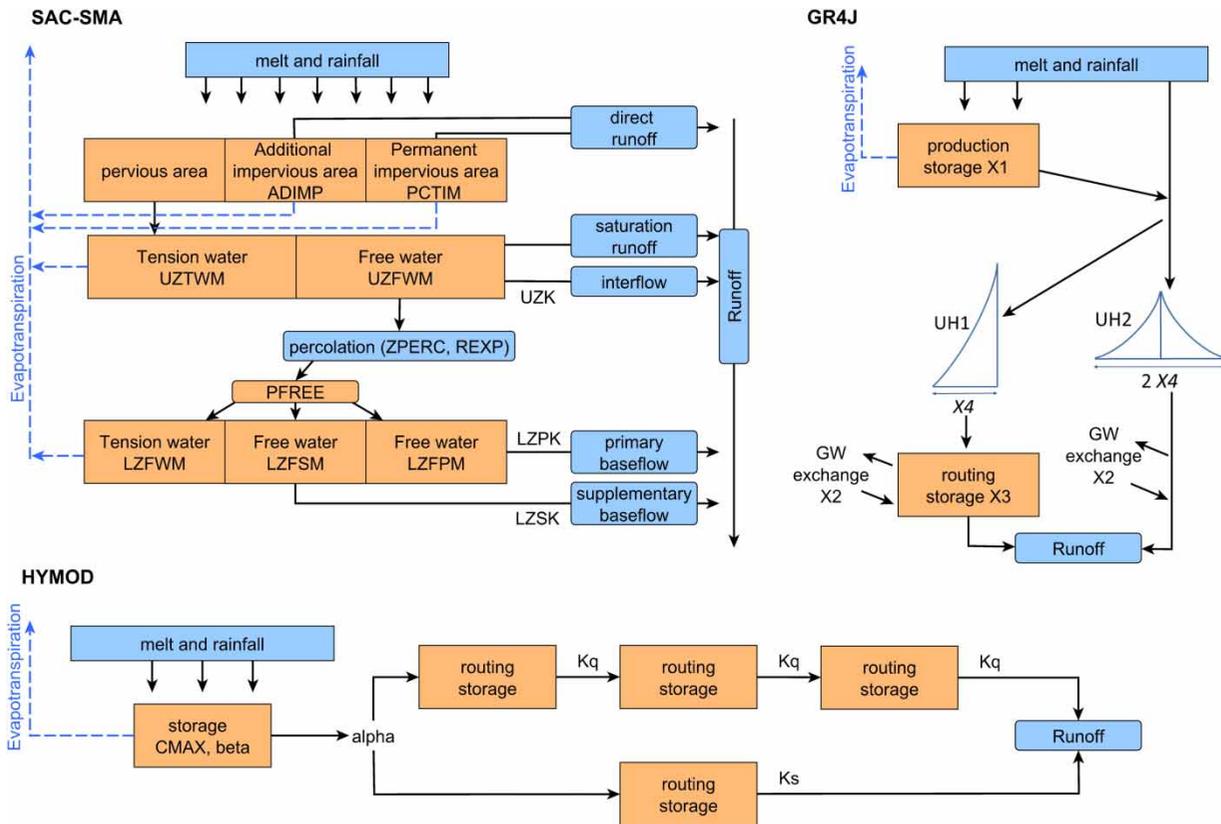
Figure 1 | Schematic structure of the GR4J, Hymod and SAC-SMA models.

supplied by the modeller (Figure 1). *Cmax* is the maximal storage capacity in the catchment and *beta* is a parameter describing the spatial distribution of this storage capacity. When precipitation input exceeds the storage capacity, the excess water is routed either to a set of three storages in series for modelling quick flow or to one storage modelling slow discharge. The recession coefficients of the quick and slow flow storages are *Kq* and *Ks*, respectively. The proportion of water allocated to each path is determined by the parameter *alpha*.

Sensitivity analysis with eFAST

Sensitivity indices indicate how changes in the model inputs are reflected in the model outputs. This study looks specifically at how the total discharge (model output) varies with changes in the model parameter values (model inputs) using a global sensitivity method, which means that the whole variation range of the parameters is considered. The method used here belongs to the group of variance decomposition-based sensitivity analysis methods in which the total output variance is apportioned to the different parameters and parameter combinations:

$$Var = \sum_{i}^{n} Var_i + \sum_{i}^{n} \sum_{j=i+1}^{n} Var_{ij} + \sum_{i}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} Var_{ijk} + \ldots + Var_{12\_n} \qquad (4)$$

with *Var* standing for the total output variance (e.g., discharge) observed when varying the parameters i, j and k. Var_i is the variance of the model output explained by variations in the parameter i and n equals the number of considered parameters. The terms on the right hand side of Equation (1) are used for defining the sensitivity indices:

First order sensitivity:

$$Si = Var_i/Var \qquad (5)$$

Total sensitivity:

$$STi = 1- \sim \mathrm{Var}_i/\mathrm{Var} \qquad (6)$$

with $\sim Var_i$ standing for the variance of the model output explained by all parameters except i. The first order sensitivity indicates therefore which proportion of the output variance is explained by a given parameter on its own, while the total sensitivity considers additionally the impact of the interactions with all other parameters.

This study used the eFAST (extended Fourier Amplitude Sensitivity Test) method which needs many model runs with different parameter sets for estimating the variances. The required parameter sets are obtained using a special sampling approach that assigns each parameter a different oscillation frequency. When the model is run with these parameter sets, it results in an output that can be decomposed into a number of curves oscillating with different frequencies using a Fourier transformation (Cukier *et al.* 1973, 1978). An analysis of the frequencies with which the output oscillates provides then insights about the influence of the different parameters at different periods. The FAST method was implemented based on the descriptions provided by Saltelli *et al.* (1999) and the code published by Ekström (2005).

Study catchment

The study was carried out for the French Broad catchment located in North Carolina (USA) which is included in the MOPEX (Model Parameter Estimation Experiment) database (Duan *et al.* 2006). The catchment has an area of 2,448 km^2 and a mean annual precipitation of 1,383 mm. The runoff coefficient equals 0.58 and the potential evapotranspiration reaches 819 mm y^{-1}. More information about this catchment can be found in Gan & Burges (2006) who evaluated the performance of the SAC-SMA model in 12 catchments (including this one) and van Werkhoven *et al.* (2008) who carried out sensitivity analyses for these catchments.

