Hydrol. Earth Syst. Sci., 15, 2205–2227, 2011
www.hydrol-earth-syst-sci.net/15/2205/2011/
doi:10.5194/hess-15-2205-2011

Hydrology and Earth System Sciences

Calibration of hydrological models using flow-duration curves

I. K. Westerberg[1,2], J.-L. Guerrero[1,3], P. M. Younger[4,5], K. J. Beven[1,4], J. Seibert[6,7], S. Halldin[1], J. E. Freer[8], and C.-Y. Xu[9]

[1]Department of Earth Sciences, Uppsala University, Villavägen 16, 75236, Uppsala, Sweden
[2]IVL Swedish Environmental Research Institute, P.O. Box 210 60, 10031, Stockholm, Sweden
[3]Civil Engineering Department, National Autonomous University of Honduras, Blv. Suyapa Ciudad Universitaria, F. M. Tegucigalpa, Honduras
[4]Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK
[5]Environmental Research Consultant, 207 Eagle Heights J, Madison, WI, 53705, USA
[6]Department of Physical Geography and Quartenary Geology, Stockholm University, 10691, Stockholm, Sweden
[7]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland
[8]School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS, UK
[9]Department of Geosciences, University of Oslo, Postboks 1047 Blindern, 0316, Oslo, Norway

Received: 23 November 2010 – Published in Hydrol. Earth Syst. Sci. Discuss.: 9 December 2010
Revised: 30 June 2011 – Accepted: 1 July 2011 – Published: 14 July 2011

Abstract. The degree of belief we have in predictions from hydrologic models will normally depend on how well they can reproduce observations. Calibrations with traditional performance measures, such as the Nash-Sutcliffe model efficiency, are challenged by problems including: (1) uncertain discharge data, (2) variable sensitivity of different performance measures to different flow magnitudes, (3) influence of unknown input/output errors and (4) inability to evaluate model performance when observation time periods for discharge and model input data do not overlap. This paper explores a calibration method using flow-duration curves (FDCs) to address these problems. The method focuses on reproducing the observed discharge frequency distribution rather than the exact hydrograph. It consists of applying limits of acceptability for selected evaluation points (EPs) on the observed uncertain FDC in the extended GLUE approach. Two ways of selecting the EPs were tested – based on equal intervals of discharge and of volume of water. The method was tested and compared to a calibration using the traditional model efficiency for the daily four-parameter WASMOD model in the Paso La Ceiba catchment in Honduras and for Dynamic TOPMODEL evaluated at an hourly time scale for the Brue catchment in Great Britain. The volume method of selecting EPs gave the best results in both catchments with better calibrated slow flow, recession and evaporation than the other criteria. Observed and simulated time series of uncertain discharges agreed better for this method both in calibration and prediction in both catchments. An advantage with the method is that the rejection criterion is based on an estimation of the uncertainty in discharge data and that the EPs of the FDC can be chosen to reflect the aims of the modelling application, e.g. using more/less EPs at high/low flows. While the method appears less sensitive to epistemic input/output errors than previous use of limits of acceptability applied directly to the time series of discharge, it still requires a reasonable representation of the distribution of inputs. Additional constraints might therefore be required in catchments subject to snow and where peak-flow timing at sub-daily time scales is of high importance. The results suggest that the calibration method can be useful when observation time periods for discharge and model input data do not overlap. The method could also be suitable for calibration to regional FDCs while taking uncertainties in the hydrological model and data into account.

Correspondence to: I. K. Westerberg
(ida.westerberg@hyd.uu.se)

1 Introduction

Hydrologic models are used as a basis for decision making about management of water resources with important consequences for sectors such as agriculture, land planning, hydropower and water supply. The degree of belief we have in model predictions will normally be dependent on how well the model can reproduce observations. The choice of the likelihood measure that measures the agreement between simulated and observed data is therefore an important choice

in any modelling study. The definition of an appropriate likelihood measure is not, however, simple. Where all sources of uncertainty can be treated as if they are aleatory in nature, then a number of frameworks exist for the definition of formal statistical likelihoods (e.g. Liu and Gupta, 2007; Schoups and Vrugt, 2010; Renard et al., 2010). Where epistemic errors are important, however, treating all uncertainties *as if* they are aleatory will generally lead to overconditioning of posterior parameter distributions (Beven, 2006, 2010; Beven et al., 2008), particularly if some periods of data are disinformative (Beven and Westerberg, 2011; Beven et al., 2011). Thus, there may be scope for using other forms of likelihood or belief measures in hydrological modelling. Such informal likelihood measures have been defined based on limits of acceptability defined from evaluation-data uncertainty (Blazkova and Beven, 2009; Krueger et al., 2010; Liu et al., 2009) but also based on traditional performance measures (Freer et al., 2003). One of the most widely used performance measures in hydrology is the Nash-Sutcliffe model efficiency (R_{eff}). It is calculated as 1.0 minus the normalisation of the mean squared error by the variance of the observed data and varies between minus infinity to 1.0 (Nash and Sutcliffe, 1970). How appropriate this criterion is for measuring goodness of fit, as well as what is an acceptable R_{eff}-value, has been much debated in the literature (Krause et al., 2005; Legates and McCabe, 1999; Seibert, 2001; Criss and Winston, 2008; Smith et al., 2008; Gupta et al., 2009). Decompositions of R_{eff} have highlighted several problems associated with this criterion in model calibration (Gupta et al., 2009; Smith et al., 2008). Gupta et al. (2009) present a decomposition of R_{eff} into three components representing bias, variability and correlation and conclude that the variability has to be underestimated to maximize R_{eff} and that runoff peaks tend to be underestimated when maximizing R_{eff}. They, together with many other authors (Garrick et al., 1978; Refsgaard and Knudsen, 1996; Legates and McCabe, 1999; Seibert, 2001; Krause et al., 2005; Schaefli and Gupta, 2007; McMillan and Clark, 2009) propose modified versions of the Nash-Sutcliffe criterion or other performance measures to overcome some of these problems. However many of the problems in using lumped global performance measures remain, for instance that the measure often is more influenced by the performance at certain flow magnitudes such as high or low flows. This issue has been addressed in multi-criteria approaches where different aspects of the fit between simulated and observed discharge are evaluated. A combination of several criteria then allows an assessment of model performance with respect to the different aspects of the hydrograph (e.g. Gupta et al., 1998). Boyle et al. (2000) and later Wagener et al. (2001), suggest distinguishing between three parts of the hydrograph (driven quick flow (during events), non-driven quick flow and slow flow) and to then calculate the performance measure separately for each flow type. In a related approach, Freer et al. (2003) used several performance measures for a multi-criteria calibration in a Generalised Likelihood Uncertainty Estimation (GLUE) framework where they differentiated the dataset by season. They found no consistently identified parameters for Dynamic TOPMODEL that could represent the range of processes between seasons in the studied watershed. However, these approaches have not generally taken any explicit account of uncertainty in the observed input and evaluation data.

Hydrologic models are simplified conceptualisations of the hydrologic processes in a catchment. Such simplifications will necessarily lead to errors in the way the structure of the model represents the real-world hydrologic processes (Beven, 1989, 2009; Grayson et al., 1992; McDonnell, 2003). The temporal and spatial scales of the measured input data are also incommensurate with both the real-world quantities and the scale of the model. This source of error must be considered together with pure measurement errors (e.g. as a result of lack of calibration or accuracy of the measurement equipment) in input data. Such errors can lead to substantial uncertainty of an epistemic (knowledge) type, e.g. if there are no rain gauges in the only part of the catchment where it rains, this will create an error that is difficult or impossible to characterise in an error model. This type of uncertainty resulting from non-stationary epistemic errors should be expected in most datasets used for hydrological modelling because of the difficulties in measuring the components of the water balance for a catchment. As discussed by Beven and Westerberg (2011), such errors, if significant, should be expected to have a disinformative effect on model calibration. They suggest that the best strategy to deal with such disinformative periods of data would be to identify and remove them from the dataset independently of the model, but recognise that this identification will be difficult in many cases because of the uncertainties in the measured data. An alternative strategy could therefore be to develop model evaluation criteria that are robust to moderate disinformation to make sure that models are rejected for the right reason – i.e. poor model structure and not disinformative data. Model parameters need to be inversely estimated from data in calibration which will involve substantial uncertainty because of the effect of the types of errors discussed here and their interactions. On top of this, the performance measure that is used for the model calibration will influence which parameter-value sets are identified as being acceptable given the uncertainties in the modelling application (see e.g. Freer et al., 1996), and is therefore an important consideration.

The reported number of discharge stations in the world has gone down substantially from the peak in the late 1970's (GRDC, 2010). At the same time global precipitation and climate data such as TRMM and ERA-Interim have become available for the last 10–20 yr. Traditional model calibration is impossible if there are no overlapping periods of input and output data. In regions where the flow regime is stationary over time it would be advantageous to use discharge data from a previous period (with sufficiently long records) to overcome this temporal mismatch. Calibration approaches

that do not rely on direct time-series versus time-series comparison are useful in such situations. Prior approaches to model calibration without direct time series comparison include calibration to spectral properties (Montanari and Toth, 2007), recession curves (Winsemius et al., 2009), slope of the flow-duration curve (Yadav et al., 2007; Yilmaz et al., 2008), base-flow index (Bulygina et al., 2009) and the use of a performance measure based on specified exceedance percentages of a synthetic regional flow-duration curve (FDC) for calibration at un-gauged sites (Yu and Yang, 2000). However, in these studies uncertainties in observed discharge are not considered explicitly. Blazkova and Beven (2009) account for discharge uncertainty and use the discharge at nine exceedance percentages between 25 to 90 % exceedance for the FDC as nine out of 57 limits of acceptability in the extended GLUE approach (Beven, 2006, 2009) in flood-frequency estimation. The latter study notes the importance of the realization effect in using a discharge data record of limited length, and the effect this has on the FDC is also discussed by Vogel and Fennessey (1994). The added uncertainty to the FDC stemming from a discharge record of limited length has to be considered if discharge data from another period is used for calibration, especially if the flow regime is not stationary.

Calibrations with traditional performance measures are challenged by problems including the following: (1) uncertainty in discharge data, (2) variable sensitivity of different performance measures to different flow magnitudes, (3) influence of input/output errors of an epistemic nature and (4) inability to evaluate model performance when observation time periods for discharge and model input data do not overlap. Uncertainty in discharge data, which has been shown to be sometimes substantial (Di Baldassarre and Montanari, 2009; Pelletier, 1988; Krueger et al., 2010; Petersen-Overleir et al., 2009) and influence the calibration of hydrological models (McMillan et al., 2010; Aronica et al., 2006), is usually not accounted for in model evaluation with traditional performance measures. Novel approaches in environmental modelling that include evaluation-data uncertainty in model calibration include Bayesian calibration to an estimated probability-density function of discharge (McMillan et al., 2010), Bayesian calibration with a simplified error model (Huard and Mailhot, 2008; Thyer et al., 2009), fuzzy rule based performance measures (Freer et al., 2004) and limits-of-acceptability calibration in GLUE for rainfall-runoff modelling (Liu et al., 2009), flood mapping (Pappenberger et al., 2007), environmental tracer modelling (Page et al., 2007) and flood-frequency estimation (Blazkova and Beven, 2009). Here we explore the limits-of-acceptability GLUE approach applied to flow-duration curves, which could be a way of dealing with some of the effects of non-stationary epistemic errors on the identification of feasible model parameters in real applications (Beven, 2006, 2010; Beven and Westerberg, 2011; Beven et al., 2008). However, in order to establish the extent to which this approach is robust to such errors, a more extensive analysis than that presented here is needed. Flow-duration curves have previously been used in model calibration by Sugawara (1979), Yu and Yang (2000), as one of the criteria considered by Refsgaard and Knudsen (1996) and by Blazkova and Beven (2009), and as a qualitative measure of model performance, e.g. by Houghton-Carr (1999), Kavetski et al. (2011), and Son and Sivapalan (2007).

The aim when calibrating a hydrological model should be to find out whether the model structure can be considered an appropriate conceptualisation or hypothesis of the hydrological processes of interest in that catchment (Beven, 2010). Ideally, the reason for rejecting the model as a suitable hypothesis of these processes should therefore be because the model structure is poor and not because the calibration method does not appropriately account for the uncertainties in the input and output data (i.e. avoiding Type II false negatives). The aim of this paper was to develop a calibration method that addresses the four problems in model calibration with traditional methods outlined above, within the framework of the limits-of-acceptability approach in GLUE and with a specific focus on accurate simulation of the water balance.

2 Study areas and data

The method was first developed for a Honduran catchment characterised by shallow soils and frequent occurrence of surface runoff, the Paso La Ceiba catchment. It was then tested for a contrasting flow regime – the Brue catchment in Great Britain where run-off generation is controlled by subsurface processes on the hill slopes.

2.1 The Paso La Ceiba catchment

The 7500 km^2 Choluteca River basin is located in south-central Honduras (Fig. 1) where the Choluteca River drains to the Pacific at the Gulf of Fonseca. Two water-supply dams (constructed in 1976 and 1992) are located upstream of the capital Tegucigalpa in the upper parts of the basin. The discharge data from the station at Paso La Ceiba, with a catchment area of 1766 km^2, were used here. This catchment has soils that are shallow and eroded (often less than a metre deep) and it is mountainous with elevations ranging from 660 to 2320 m above sea level. The discharge station was destroyed in October 1998 by the flooding that occurred during hurricane Mitch and a new station was installed three kilometres upstream.

The bimodal precipitation regime in the basin is characterised by a high spatial and temporal variability with a dry season November–December to April and a rainy season (with around 80 % of the total precipitation) modulated by a relative minimum, "the midsummer drought", in July–August (Westerberg et al., 2010; Portig, 1976; Magaña et

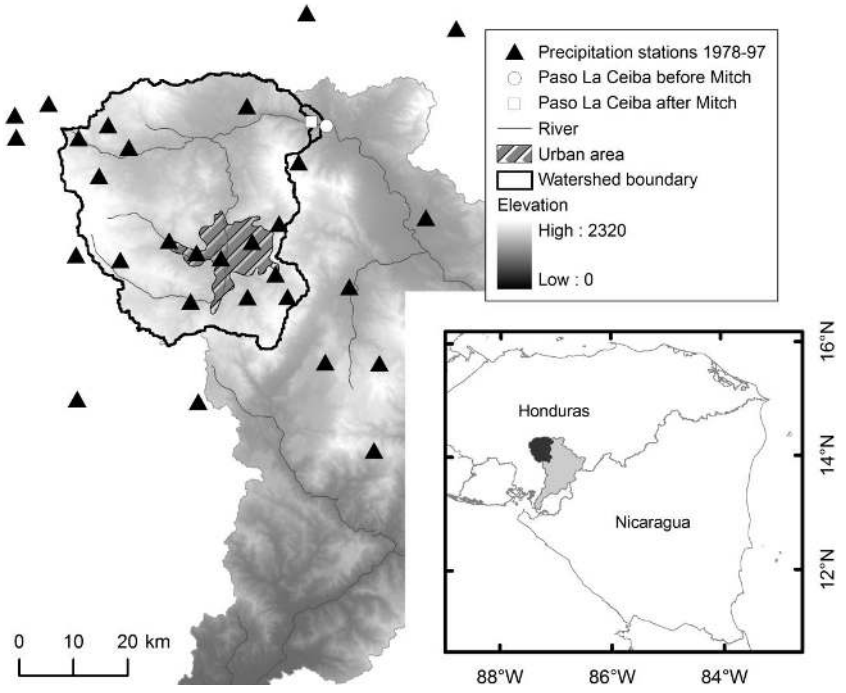

Fig. 1. The Choluteca River Basin and the Paso La Ceiba catchment, the urban area in the upper catchment represents Tegucigalpa, the Honduran capital. Black triangles represent precipitation stations with daily data in 1978–1997 within 30 km of the Paso La Ceiba catchment.

al., 1999). Characteristic of the tropics, temperature variability is low and precipitation is mainly convective. ENSO (El Niño/Southern Oscillation) and Atlantic sea-surface temperatures modulate climate variability on a longer, inter-annual time scale (Diaz et al., 2001; Enfield and Alfaro, 1999). The long dry season in combination with a fast response of run-off to precipitation and little base flow lead to a flow regime where peak flows of short duration account for a large part of the total volume of discharged water.