Study implementation

Figure 2 shows a diagram with the main steps carried out in this study. In the first place, the sensitivity indices for the parameters of the three conceptual models were estimated. The variable of interest was the total discharge, therefore, the obtained sensitivity indices describe how changes in the parameter values affect the total modelled discharge. The sensitivity analysis was done with the eFAST method using a sample of 10,000 parameter sets for each conceptual model, after it was confirmed that this sample size resulted in stable sensitivity indices. The parameter limits for the SAC-SMA model were taken from van Werkhoven *et al.* (2008) and Anderson (2002); the parameter ranges for the GR4J model come from Perrin *et al.* (2003); and the Hymod model used the limits provided by Bos & Vreng (2006).

The obtained total sensitivity indices were used for developing hybrid models combining model trees with conceptual hydrological models. For developing the model trees it was necessary to look first at the potential model inputs and decide on the predictor variables. This analysis was done using Spearman's rank correlation and focused on the correlation between the discharge and different variables (precipitation, evapotranspiration, temperature, sensitivity indices) considering various time lags. The results of this analysis indicated that precipitation had the highest correlation with the discharge when a 2-day lag was considered; the correlation for a 1-day lag was, however, almost as important. With respect to the remaining climatic variables, the analysis showed that the maximum temperature had a higher correlation with discharge than the potential evapotranspiration. Also here, the 2-day lag correlation was highest, slightly lower than the 1-day lag correlation. Since test results indicated, however, that temperature had no impact on the results of the model trees, it was not considered as a predictor variable for the model trees, leaving the precipitation at *t*-1 and *t*-2 as the only climatic inputs to the model.

In order to be able to model discharge time series adequately, data-based models require further information about the state of the system at each time step. This is usually provided through the discharge at previous days, but also other alternatives have been explored. This study uses three types of variables for describing the state of the system: the discharge of previous days, the antecedent precipitation index and sensitivity indices. The correlation analysis for the sensitivity indices showed that for about
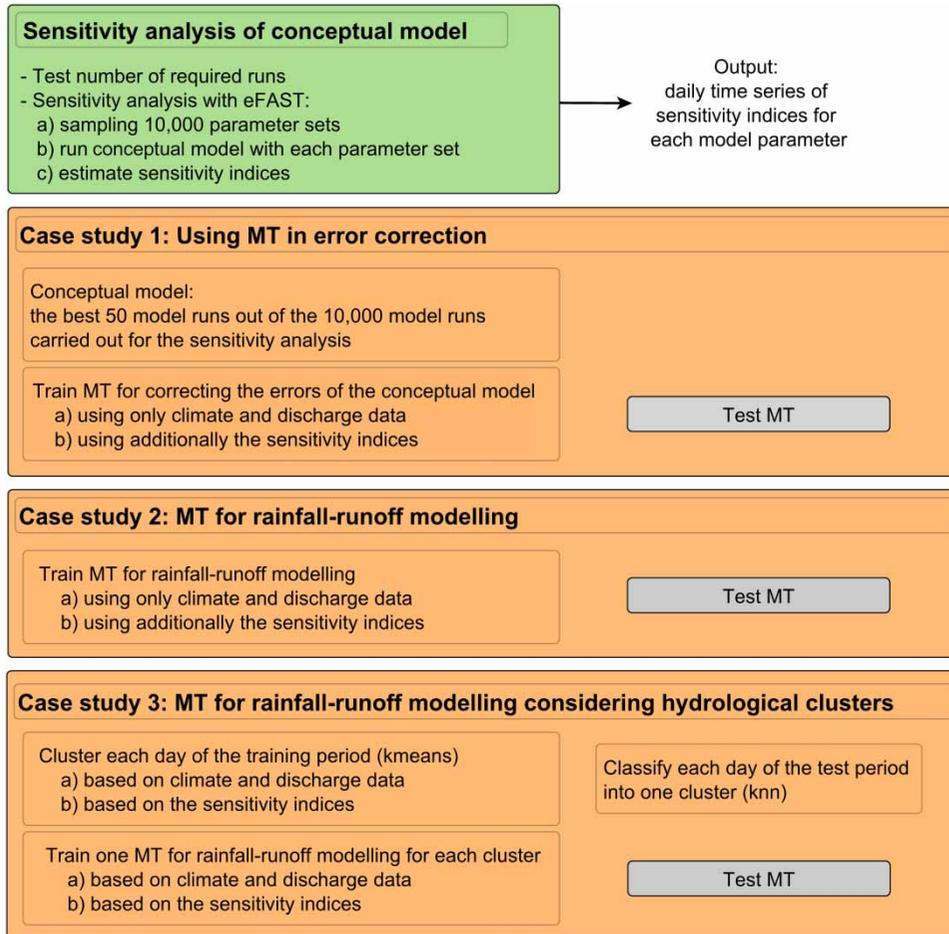
Figure 2 | Implementation of the study.

two-thirds of the parameters their parameter sensitivities had the highest correlation with the discharge at time t. For most other parameters the sensitivities with a 1-day lag were higher. It could be seen that the parameters with higher correlations at t-1 are in many cases related to the slower processes; for instance, the parameter Ks (Hymod), the two lower zone recession coefficients (SAC) and $ADIMP$ and $PCTIM$ which do have an important influence on the amount of water percolating into the soil. For each parameter the sensitivity time series with the lag that resulted in the highest correlation with the discharge was used as input to the model trees.

The combinations of hydroclimatic inputs considered in the case studies were:

No_Q: Precipitation at t-1 and t-2
Q_1: Precipitation at t-1 and t-2, discharge at t-1

Q_2: Precipitation at t-1 and t-2, discharge at t-2
AQ_2: Precipitation at t-1 and t-2, discharge at t-2, API_2
Q_4: Precipitation at t-1 and t-2, discharge at t-4
AQ_4: Precipitation at t-1 and t-2, discharge at t-4, API_4
API: API_2, API_4, API_8

with API_n standing for the antecedent precipitation index for the n previous days as defined by Kohler & Linsley (1951). The analyses were done once for the seven combinations of hydroclimatic inputs mentioned above and then repeated using additionally the time series of the sensitivity indices as inputs.

As model trees are used for different purposes, this study looked at the effect of combining MT with sensitivity information for two common applications of data-based modes: as error correctors (case study a) and as rainfall–runoff models (case studies b and c). For the rainfall–runoff MT,

the analysis looked at two ways of incorporating the hydro-logical information provided by the sensitivity indices into the model: first by using them as predictor variables and then by using them for separating the time series into different hydrological relevant clusters and then developing one MT for each cluster. More details about each case study are provided below.

(a) Error correction: Since the sensitivity analyses required 10,000 model runs for each model, it was decided to reuse the 50 model runs with the lowest mean absolute error (MAE) for this study. A model tree error correction model was then constructed for each model run. Since the conceptual models simulate the processes believed to take place in the catchment, there is no need to provide additional hydrological information to the MT in the form of sensitivity indices in this specific case. The approach presented here can be, however, applied for correcting other models which lack hydrological knowledge.

(b) Rainfall–runoff modelling: Rainfall–runoff modelling using model trees, where the results obtained with and without using sensitivity indices as predictor variables were compared.

For these two case studies (a and b), it was also analysed which of the sensitivity time series was able to provide most information to the model trees. For this, the model trees were fitted first using only the sensitivity time series for one parameter (besides the hydroclimatic input time series). The parameter whose sensitivity time series resulted in the lowest MAE was then added to the set of hydroclimatic input time series. This was repeated, including each time, one more sensitivity time series until no sensitivity time series could lower the MAE in the test period.

(c) Modular model trees: While model trees already follow a modular approach (since they split the input space into different areas each of which is modelled with a different equation), the splitting is not based on hydrological knowledge. This case study shows therefore how such knowledge can be incorporated by developing one MT for each hydrological cluster. This required first the clustering of each day of the training period using the k-means algorithm. The clustering was done first based on hydroclimatic information (precipitation at t-2, difference between the discharge at t-1 and t-2 and the antecedent precipitation during the last 4 days) and then repeated using the sensitivity

indices instead. The next step was to develop a rainfall–runoff MT for each cluster and to aggregate afterwards the results from the different clusters for obtaining the complete time series. For the test periods it was necessary to classify each day into one of the previously defined clusters, which was achieved with the knn algorithm.

For identifying how many clusters should be used, the model trees were developed for a varying number of clusters ranging from 2 to 20. It was seen that for the clustering based on the hydroclimatic variable set and the SAC sensitivity indices the total sum of squares within each cluster did not decrease much when more than 14 clusters were considered. For the GR4J and Hymod models 10 clusters were regarded as being adequate.

The model trees were developed for a training period (1 November 1952–31 October 1961) and then used for prediction during two test periods (1 November 1961–31 October 1967 and 1 November 1967–31 October 1973). Since the results did not reveal important differences between the two test periods, only the outcome of the analysis for the first test period is presented here.

RESULTS AND DISCUSSION

Error correction using model trees and sensitivity indices

This first case study used model trees for correcting the errors of the 50 runs with lowest mean absolute error (MAE) that were used for estimating the sensitivity indices. The first row in Table 1 shows the average MAE for the 50 model runs before the error correction. It is seen that the GR4J model has the lowest MAE, followed by the SAC and then the Hymod model.

Table 1 also shows that model trees not including sensitivity indices do not improve the estimations when previous discharge is not used as model input, as it is the case with the input sets No_Q and API. This is understandable since in these cases there is only information about the precipitation in previous days (directly as precipitation input or as a precipitation index) and it is not possible to relate this to specific discharge levels. When given information about the discharge at previous days (t-1, t-2 or t-4) model trees

Table 1 | Mean absolute error (MAE) for the 50 best parameter sets of each conceptual model (first row). MAE after correction using model trees with different sets of hydroclimatic (HC) inputs (rows 2–8) and after including also the sensitivity time series as predictors to the model trees (rows 9–15)

		GR4J		Hymod		SAC	
		Train	Test	Train	Test	Train	Test
	Initial	0.29	0.35	0.59	0.69	0.42	0.50
Model trees with HC data only	No_Q	0.64	0.76	0.65	0.74	0.63	0.73
	Q_1	0.23	0.30	0.24	0.30	0.23	0.31
	Q_2	0.30	0.41	0.30	0.40	0.29	0.40
	AQ_2	0.20	0.30	0.20	0.25	0.21	0.28
	Q_4	0.36	0.49	0.38	0.49	0.36	0.48
	AQ_4	0.26	0.37	0.24	0.31	0.26	0.36
	API	0.51	0.63	0.46	0.53	0.53	0.62
Model trees with sensitivity indices	No_Q	0.40	0.54	0.43	0.56	0.24	0.44
	Q_1	0.17	0.25	0.20	0.28	0.15	0.23
	Q_2	0.22	0.34	0.25	0.36	0.18	0.32
	AQ_2	0.17	0.27	0.16	0.23	0.16	0.26
	Q_4	0.25	0.38	0.30	0.43	0.21	0.35
	AQ_4	0.21	0.34	0.20	0.29	0.17	0.28
	API	0.32	0.48	0.29	0.39	0.20	0.38

are able to improve the discharge values provided by the SAC and Hymod models. The results of the GR4J model, which are already good, can only be improved with discharge information of the previous day (t-1) or when both the discharge at t-2 and API_2 are used as inputs. It can also be seen that the inclusion of the antecedent precipitation improves considerably the results (e.g., comparing Q_2 and Q_4 with AQ_2 and AQ_4, respectively) which are as good or better than the ones obtained with the t-1 discharge time series.

Considering the sensitivity indices as additional predictor variables in the model trees further improves their performance, resulting in almost all cases in models with lower MAE than the conceptual models. The only exceptions are observed for the GR4J model using the No_Q and API input sets, where the MAE for the GR4J could not be reduced even with model trees using sensitivity information.