The WASMOD model was driven with daily data of precipitation and potential evaporation. Precipitation data for 1978–1997 from 29 stations within a 30 km distance of the Paso La Ceiba catchment (Fig. 1) were interpolated with inverse-distance weighting, this method was chosen because of the low correlation between daily precipitation data from different stations and the varying station density (Westerberg et al., 2010). There were almost twice as many active precipitation stations in the end of the 90's as in the early 80's implying that there could potentially be time-varying biases in the interpolated series. Another potential source of data commensurability errors resulted from the fact that precipitation is measured at 7 a.m. but registered on the previous day. Since the delay time from rainfall in the upper catchment to a peak in run-off at the Paso La Ceiba station is less than 24 h and precipitation has a clear diurnal variability with a peak during the second half of the day, the registration of rainfall had to be changed to the day of the actual measurement to agree with the daily time step in the model. The mean annual areal precipitation for the catchment equalled 1060 mm yr^{-1}, with a minimum of 810 mm yr^{-1} and a maximum of 1450 mm yr^{-1} for the studied period.

Potential evaporation was calculated with the Penman-Monteith equation (Monteith, 1965; Allen et al., 1998) using daily data of temperature, wind speed, relative humidity and sun hours from the Toncontín station in Tegucigalpa. There was a decrease in the measured relative humidity around 1984 because of a relocation of the station from a roof-top to the ground and these data were therefore corrected by the difference in mean value between the first and the second period. There was also a clear shift in the relative humidity data when the calculation method was changed from lookup tables to formula in 1 November 1999, which was adjusted for in the same way. Missing meteorological data were filled with daily values for a mean year. The correction of the data was deemed necessary since there was only one station available with data covering the entire modelling period.

The discharge and uncertainty in discharge was previously calculated with a fuzzy linear regression of rating data based on the estimated uncertainty in single discharge and gauge-height measurements by Westerberg et al. (2011) and only the key points are given here. The method accounted for the non-stationarity in the stage-discharge relationship which was substantial in the alluvial Choluteca River, as well as the commensurability error in only having a limited number of gauge-height measurements per day for the calculation of mean daily discharge. The added uncertainty from

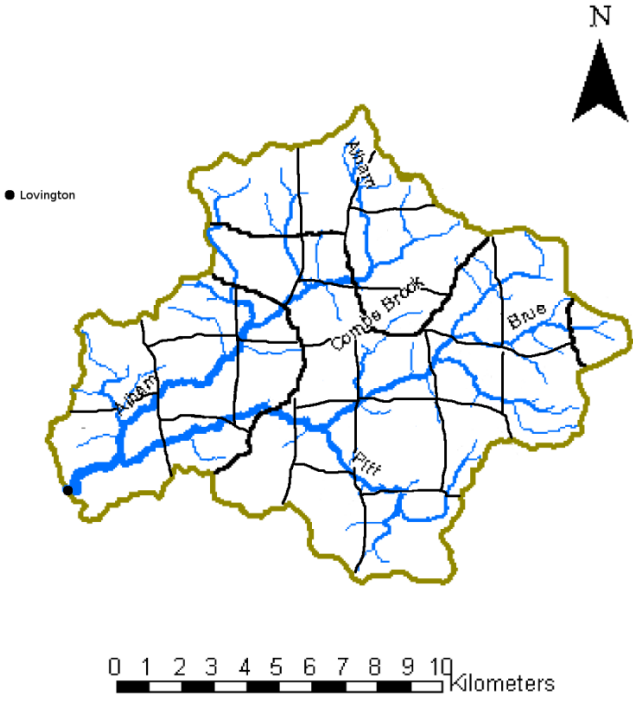

Fig. 2. The Brue catchment and the location of the 28 rain areas (black lines) and the Lovington flow gauge (black dot).

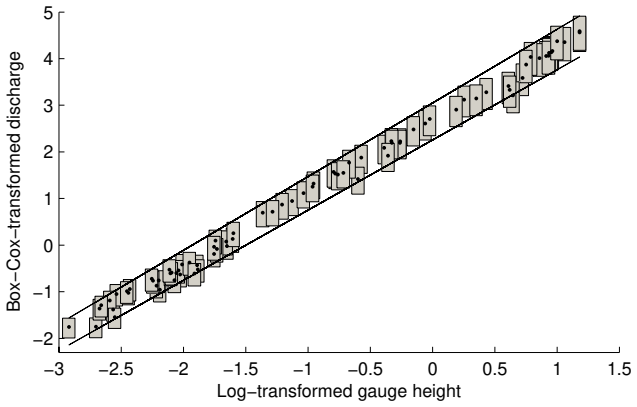

Fig. 3. Uncertain rating curve for the Lovington gauging station in the Brue catchment derived from the stage-discharge measurements from 1990–1998 (stage in m and discharge in $m^3 s^{-1}$ before transformation). The dots represent the measured values and the grey boxes the fuzzy representation of the estimated uncertainty in the measurements. The upper and lower lines represent the uncertainty limits for the fitted rating curve.

this commensurability error was estimated at 17 %, a factor that represented 95 % of the errors from calculations using high temporal resolution stage data for a later period. Larger uncertainties could occur at some events if flow peaks pass between the stage readings, but are not easily estimated. The data included 1216 ratings for 1980–1997 at the Paso La Ceiba station and gauge-height measurements three times-a-day, at 06:00, 12:00 and 18:00. Estimated discharge uncertainty was in the form of a time series of triangular fuzzy numbers consisting of a crisp (best-estimate) discharge and a lower and upper limit.

2.2 The Brue catchment

The 135 km^2 Brue catchment in south-west England (Fig. 2) is characterised by low hills (up to 300 m above sea level) and alternating bands of permeable and impermeable rocks beneath clayey soils on top of which the land use is dominated by grasslands (74 %). An extensive precipitation data set from the HYREX (HYdrological Radar EXperiment) project (Moore, 2000; Wood et al., 2000) includes 49 gauges as well as radar data with a 15-min resolution. The mean areal precipitation for the period 1 January 1995 to 31 December 1997 equalled 770 mm yr^{-1}. Potential evaporation data from the HYREX project that had been calculated using data from an automatic weather station in the lowland part of the catchment were used and periods with missing data were filled using a sine-wave function. Flow data were from the Lovington gauging station, for which the rating curve data from the UK Environmental Agency showed considerable spread. Discharge uncertainty limits were calculated with the same method as for the Paso La Ceiba catchment, but here the rating curve was assumed stationary and 15-min stage data were available for the whole period so no temporal commensurability error needed to be estimated. Discharge and the uncertainty limits were calculated using 79 simultaneous stage-discharge measurements from 1990–1998 that covered the flow range well. The gauge heights (in m) were log-transformed and the discharges (in $m^3 s^{-1}$) were Box-Cox-transformed to obtain a linear relationship (Fig. 3). The Box-Cox lambda parameter was optimized to obtain the highest degree of linearity and a lambda-value of 0.0946 gave a correlation of 0.998. The same uncertainties in the stage and discharge measurements as for the Honduran data were assumed (5 % for gauge height and 25 % for discharge), as the fitted curve encompassed the uncertainty in the ratings well (Fig. 3).