A comparison between the conceptual models shows that the results of the corrected models are similar when only the hydroclimatic information is used as predictor variables for the model trees, indicating that model trees can compensate for the differences between the hydrological models. When including additionally the sensitivity indices

as predictors, the differences between the models become evident again.

Figure 3 shows the hydrographs for the GR4J model for the 50 conceptual model runs as well as for the corrected models with MT using the set of hydroclimatic variables Q_4. It is seen that the conceptual model has problems modelling the low flow period (around 15.5), a situation that is improved using the model trees error correction models. However, it is also seen that the model trees exhibit unexpected oscillations (e.g., during the low flow period). This indicates that, while average discharge values of the ensemble improve through the model tree correction, this does not necessarily hold for individual models or at a given point in time.

Finally, it was analysed which sensitivity indices are responsible for the largest decreases in the MAE. For the GR4J model parameter $X4$, describing the time base of the unit hydrograph was the most important in all cases. The remaining parameters also had an effect, but it was not possible to rank them. The sensitivities of the Hymod model showed a clear ranking and decrease in importance from *beta*, a*lpha*, *Kq* to *Cmax*. The parameter *Ks* did not decrease the MAE in the testing periods, and should thus not be included as predictor variable. For the SAC model, the important parameters were *REXP*, *LTWM* and *UZFWM*, while *UZTWM*, *RIVA* and *LZFSM* had no impact on the results of the test periods.

Rainfall–runoff modelling using model trees and sensitivity indices

The first two rows of Table 2 show the MAE when only hydroclimatic information is used as input to the model trees, while the following rows show the results when also the sensitivity indices are considered as predictors. The best results are obtained when the SAC sensitivity indices are included as input. This might be explained by the higher number of parameters in the SAC model which might result in more detailed information given to the model trees.

When comparing the runoff estimations of the SAC and Hymod models (first row Table 1) with the results of the model trees (Table 2), it is seen that model trees using sensitivity indices give better results than the conceptual models even if no discharge information is used as input. When no
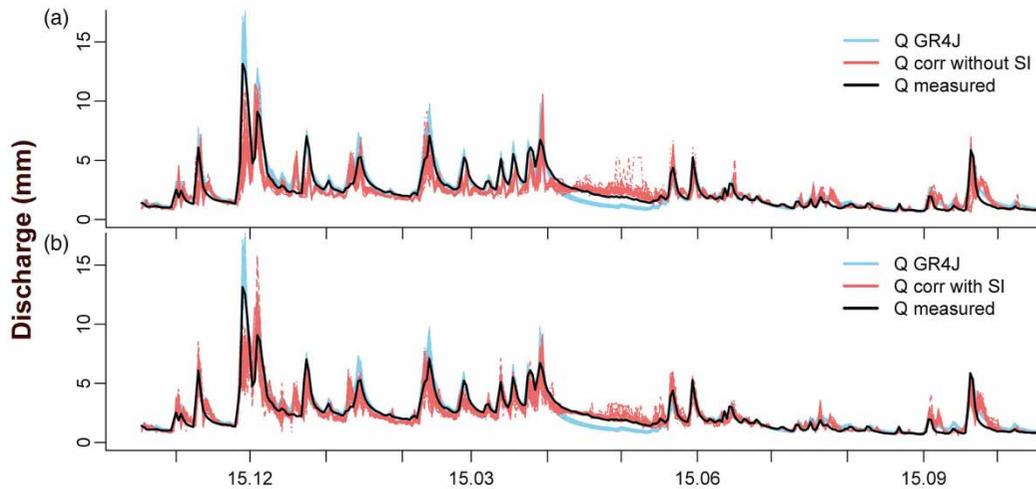
Figure 3 │ Measured discharge and discharge for the best 50 parameter sets of the GR4J model. Additionally, the MT corrected discharge using only the Q_4 set of inputs (a) and including the sensitivity time series (b) is shown.

Table 2 │ MAE of the model trees used as rainfall–runoff models. The columns show the results of using different sets of hydroclimatic inputs (HC) as predictor variables. The first two rows use only the HC inputs; rows 3–8 also includes sensitivity indices time series for the GR4J, Hymod and SAC models (rows 4–12) during training and testing

		No_Q	Q_1	Q_2	AQ_2	Q_4	AQ_4	API
HC only	Train	0.73	0.16	0.22	0.22	0.32	0.26	0.60
	Test	0.79	0.22	0.30	0.32	0.40	0.37	0.68
GR4J	Train	0.43	0.14	0.18	0.19	0.26	0.23	0.40
	Test	0.56	0.23	0.29	0.28	0.38	0.34	0.54
Hymod	Train	0.47	0.13	0.18	0.19	0.27	0.21	0.37
	Test	0.56	0.21	0.28	0.27	0.40	0.30	0.51
SAC	Train	0.27	0.13	0.17	0.17	0.23	0.19	0.25
	Test	0.49	0.20	0.28	0.29	0.35	0.31	0.47

sensitivity indices are used (first two rows in Table 2), model trees perform better only when they also include information about previous discharge. The use of the sensitivity indices of the GR4J model as input to the model trees gave better results only when also previous discharge data were used as model input regardless of the use of sensitivity information. This is probably the case because the GR4J conceptual model is able to provide already very good results. It must be mentioned, however, that the other two conceptual models might also have been able to provide better discharge estimations if their parameter values were calibrated (instead of selecting the models with lowest MAE from the model runs used for the sensitivity analysis), something that holds especially for the SAC model which has 14 parameters.

The impact of including or not sensitivity indices as predictor variables depends a lot on the considered set of hydroclimatic variables. The largest decrease in the error was observed for models not considering any previous discharge as input to the model trees (No_Q and API), while models using the discharge at t-1 and t-2 showed only a marginal decrease in the MAE. Since the previous discharge is a very good predictor of the current discharge, as shown by the high correlation coefficients obtained (0.95 for t-1 and 0.88 for t-2), there is not much additional information that could be provided by the sensitivity indices (Figure 4). On the contrary, when there is more limited information, the sensitivity information can increase the quality of the results and in these cases it is also possible to detect differences
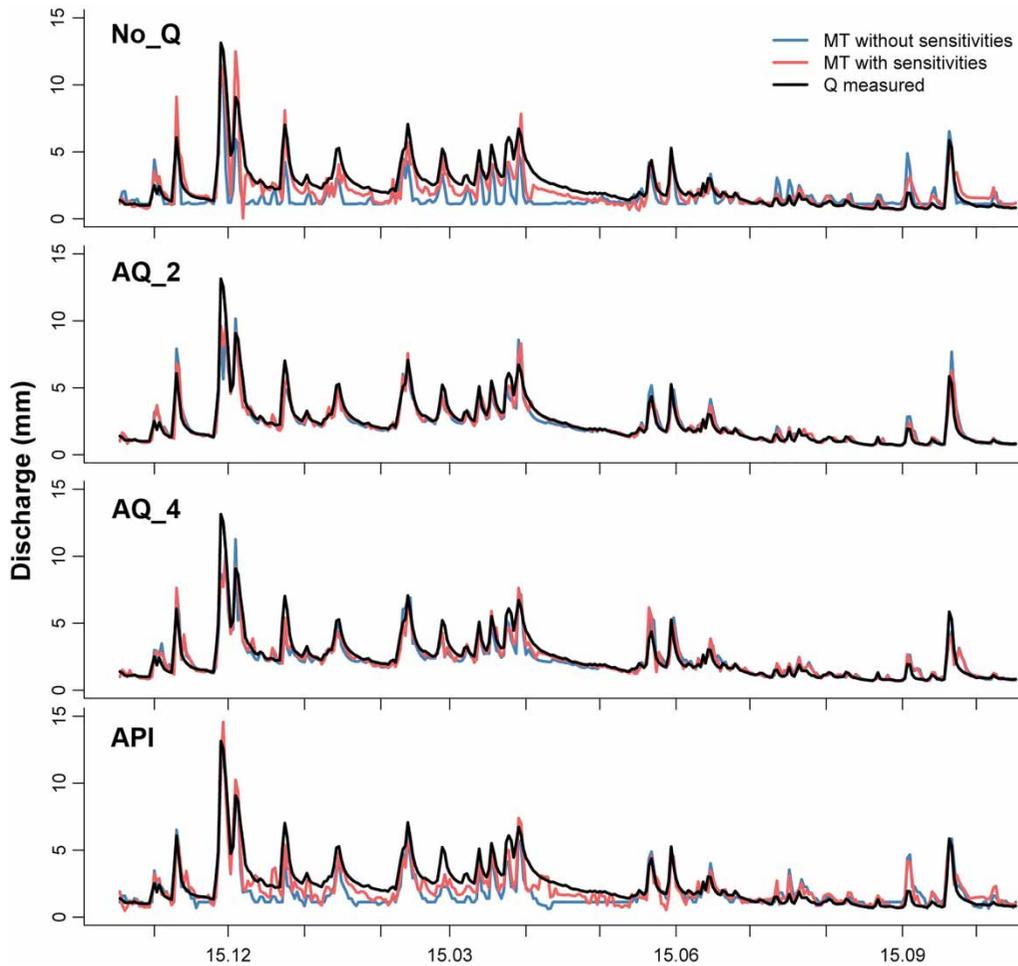
Figure 4 │ Discharge for the testing period 1 November 1961 to 31 October 1962 modelled with model trees using different combinations of hydroclimatic inputs and also including the sensitivity time series of the GR4J conceptual model as predictor variables in the model trees. The black line represents the measured discharge.

between the sensitivity information obtained with different conceptual models.

With respect to the impact of different parameters it was found that, similarly to the previous case study, the sensitivity of parameter $X4$ as predictor variable results in the largest reduction in the MAE for the GR4J model. The sensitivities of the other parameters also decrease the error during the testing period, but there was no clear ranking of their importance. The parameters of the Hymod model also show the same trend as in the previous case: the parameter *beta* has the largest impact, followed by the parameters *alpha* and *Cmax*. For the SAC model the ranking was not that clear, but the parameter *UZFWM* was among the most important ones. The other important

parameters showed large variations between the training and test periods.

Modularization using sensitivity indices-based clustering

Model trees separate the input space into different parts and then estimate the result for each part independently. One important advantage of this is that it is possible to use expert knowledge for defining areas which should be estimated with the same equations (Solomatine & Ostfeld 2008), allowing in this way the inclusion of hydrological knowledge into the models.

In this study, the time series were separated into different clusters defined first using only the hydroclimatic time series as input to the cluster algorithm and then repeated using the sensitivity indices instead. After that, one model tree was developed for each cluster.

Only the results for the Hymod model are presented here, but the main points are also valid for the results obtained when using the sensitivity indices for the other two conceptual models instead. Figure 5 shows the results of the clustering using the Hymod sensitivities for a 1.5-year period. Figure 6 shows the final model tree for cluster 4 as an example of what the structure of the model trees looks like and for explaining the terminology. It is seen that the input space is divided first according to the previous discharge (t-2) and in a second step according to the amount of precipitation during the previous day, Pp(t-1). These two variables define therefore the conditions in this model tree but also the usage for the two leaves on the left side since these are the variables used in the regression. Since the regression on the right side also makes use of the sensitivities of *alpha*, *beta*, *Cmax* and Pp(t-2), these are also usage parameters of the model developed for cluster 4.

Table 3 shows some statistics for each cluster's model tree. The first two columns give information about the MAE and the correlation coefficient, while the remaining columns provide information about the appearance of each predictor variable as a condition in the model tree or a regression variable (usage). It must be noted here that the results for cluster 4 do not agree with the model presented in Figure 6. This is because Figure 6 considers the final model without applying any smoothing, while Table 4 also includes the variables used in the course of the smoothing procedure.

As a next step, the internal structure of the model trees was looked at with the intention of linking it to the hydrological characteristics of the respective cluster. The fact that the model discharge was insensitive to changes in the Ks parameter (reflected in sensitivity indices close to zero on almost all days) explains why the sensitivity for the Ks parameter (slow recession coefficient) is not used by any model. This might be explained by the relatively high wetness of the catchment, suggesting that it never gets to the point when this parameter determines the outflow. This might also indicate that the parameter ranges considered for the sensitivity analysis were not suitable for modelling the wetness observed in this catchment.