3 Hydrological models

Two hydrological models with different time scales but relatively parsimonious conceptualisations of the dominant hydrological processes in the two catchments were chosen, WASMOD (Xu, 2002) for the Honduran catchment and Dynamic TOPMODEL (Beven and Freer, 2001) for the British catchment.

Table 1. List of equations, parameters and their sampling ranges for the version of WASMOD used in this study.

Model equation	Description	Parameter	Units	Sampling range
$e_t = \min(\mathrm{ep}_t(1 - A_{\mathrm{et}}^{w_t/(\mathrm{ep}_t \times \Delta t)}), w_t/\Delta t)$ where $w_t = p_t \times \Delta t + \mathrm{sm}_{t-1}$ is available water for evaporation, p_t is mean areal precipitation for day t, ep_t is potential evaporation, and sm_{t-1} is soil moisture storage at day $t-1$	Actual evaporation	A_{et}	[-]	[0, 1]
$s_t = S_{\mathrm{f}}(\mathrm{sm}_{t-1})^{0.5}$	Slow flow	S_{f}	[$\mathrm{mm}^{0.5}\,\mathrm{day}^{-1}$]	[e^{-9}, 1]
$f_t = F_{\mathrm{f}} \times \mathrm{sm}_{t-1} \times n_t$ where n_t is active precipitation	Fast flow	F_{f}	[mm^{-1}]	[e^{-7}, e^{-4}]
$n_t = p_t - \mathrm{ep}_t(1 - e^{-\frac{p_t}{\mathrm{ep}_t}})$ if $\mathrm{ep}_t > 1$ $n_t = p_t - \mathrm{ep}_t$ if $\mathrm{ep}_t \leq 1$				
$\mathrm{sc}_t = \mathrm{sc}_{t-1} + f_t \times \Delta t$ $r_t = R_{\mathrm{f}} \times \mathrm{sc}_t$ $\mathrm{sc}_t = \mathrm{sc}_t - r_t \times \Delta t$ where sc_t is the routing storage for day t	Routing of fast flow	R_{f}	[day^{-1}]	[0, 1]
$d_t = \min(s_t + r_t, w_t - e_t)$	Total runoff			
$\mathrm{sm}_t = \mathrm{sm}_{t-1} + (p_t - e_t - d_t) \times \Delta t$	Water balance equation			

3.1 The model used in the Paso La Ceiba catchment – WASMOD

The lumped conceptual water-balance model WASMOD has been applied to many catchments with different climatic conditions and has been used at various spatial scales – e.g. Widen-Nilsson et al. (2007) and Xu and Halldin (1997). Here it was used for the Honduran catchment with a daily time step and a model formulation for snow-free catchments with potential evaporation and precipitation as input data. This version of the model, identical to the snow-free part of the monthly WASMOD model except for the routing scheme, had four parameters for fast flow, slow flow, actual evaporation and routing (Table 1). This was the first application of this model version using a daily time step. The model was evaluated in a split-sample test for 1980–1988/1989–1997, where it was first calibrated in the first period and evaluated in the second and then the reverse. The two years prior to 1980 were used as a warming-up period.

3.2 The model used in the Brue catchment – Dynamic TOPMODEL

In the Brue catchment the semi-distributed Dynamic TOPMODEL was run using a 15-min simulation time step. The simulated runoff series were aggregated to a mean hourly time step before the computation of any goodness-of-fit measure or other analysis of the simulated results. Compared to the original TOPMODEL (Beven and Kirkby, 1979), the dynamic version enables the distributed response to be represented more explicitly through functional units of the landscape. These functional units are not only defined by the topographic index (as in the original TOPMODEL version) but also by similarity in land use, differences in rainfall inputs or other spatial characteristics. In this application, which was the same as in Younger et al. (2009), land use was considered homogenous and the functional units were a function of slope and contributing area (i.e. the topographic index was split up to allow dynamic changes in the upslope contributing area) as well as the spatiotemporal variability in rainfall (see also the previous application of the Probability Distributed Model (PDM) and Grid to Grid models to the Brue in Bell and Moore, 2000). Data from rainfall stations within the same 2 km grid cell were averaged so that 28 "rain areas" were created from the 49 gauges via a nearest-neighbour approach. The parameter intervals for the Monte Carlo sampling are given in Table 2. The model was evaluated in a split-sample test for 1995–1996/1997–30 June 1998, first with the first period for calibration and the second for prediction and then the reverse, 1994 was used as a warming-up period.

4 Flow-duration curve calibration

Monte Carlo runs were performed for both test catchments as a basis for the subsequent calibration. For the Paso La Ceiba catchment 100 000 parameter-value sets were generated and used to simulate runoff series with WASMOD. For the Brue catchment TOPMODEL was run 50 000 times. For calibration (i.e. the selection of the behavioural parameter-value sets and their weights for GLUE) the FDCs of these simulated time series were then evaluated in a comparison

Table 2. Sampling ranges for dynamic TOPMODEL parameters.

Parameter	Units	Sampling range	Description
SZM	[m]	[0.01, 0.1]	Form of the exponential decline in saturated hydraulic conductivity with depth
ln (T_0)	[ln(m^2 h^{-1})]	[-8, 0]	Effective lateral saturated transmissivity
SR_{max}	[m]	[0.005, 0.1]	Maximum soil root zone deficit
SR_{init}	[m]	[0, 0.01]	Initial root zone deficit
CHV	[m h^{-1}]	[500, 2500]	Channel routing velocity
Td	[h]	[0.1, 40]	Unsaturated zone time delay
$\Delta\Theta$	[-]	[0.3, 0.7]	Effective porosity
S_{max}	[m]	[0.1, 0.8]	Maximum effective deficit of the subsurface storage zone

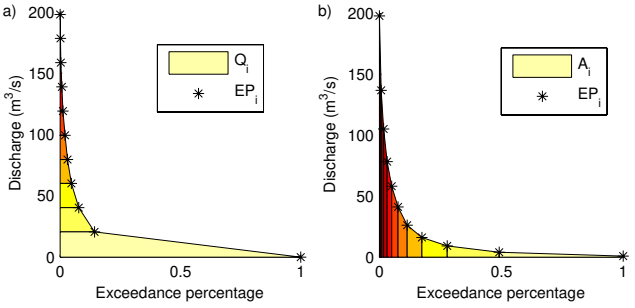

Fig. 4. **(a)** Selection of EP values using equal intervals of crisp discharge (FDC-Q); **(b)** selection of EP values using equal intervals of the area under the FDC (i.e. using equal intervals of water volume contributed by flows in a certain magnitude range (FDC-V).

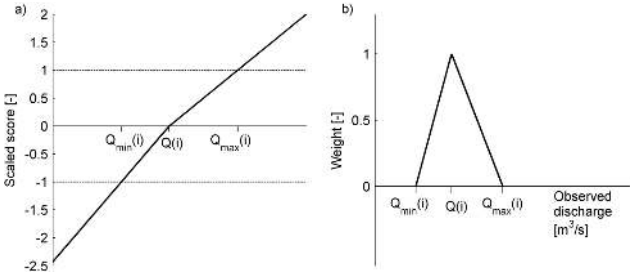

Fig. 5. **(a)** Calculation of the scaled scores, $Q_{\min}(i)$ is the lower limit for the discharge uncertainty at the i:th evaluation point (EP), $Q_{\max}(i)$ the upper limit and $Q(i)$ the crisp discharge. A simulated value that is at the crisp value gets a scaled score of 0, if the value is at the lower limit a scaled score of -1 and at the upper limit it is 1, values within or outside are linearly inter- or extrapolated; **(b)** triangular weighting function applied at each EP such that weights are zero for scaled scores outside the range $[-1, 1]$.

with the observed FDCs. The observed FDCs together with limits of acceptability were constructed from the discharge time series and the estimated uncertainty bounds. The FDC of each simulated discharge series from the Monte Carlo runs was compared to the limits of acceptability for the observed FDC at selected evaluation points (EPs) along the FDC. All simulated FDCs which were inside the limits of acceptability for all EPs were considered behavioural and a performance measure was calculated using a triangular evaluation function at each EP. This performance measure was used as an informal likelihood measure for each behavioural parameter-value set. This FDC calibration was compared to that using the model efficiency (Nash and Sutcliffe, 1970) with different behavioural threshold values. Furthermore, the model performance when using an observed FDC from a time period different to the simulated one was evaluated in the Paso La Ceiba catchment to assess the ability of the method to address mismatching observation time periods. These are called "time-shift" calibrations below. Finally, in a posterior analysis the simulated discharge uncertainty ranges, which resulted from using the different performance measures, were compared to the observed discharge uncertainties for the simulated periods.