The lowest MAE is observed for clusters 5 and 10, which also have the highest correlation coefficients. A glance at Figure 5 shows that these clusters characterize the recessions and low flow periods. An analysis of the cluster centroids indicated that cluster 5 has the highest value for the Kq sensitivities (0.75) and a relatively high value for
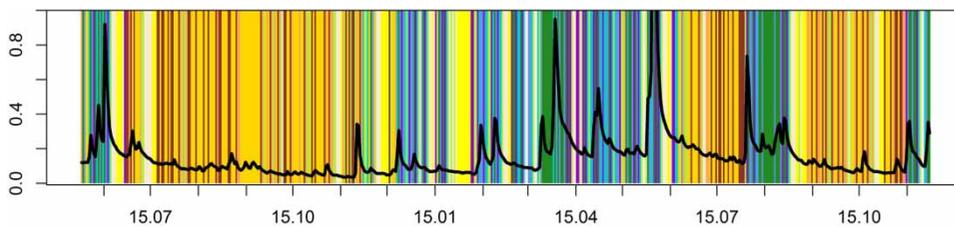

Figure 5 │ Distribution of the 10 clusters obtained by clustering the sensitivity time series of the Hymod model for the period between 1 June 1951 and 30 November 1952.
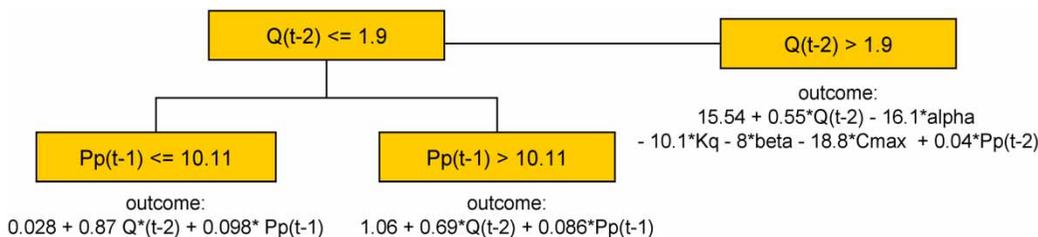

Figure 6 │ Structure of the model tree for cluster 4.

Table 3 | Mean absolute error (MAE) and correlation coefficient (CC) of modular model trees developed for each cluster. The last nine columns of the table show for each parameter its appearance in the model structure as a condition or a variable in the regressions

Clusters		MAE	CC	Q-2	Pp-1	Pp-2	API$_2$	SCmax	Salpha	Sbeta	SKs	SKq
1	Cond	0.45	0.94	0.99			0.96	0.24				
	Usage			0.95	0.95	0.95	0.04			0.18		0.18
2	Cond	0.28	0.9	1.00	0.59			0.56	0.03			
	Usage			0.99	0.99	0.61		0.80	0.32	0.32		0.32
3	Cond	0.43	0.95	1.00			1.00					
	Usage			1.00	1.00	1.00	0.06	0.34	0.47	0.47		0.47
4	Ccond	0.14	0.95	0.99	0.72							
	Usage			1.00	0.72	0.28		0.28	0.28	0.28		0.28
5	Cond	0.04	0.99	1.00					0.10			0.28
	Usage			1.00				0.18	0.18	0.26		0.18
6	Cond	0.16	0.92	0.89			0.97					
	Usage			1.00	1.00	1.00		0.87	0.87			
7	Cond	0.12	0.95	1.00			0.96					
	Usage			1.00	0.98	0.98	0.22	0.22	1.00	0.22		0.22
8	Cond	0.15	0.92	0.97		0.08	0.98					
	Usage			0.98	0.90	0.98	0.08		0.98			
9	Cond	0.35	0.96	0.99			1.00			0.07		
	Usage			1.00	0.91	0.95	0.15	0.91	0.87	0.97		0.43
10	Cond	0.06	0.99	1.00				0.35	0.84			
	Usage			1.00		0.05	0.16	0.23	0.24	0.24		0.07

the *alpha* sensitivity (0.41). Cluster 10 has a similar pattern, with the second highest sensitivity value for *Kq* (0.63) and slightly higher values for *alpha* (0.49). In Figure 5 it can be seen that cluster 5 is characterized by having drier days than cluster 10. This agrees with the higher sensitivity to *Kq* and lower sensitivity to *alpha* observed for cluster 5, since *alpha* has an impact just after the water leaves the soil storage when it is either routed to a fast or slower path, while *Kq* has an effect later on, once the water reaches the quick storages. The model trees for both clusters use the discharge from the previous days as the main variable defining the conditions (Table 3). For cluster 5 the sensitivity to *Kq* also plays a role, while for cluster 10 the sensitivity of *alpha* is also important for separating the input space into different areas. When looking at the variables considered in the regressions it is seen that the value of the previous discharge is by far the most important variable. This is not really surprising since these periods describe the recessions, when there is no precipitation.

The clusters with the worst results are clusters 1 and 3 which are related to high discharge periods. These are the clusters with the highest sensitivities for *alpha* (0.8 and

0.89 for the centroids of clusters 1 and 3, respectively), meaning that the parameter defining the discharge at this point is the one separating the water into a fast and slow path. The input space is separated into different areas using the previous discharge and the API. In contrast to the recession and low flow clusters, there is precipitation taking place in these clusters, which is why the moisture content (API) of the catchment is important. The regressions use, besides the API and previous discharge, the precipitation at the previous days. Cluster 3 also uses the sensitivities to the parameters *Cmax*, *alpha*, *beta* and *Kq* as regression variables.