4.1 Selection of evaluation points

The selection of the exceedance percentages that were used as evaluation points (EPs) – i.e. the points where the simulated FDC was compared to the observed – was an important choice for the FDC calibration. The high-flow part of the FDC, which describes the dynamic response of the catchment to the effective precipitation input, usually contains most of the information about catchment response and many parameters are therefore sensitive with respect to these high flows. Sufficient points on this part of the FDC therefore needs to be set in order to constrain these parameters. Here we explored two methods for EP selection which each emphasized different aspects of the FDC (Fig. 4). For the first method the crisp discharge values (i.e. the best estimate of the uncertain discharges) were classed into N equal classes (Fig. 4a). The minimum and maximum discharge values of the entire FDC were excluded and the remaining $N-1$ discharge class boundary values were used to calculate the corresponding EPs. Here $N = 20$ intervals were used resulting in 19 EPs. Different ways can be used to calculate specific exceedance percentages or discharge values for the FDC, but the choice of method is negligible in cases where the FDC is based on thousands of daily discharges as was the case here (Vogel and Fennessey, 1994). We calculated exceedance percentages from the sorted discharges based on the percentile values 100(0.5/n), 100(1.5/n), ..., 100([n-0.5]/n), where n is the number of discharge values. Linear interpolation was used between the sorted observed discharge values. This calculation was first reversed to calculate EPs in terms of exceedance percentages for the discharge class boundary values for the crisp observed discharge. It was finally used to

calculate discharge for the upper and lower acceptability limits and for the simulated discharge at these EPs, which were then used in the calculation of the performance measures. The second method for EP selection consisted of re-scaling the FDC so that it represented the total volume of water contributed by flows smaller than or equal to a given magnitude. These volumes were then divided into N equal classes and the EPs were calculated in the same way, again excluding the minimum and maximum discharge values. As the area under the normal FDC represents the volume of water discharged during the time for which the FDC was calculated, this approach equalled a weighting using N intervals of equal area below the curve for the crisp discharge (Fig. 4b). Since we used $N = 20$ this resulted in volume increments of 5 %. The expectation was that the volume-based EP selection would provide a more appropriate evaluation with respect to the entire FDC than the discharge-based selection, because the latter meant that the low flows were not constrained for the types of flow regimes considered here. The volume method was therefore expected to be well-suited for water-balance studies, whereas the discharge method was more focused on high-flow performance.

4.2 Performance measures

Two performance measures $R_{\mathrm{FDC-Q}}$ (for EP selection based on discharge intervals) and $R_{\mathrm{FDC-V}}$ (for EP selection based on volume intervals) were calculated using the sum of a triangular weighting function based on the observed discharge and its limits of acceptability at each EP (Fig. 5b). Scaled scores were calculated to evaluate the deviations of the simulated discharge with respect to the limits of acceptability. If the simulated discharge value equalled the crisp discharge for a certain EP, the scaled score was zero; if it was at the upper or lower limit the score was 1 and -1 respectively. Values between and outside these values were calculated based on linear inter- or extrapolation (Fig. 5a).

In this study behavioural simulations were required to be inside the limits of acceptability (i.e. to have an absolute scaled score ≤ 1) at all EPs. The performance measures $R_{\mathrm{FDC-V}}$ and $R_{\mathrm{FDC-Q}}$ were calculated as:

$$R_{\mathrm{FDC}} = 1 - \frac{\sum_{i=1}^{N-1} |S_i|}{N-1} \text{ where } -1 \leq S_i \leq 1, i = 1, 2, ..., N-1 \quad (1)$$

where $N-1$ was the number of EPs and S_i the scaled score at EP i. This means that a simulation with a perfect fit to the crisp discharge at all EPs received a value of 1 and if the simulated discharge was at either limit for all EPs, this resulted in a value of 0. There were no values lower than 0 as simulations were classed as non-behavioural if the absolute scaled score was larger than 1 for any EP (Fig. 5b). These performance measures were compared to the model efficiency (R_{eff}) calculated based on the crisp discharge (with different behavioural thresholds). This form of triangular weighting function based on scaled scores has been used before, for example by Blazkova and Beven (2009) and Liu et al. (2009) and is analogous to the fuzzy measures used by Pappenberger et al. (2007) and Page et al. (2007).

4.3 Posterior analysis of simulated and observed discharges

In a posterior analysis the time series of observed uncertain discharge were compared to the simulated results from the calibration and prediction with the two models. A simple measure of how well the simulated and observed uncertain discharge agree, is given by the calculation of the percentage of time that the observed and simulated uncertainty bounds overlap (here termed OP). A similar measure, called *reliability*, has been used previously for single-valued observed discharge (Yadav et al., 2007). The overlap measure can be high simply because the simulated uncertainty is overestimated. Therefore a combined overlap percentage (COP) was calculated as the mean of the percentage of the overlapping range between the observed and simulated discharge relative to the observed and relative to the simulated discharge range (Eq. 2).

$$\mathrm{COP} = \frac{\sum_{t=1}^{T} \left(\mathrm{mean} \left(\frac{QR_{\mathrm{overlap}}}{QR_{\mathrm{obs}}}, \frac{QR_{\mathrm{overlap}}}{QR_{\mathrm{sim}}} \right) \right)}{T} \quad (2)$$

T is the number of time steps, QR_{overlap} the intersection between the simulated and observed discharge ranges, $\mathrm{QR_{obs}}$ the observed discharge range and QR_{sim} the simulated discharge range. A perfect match of 100 % can then not be achieved if the simulated uncertainty is overestimated.

More complex measures, such as a PQQ-plot (Thyer et al, 2009) or a rank histogram, analyse the quantiles of the observed value in the simulated distribution. The generalised rank histogram (McMillan et al., 2010) is an extension of the rank histogram that compares two uncertain distributions so that uncertainty in the observed data can be accounted for. However, the generalised rank histogram does not relate how far simulated values that are outside the observed distribution lie. We therefore chose to analyse scaled scores to the limits of acceptability for the time series of simulated values. These were calculated in the same way as the scaled scores used in the calculation of $R_{\mathrm{FDC-V}}$ and $R_{\mathrm{FDC-Q}}$, but for each time step instead of each EP in the FDC. The scaled scores of all the behavioural simulations were analysed for different flow types: base flow, rising limbs, falling limbs, peaks and troughs, to be able to identify differences in the simulation of different parts of the hydrograph between the criteria. For each performance measure the histograms of scaled scores were normalised to the number of behavioural simulations to facilitate comparison. The classification of discharge into different flow types was made in the same way as by Younger et al. (2011) for the Brue catchment. However, we used different threshold values since the hydrographs were analysed at an hourly instead of 15-minute time step. The observed flow Q_t at time t was classified as:

baseflow if $Q_t < Q_b$

rising limb if $Q_{t-T} < Q_t < Q_{t+T}$ and $Q_t > Q_b$

falling limb if $Q_{t-T} > Q_t > Q_{t+T}$ and $Q_t > Q_b$

peak if $Q_{t-T} < Q_t$ and $Q_t > Q_{t+T}$ and $Q_t > Q_b$

trough if $Q_{t-T} > Q_t$ and $Q_t < Q_{t+T}$ and $Q_t > Q_b$

The values of Q_b and T were determined through visual inspection of the classified hydrographs. The values were determined to $Q_b = 1.7\,\mathrm{m^3\,s^{-1}}$ ($= 13\,\mathrm{l\,s^{-1}\,km^{-2}}$) and $5\,\mathrm{m^3\,s^{-1}}$ ($= 2.8\,\mathrm{l\,s^{-1}\,km^{-2}}$) and $T = 4$ h and 3 days for the Brue and Paso La Ceiba catchment respectively. Plots of the time series of mean scaled scores for each performance measure together with the simulated and observed discharge were also used to analyse the simulated results, especially the periods where the simulations were outside the uncertainty in the observed discharge.

5 Results

5.1 Observed uncertain FDCs

The FDCs for the two catchments illustrate the differences in flow regime. In the Honduran catchment base flow was very low and a larger part of the total volume of water was contributed by high flows than in the British catchment (Fig. 6). At Paso La Ceiba the flow regime (as illustrated by the FDCs) was more or less stable in-between the calibration and evaluation periods. In the Brue catchment, where the discharge record was much shorter, the low-flow part of the FDC was not as stable as the high-flow part between the two periods. If a model is calibrated with data from another time period (a "time-shift" calibration) and the FDC is not stable, there could be a realisation effect in using a limited sample of discharge data. Therefore the extremes from a bootstrap of FDCs for successive nine- and two-year periods of discharge data (for the Paso La Ceiba and Brue catchment respectively) were plotted to illustrate the extra uncertainty from this realisation effect – that should be accounted for if the stationarity of the FDC is unknown. As would be expected, the realisation effect was larger for the Brue compared to Paso La Ceiba. Factors affecting the magnitude of the realisation effect include the length of the record, the nature of the climate variability and the non-stationarity of the hydrological regime. The estimated uncertainty in discharge ranged between −43 to +73 % of the best discharge estimate at Paso La Ceiba (Westerberg et al., 2011) and ±34 % in the Brue catchment. The EPs of the FDCs ranged from a fraction of flow equalled or exceeded of 0.004 to 0.70 for $R_{\mathrm{FDC-V}}$ and from 0.0002 to 0.30 for $R_{\mathrm{FDC-Q}}$ for the two periods in the Brue and from 0.003 to 0.69 for $R_{\mathrm{FDC-V}}$ and from 0.0003 to 0.17 for $R_{\mathrm{FDC-Q}}$ for the two periods at Paso La Ceiba. The very low values included here reflect the fact that the high flows represent a small fraction of all flows.

5.2 Number of behavioural parameter-value sets

The identification of behavioural parameter-value sets using the performance measures based on the FDC evaluation points resulted in more behavioural parameter-value sets for the discharge-interval selection compared to the volume-interval selection for both catchments (Table 3). The numbers of behavioural parameter-value sets are those that survived the limits of acceptability for all the EPs considered, of the 100 000 simulations for Paso La Ceiba and 50 000 simulations for the Brue. The time-shift calibration results for Paso La Ceiba use the FDC from one period, to provide limits of acceptability for the other period (which in this case is assumed to have no observed discharges available). The column labelled prediction shows the percentage of parameter-value sets calibrated in the second period which were behavioural for the first period based on the two FDC criteria. For the Brue catchment the performance for the two periods was quite different and only 3 % ($R_{\mathrm{FDC-V}}$) and 13 % ($R_{\mathrm{FDC-Q}}$) of the parameter-value sets in the second period were also behavioural in the first. The percentages were higher for the Paso La Ceiba with almost 50 % of the parameter-value sets behavioural in both periods for both criteria. This is likely a result of the higher uncertainty in discharge combined with the less complex rainfall-runoff relationship in this catchment compared to the Brue, especially since a simpler model and more uncertain precipitation data were used compared to the semi-distributed model set-up and dense rain-gauge network in the Brue. It might also provide an indication that the more complex Dynamic TOPMODEL has been over-fitted to responses and errors in the calibration period that are then rather different in the evaluation period.

Table 4 shows the results based on the Nash-Sutcliffe efficiency performance measure, using different thresholds to define the behavioural parameter-value sets, and also with an additional constraint based on the absolute volume error (VE) in predicted discharge. With higher thresholds there was a greater chance that the sets of behavioural parameter values for the two periods would be non-overlapping, while the maximum values for the Brue were generally lower than at Paso La Ceiba. In the Paso La Ceiba catchment the addition of the VE had a large constraining effect on the number of behavioural parameter-value sets but not in the Brue catchment. The time-shift calibration was not possible with this performance measure.

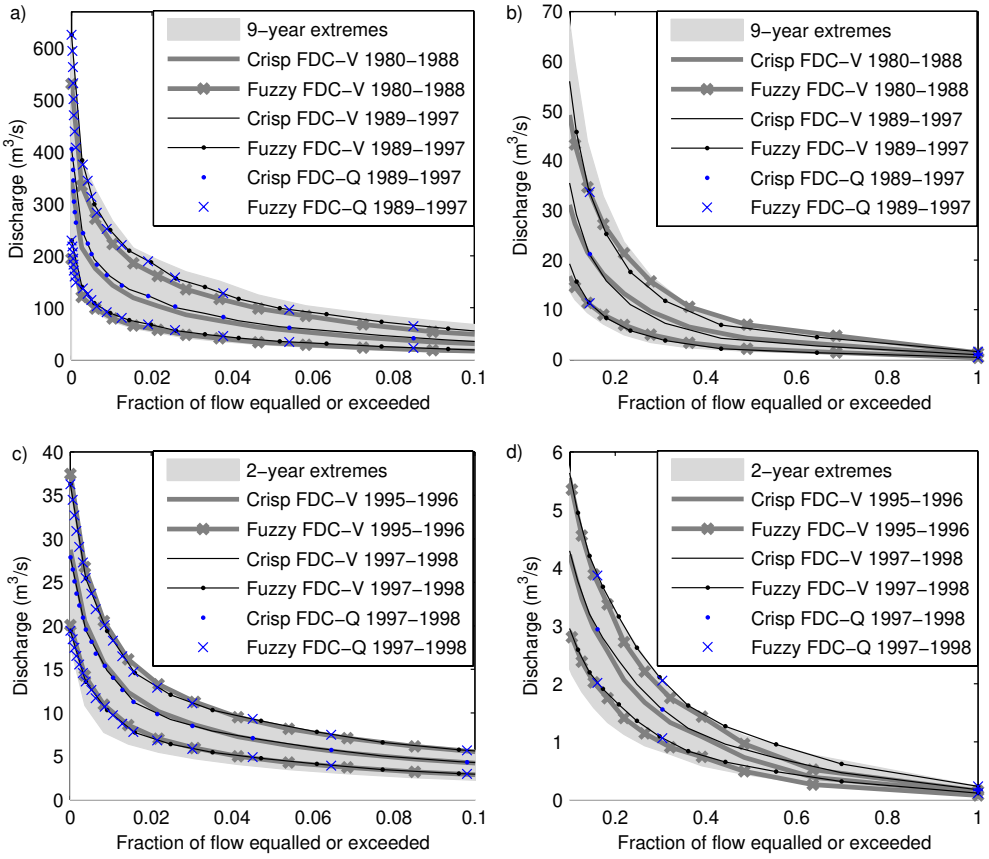

Fig. 6. Observed crisp and uncertain FDCs for the Paso La Ceiba catchment, **(a–b)** upper and lower flow range respectively and for the Brue catchment, **(c–d)** upper and lower flow range respectively. The extreme FDC represents the maximum and minimum uncertain FDC for all consecutive 9- and 2-yr periods for the Paso La Ceiba and Brue catchment respectively. The FDC-V represents volume interval EPs and FDC-Q discharge interval EPs (only plotted for the last period in each catchment). The high and low flows of the FDCs are plotted separately for better visualisation; note the difference in scale on the y-axis.