Clusters 6 to 8 are mostly found at the end of summer and in autumn. Their centroids are similar, all with high sensitivities for *alpha* (0.61–0.78). It is also these clusters which have the highest sensitivities for *beta*. The main differences between the centroids of the clusters are in their sensitivities to *Kq*, with cluster 7 showing the highest sensitivities followed by cluster 8 and then cluster 6. For all three clusters the input space is separated into different areas using the values of the previous discharge and the API. The variables used in the regression are the previous discharge,

Table 4 | Mean absolute error of modular model trees. The clustering was done using hydroclimatic information (rows 1–7) or the sensitivity indices of the Hymod model (rows 8–14). The model trees were constructed using seven different sets of hydroclimatic information only (first column labelled 'No sensitivities') and then also including the sensitivity indices of the conceptual models

| | | Model trees | | | | | | | |
| | | No sensitivities | | GR4J sensitivities | | Hymod sensitivities | | SAC sensitivities | |
		Train	Test	Train	Test	Train	Test	Train	Test
Clustering HC	No Q	0.67	0.75	0.40	0.57	0.43	0.57	0.28	0.54
	Q-1	0.16	0.23	0.15	0.23	0.14	0.22	0.14	0.23
	Q-2	0.21	0.33	0.19	0.31	0.19	0.30	0.19	0.31
	AQ_2	0.21	0.31	0.19	0.31	0.19	0.30	0.19	0.32
	Q-4	0.29	0.42	0.22	0.38	0.23	0.39	0.20	0.36
	AQ_4	0.26	0.41	0.21	0.39	0.22	0.38	0.19	0.36
	API	0.63	0.72	0.40	0.57	0.38	0.57	0.27	0.49
Clustering Hymod	No Q	0.59	0.72			0.46	0.59		
	Q-1	0.14	0.24			0.12	0.24		
	Q-2	0.20	0.33			0.17	0.34		
	AQ_2	0.19	0.31			0.17	0.29		
	Q-4	0.30	0.45			0.26	0.41		
	AQ_4	0.23	0.37			0.20	0.35		
	API	0.52	0.65			0.36	0.55		

precipitation in the previous days, API, as well as the sensitivity to the parameter *alpha*. The differences between these clusters lie in the consideration of the sensitivity of *Cmax*, which is not considered in cluster 8, has some importance in cluster 7 and is very important in cluster 6. While it cannot be seen in Figure 5, it might be speculated that cluster 6 is observed when there is some precipitation, while clusters 7 and 8 group days with no precipitation. This could explain the lower importance of the previous discharge for cluster 6 in the conditions, as well as the higher impact that *Cmax* has in the regressions. For the other clusters the interpretation was not that clear, mainly because they characterize periods that are neither clearly wet nor dry.

Table 4 shows the results for the complete modular model trees. The results for GR4J are not presented, since they are similar to the Hymod results. This is also true for the SAC model as long as discharge information is used as predictor variable for the model trees. When this is not the case (No_Q and API), the SAC results are considerably better than the ones obtained with the sensitivities of the GR4J and Hymod models.

In general, it does not make a significant difference if the clustering is done using hydroclimatic or sensitivity time series. In a similar way, the differences between using and not using sensitivity indices are not that important when the model tree is given information about the previous discharge as predictor variable.

SUMMARY AND CONCLUSIONS

This study combines the sensitivity indices estimated for three conceptual models with M5 model trees to assess the usefulness of sensitivity indices as inputs to model trees. This allows the strengths of both type of models to be taken advantage of: the ability to extract efficiently information from the available time series (data-driven models) and the ability to describe the hydrological processes (conceptual models).

The results of the case studies indicate that sensitivity indices can support data-based models. The degree to which this occurs depends, however, on many factors. First it is important to note that the influence of adding sensitivity information is higher when the model trees are given information only sparingly. More specifically, it was seen that when the M5 model tree was given information about the previous discharge (at t-1, t-2 or t-4), the sensitivity information did not make a large difference in the results compared to the cases where this information was not

provided (No_Q and API). When discharge data were not provided to the model trees, the incorporation of sensitivity information resulted in an improvement of the discharge estimations. In most cases, however, it was seen that previous discharges were better predictor variables than the sensitivity indices.

It was also observed that the sensitivity information provided by the different conceptual models was not equivalent. This is not surprising since the conceptual models are based on different conceptualizations of the catchment and the processes taking place in it. It can be therefore expected that they do not provide the same information. The sensitivities estimated from the SAC model gave the best results for the training period for all hydroclimatic input sets (e.g., Q-1, Q-2, …) and all case studies (error correction, rainfall–runoff modeling, modular model trees). The picture becomes less clear though when the test period is considered, since the other models sometimes showed slightly better results. A clear ranking between the sensitivities of the GR4J and the Hymod models is not possible, since there is a large variability in their relative performances. In a similar way, the sensitivity information provided by different parameters is variable. There were large differences in the impact that individual parameters had on the capabilities of the model trees for estimating the discharge. For the SAC model it was further seen that the importance of a given parameter also depended on the case study considered.

Overall, it was interesting to see that all conceptual models, regardless of their conceptualization of the catchment, provided sensitivity information that could be used by the model trees. This stresses that it is an important task of the modeller to ensure that the conceptual model used provides a good representation of the processes taking place at the catchment.

Finally, it was explored whether sensitivity indices were able to provide useful hydrological information to the model trees for increasing the hydrological knowledge contained in the data-based model. An analysis of the model structure of each cluster showed that they agree in general with the variables that would be expected to be important based on the hydrologic characteristics of the clusters. This indicates that sensitivity indices could be used by model trees for increasing the amount of hydrological knowledge included in the model and it is probable that this could also be the

case for other data-based models. Sensitivity indices could, in this way, be a good alternative for increasing the confidence and acceptance of data-based models.

REFERENCES

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E. & Wilby, R. 2012 Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Prog. Phys. Geogr.* **36** (4), 480–513.

Anderson, E. 2002 *Calibration of conceptual hydrologic models for use in river forecasting*. Report NOAAs National Weather Service, Hydrology Laboratory Silver Spring, MD, USA. http://www.nws.noaa.gov/oh/hrl/modelcalibration/1.%20Calibration%20Process/1_Anderson_CalbManual.pdf (accessed 27 July 2014).

Baker, L. & Ellison, D. 2008 The wisdom of crowds – ensembles and modules in environmental modelling. *Geoderma* **147**, 1–7.

Bos, A. & Vreng, A. 2006 *Parameter optimization of the HYMOD model using SCEM-UA and MOSCEM-UA*. Report Computational Bio- and Physical Geography, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands.

Burnash, R. J. C., Ferral, R. L. & McGuire, R. A. 1973 *A generalized streamflow simulation system – Conceptual modeling for digital computers*. Technical Report, Joint Federal and State River Forecast Center, US National Weather Service and California Department of Water Resources, Sacramento, CA, USA, 204 pp.

Casper, M., Gemmar, P., Gronz, O., Johst, M. & Stüber, M. 2007 Fuzzy logic-based rainfall–runoff modelling using soil moisture measurements to represent system state. *Hydrolog. Sci. J.* **52** (3), 478–490.

Corzo Perez, G. A. 2009 Hybrid model for hydrological forecasting: Integration of data driven and conceptual modelling techniques. PhD dissertation, Delft University of Technology, Delft, The Netherlands.

Cukier, R. I., Fortuin, M. C., Shuler, K. E., Petschek, A. G. & Schaibly, J. H. 1973 Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. Theory. *J. Chem. Phys.* **59** (8), 3873–3878.

Cukier, R. I., Levine, H. B. & Shuler, K. E. 1978 Nonlinear sensitivity analysis of multiparameter model systems. *J. Comput. Phys.* **26**, 1–42.

de Vos, N. J., Rientjes, T. H. M. & Gupta, H. V. 2010 Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering. *Hydrol. Process.* **24** (20), 2840–2850.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. & Wood, E. F. 2006 Model Parameter Estimation Experiment (MOPEX):

An overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320** (1–2), 3–17.

Ekström, P.-A. 2005 *Eikos. A simulation toolbox for sensitivity analysis*. Uppsala University, Uppsala, Sweden. http://www.facilia.se/products/docs/Eikos_A_Simulation_toolbox_for_Sensitivity_Analysis.pdf (accessed 14 May 2012), p. 42.

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. P. 2010a Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 1: Concepts and methodology. *Hydrol. Earth Syst. Sci.* **14** (10), 1931–1941.

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. P. 2010b Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 2: Application. *Hydrol. Earth Syst. Sci.* **14** (10), 1943–1961.

Gan, T. Y. & Burges, S. J. 2006 Assessment of soil-based and calibrated parameters of the Sacramento model and parameter transferability. *J. Hydrol.* **320** (1–2), 117–131.

Herman, J. D., Kollat, J. B., Reed, P. M. & Wagener, T. 2013a From maps to movies: high-resolution time-varying sensitivity analysis for spatially distributed watershed models. *Hydrol. Earth Syst. Sci.* **17** (12), 5109–5125.

Herman, J. D., Reed, P. M. & Wagener, T. 2013b Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resour. Res.* **49** (3), 1400–1414.

Jain, A. & Kumar, S. 2009 Dissection of trained neural network hydrologic models for knowledge extraction. *Water Resour. Res.* **45** (7), W07420.

Jung, N.-C., Popescu, I., Kelderman, P., Solomatine, D. P. & Price, R. K. 2010 Application of model trees and other machine learning techniques for algal growth prediction in Yongdam reservoir, Republic of Korea. *J. Hydroinform.* **12** (3), 262–274.

Kohler, M. A. & Linsley, R. K. 1951 *Predicting the Runoff from Storm Rainfall*. Research Paper 34, Division of Climatological and Hydrologic services, US Weather Bureau, Washington, DC, USA.

Kuhn, M., Weston, S., Keefer, C. & Coulter, N. 2014 C code for Cubist by Ross Quinlan. Cubist: Rule- and Instance-Based Regression Modeling. R package version 0.0.18. http://CRAN.R-project.org/package=Cubist.

Massmann, C. & Holzmann, H. 2015 Analysing the sub-processes of a conceptual rainfall-runoff model using information about the parameter sensitivity and variance. *Environ. Model. Assess.* **20** (1), 41–53.

Massmann, C., Wagener, T. & Holzmann, H. 2014 A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales. *Environ. Modell. Softw.* **51**, 190–194.

Moore, R. J. 1985 The probability-distributed principle and runoff production at point and basin scales. *Hydrolog. Sci. J.* **30** (2), 273–297.

Pal, M. & Deswal, S. 2009 M5 model tree based modelling of reference evapotranspiration. *Hydrol. Process.* **23**, 1437–1443.

Pathiraja, S., Westra, S. & Sharma, A. 2012 Why continuous simulation? The role of antecedent moisture in design flood estimation. *Water Resour. Res.* **48** (6), doi:10.1029/2011WR010997.

Perrin, C., Michel, C. & Andréassian, V. 2003 Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **279** (1–4), 275–289.

Quinlan, J. R. 1992 Learning with continuous classes. In: *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence (AI'92)*, Tasmania, Australia, November 1992, Vol. 92 (A. Adams & L. Sterling, eds). World Scientific Publisher, Singapore, pp. 343–348.

Saltelli, A., Tarantola, S. & Chan, K. P.-S. 1999 A quantitative model independent method for global sensitivity analysis of model output. *Technometrics* **41** (1), 39–56.

Sattari, M. T., Pal, M., Apaydin, H. & Ozturc, F. 2013a M5 model tree application in daily river flow forecasting in Sohu Stream, Turkey. *Water Resour.* **40** (3), 233–242.

Sattari, M. T., Pal, M., Yurekli, K. Y. & Ulukara, A. 2013b M5 model trees and neural network based modelling of ET0 in Ankara, Turkey. *Turk. J. Eng. Environ. Sci.* **37**, 211–219.

Solomatine, D. P. & Dulal, K. N. 2003 Model trees as an alternative to neural networks in rainfall-runoff modelling. *Hydrolog. Sci. J.* **48** (3), 399–411.

Solomatine, D. P. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *J. Hydroinform.* **10** (1), 3–22.

Stravs, L., Brilly, M. & Sraj, M. 2008 Precipitation interception modelling using machine learning methods – The Dragonja River case study. In: *Practical Hydroinformatics. Computational Intelligence and Technological Developments in Water Applications* (R. J. Abrahart, L. M. See & D. P. Solomatine, eds). Springer, Berlin, Heidelberg, Germany, pp. 347–358.

van Werkhoven, K., Wagener, T., Reed, P. & Tang, Y. 2008 Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resour. Res.* **44** (1), W01429.