Table 3. Number of behavioural parameter-value sets for the different FDC performance measures.

Catchment (model)	Paso La Ceiba (WASMOD)					Brue (Dynamic TOPMODEL)		
Performance measures	Calibration		Time-shift Calibration[1]		Prediction[2]	Calibration		Prediction[2]
	1980–1988	1989–1997	1980–1988	1989–1997		1995–1996	1997–1998	
R_{FDC-Q}	17 085	24 166	21 932	22 853	48 % (11 575)	983	477	13% (123)
R_{FDC-V}	758	1430	871	1408	47 % (673)	360	42	3 % (12)

[1] Calibration using the FDC from the previous/later period [2] Percentage (number) of behavioural parameter-value sets calibrated in the second period that were also behavioural in the first period.

5.3 Parameter identifiability

5.3.1 The Paso La Ceiba catchment – WASMOD

In this catchment the performance measures based on the FDC resulted in more overlapping sets of behavioural parameter values between calibration and prediction compared to the calibration with R_{eff} (Tables 3 and 4). The FDC criterion based on volume EPs, R_{FDC-V}, resulted in much fewer behavioural parameter-value sets than R_{FDC-Q}. The largest difference in parameter identifiability was seen for the evaporation and slow-flow parameters which mainly control simulated discharge for low flows and recession periods (Fig. 7). They were better constrained for the R_{FDC-V}

Table 4. Number of behavioural parameter-value sets for different Nash-Sutcliffe based performance measures.

Catchment (model)	Paso La Ceiba (WASMOD)			Brue (Dynamic TOPMODEL)		
Performance measures[1]	Calibration		Prediction[2]	Calibration		Prediction[2]
	1980–1988	1989–1997		1995–1996	1997–1998	
$R_{\text{eff}} > 0.7$ & VE < 20%	796	12 477	4 % (464)	2299	240	4 % (82)
$R_{\text{eff}} > 0.7$ & VE < 10%	365	6399	2 % (147)	1128	127	0 % (0)
$R_{\text{eff}} > 0.7$	1473	28 455	5 % (1,473)	2696	240	4 % (108)
$R_{\text{eff}} > 0.75$	89	20 046	0.4 % (89)	985	13	0.4 % (4)
$R_{\text{eff}} > 0.8$	0	11 101	0 % (0)	140	0	0 % (0)
$R_{\text{eff}} > 0.85$	0	2246	0 % (0)	3	0	0 % (0)

[1] VE is the absolute volume error [2] Percentage (number) of behavioural parameter-value sets calibrated in the second period that were also behavioural in the first period.

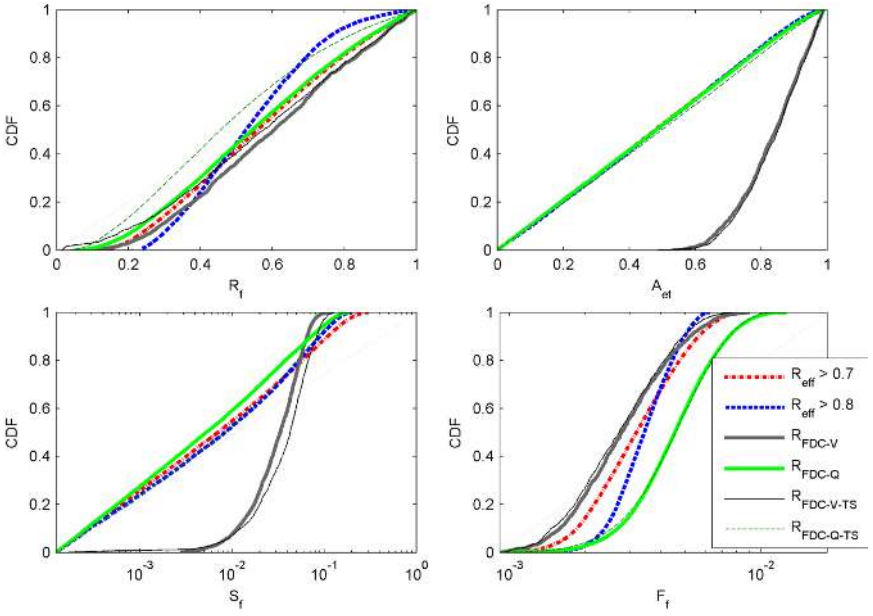

Fig. 7. Cumulative informal likelihood distributions for all WASMOD model parameters (R_{f} – routing of fast flow, A_{et} – evaporation, S_{f} – slow flow, and F_{f} – fast flow). The informal likelihood weights for each performance measure were calculated for the calibration in 1989–1997 for R_{eff}, $R_{\text{FDC-Q}}$ and $R_{\text{FDC-V}}$, and for the calibration in 1989–1997 using the FDC for 1980–1988 for $R_{\text{FDC-Q-TS}}$, and $R_{\text{FDC-V-TS}}$ in the Paso La Ceiba catchment.

measure compared to the $R_{\text{FDC-Q}}$ and R_{eff} measures, which mostly constrained model performance at medium to high-flows. The behavioural parameter-value sets obtained from calibrating the model for 1989–1997 using the "time-shift" FDC for 1980–1988 did not differ much from calibration with the FDC from 1989–1997, especially for the volume EP criterion, as the flow regime did not change substantially in-between the two periods (Fig. 6–7).

5.3.2 The Brue catchment – Dynamic TOPMODEL

As in the Paso La Ceiba catchment, the largest difference in parameter identifiability between the R_{eff} and $R_{\text{FDC-V}}$ measures could be seen for the parameters controlling the recession/slow flow and the evaporation in the model (Fig. 8). In Dynamic TOPMODEL the SZM parameter describes the exponential decline in saturated hydraulic conductivity with depth and controls the shape of the hydrograph in the recession periods. It was constrained to much lower values for $R_{\text{FDC-V}}$ compared to the other measures. The SR_{max} parameter, which controls the water available for evaporation, was also more constrained for $R_{\text{FDC-V}}$. The best simulations for R_{eff} ($R_{\text{eff}} > 0.8$) showed more constraint on the CHV and S_{max} parameters. In the case of CHV, the channel-routing velocity parameter, this reflects the sensitivity of the R_{eff} measure to timing errors in the higher peak hydrographs. The sensitivity of S_{max}, which controls the root zone deficit due to actual evapotranspiration, might reflect the effect of

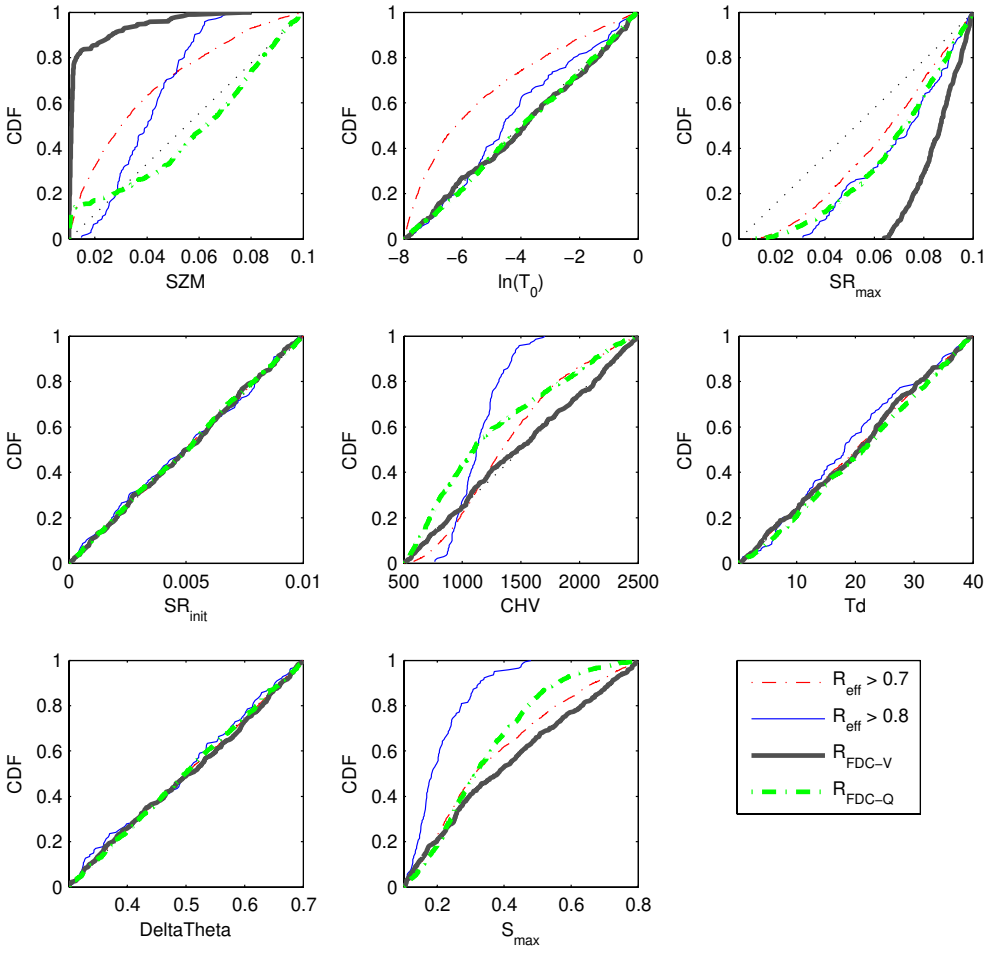

Fig. 8. Cumulative informal likelihood distributions for all Dynamic TOPMODEL parameters (the parameter names are explained in Table 2). The informal likelihood weights for each performance measure were calculated for the calibration in 1995–1996 in the Brue catchment.

antecedent conditions on peak flow magnitude and timing that is not so important for the R_{FDC} measures.

5.4 Simulated flow-duration curves

5.4.1 The Paso La Ceiba catchment – WASMOD

The $R_{\mathrm{FDC-V}}$ measure gave simulated FDCs that most closely resembled the observed FDC for the whole flow range in both calibration and prediction. The largest difference between the performance measures occurred at low flows for both the calibration and evaluation periods (Fig. 9). Here almost all of the simulations for the R_{eff} and $R_{\mathrm{FDC-Q}}$ measures underestimated the discharge, but there were a number of simulations that had a large overestimation in this flow range. The $R_{\mathrm{FDC-V}}$ simulations were more evenly distributed within the range of the uncertain observed FDC at the low-flow EPs. This difference at low flows was not surprising since the largest difference in the parameter identifiability (Fig. 7) was seen for the evaporation and slow-flow parameters that control this part of the FDC. For the $R_{\mathrm{FDC-Q}}$ measure this lack of constraint was not surprising as there were no low-flow EPs. For the R_{eff} calibration the low-flow simulation even for behavioural parameter-value sets with the highest R_{eff} values resulted in consistent errors for low flows. The calibration in 1989–1997 using the "time-shift" FDC in 1980–1988 with the $R_{\mathrm{FDC-V}}$ measure gave results similar to when the 1989–1997 FDC was used for the same measure. The $R_{\mathrm{FDC-Q}}$ measure gave good high-flow performance but the poorest performance for low flows as seen when plotted for the volume EPs.

In prediction 1989–1997 R_{eff} gave more consistent underestimation for high flows compared to $R_{\mathrm{FDC-V}}$ and $R_{\mathrm{FDC-Q}}$. As in the calibration period, the low-flow performance was much poorer for R_{eff} and $R_{\mathrm{FDC-Q}}$ compared to $R_{\mathrm{FDC-V}}$, which was largely consistent with the observed FDC. Note that in calibration the lowest EP for which the $R_{\mathrm{FDC-Q}}$ was evaluated in the current study was at a crisp discharge of 21 $\mathrm{m^3\,s^{-1}}$. Figure 9 shows that this still allows sufficient freedom for the behavioural simulations to depart from the observed FDC limits at lower flows, in this case for 86 %

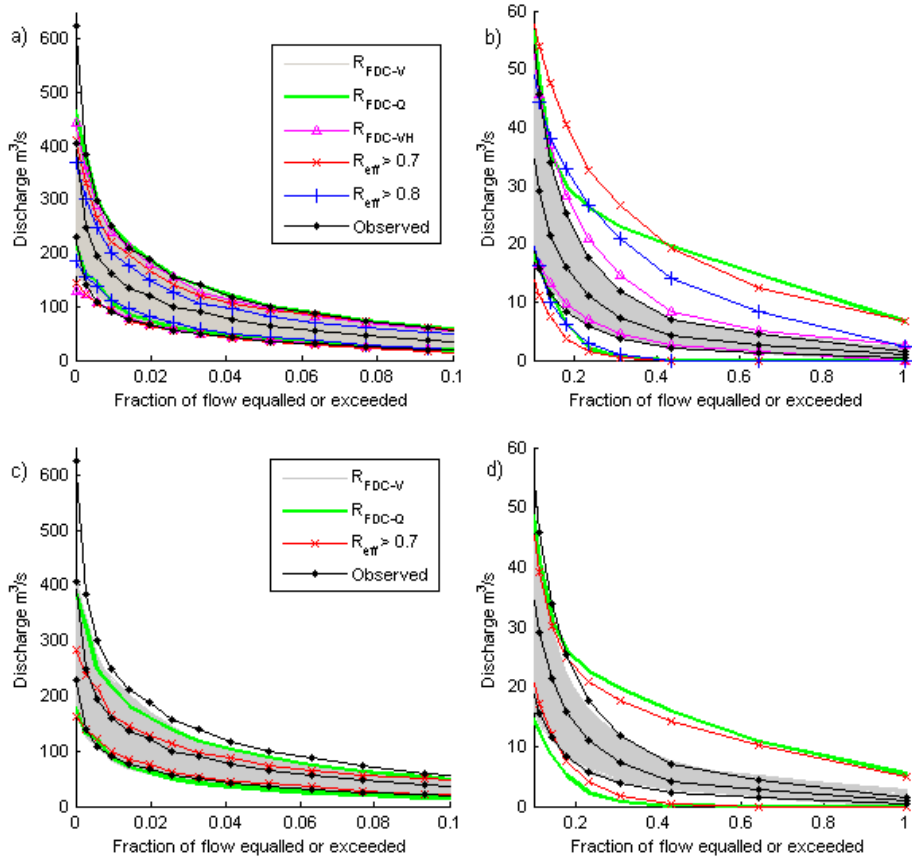

Fig. 9. **(a)** and **(b)** FDCs for behavioural parameter-value sets for WASMOD in the Paso La Ceiba catchment for calibration in 1989–1997 using $R_{\mathrm{FDC-V}}$ (all FDCs plotted as grey lines), R_{eff}, $R_{\mathrm{FDC-Q}}$, and $R_{\mathrm{FDC-V-TS}}$ (maximum and minimum FDC values plotted as lines) and observed crisp, upper-limit and lower-limit discharge; **(c)** and **(d)** FDCs for prediction in 1989–1997 using behavioural parameter-value sets for $R_{\mathrm{FDC-V}}$ (all FDCs plotted as grey lines), R_{eff} and $R_{\mathrm{FDC-Q}}$ calibrated 1980–1988 (maximum and minimum FDC values plotted as lines) and observed crisp, upper limit and lower limit discharge. The FDCs are split in two plots (left – high flows and right – low flows) at 10 % exceedance. All FDCs are plotted for the volume interval EPs.

of the time, and that these simulated results were similar to those of the R_{eff} calibration.

5.4.2 The Brue catchment – Dynamic TOPMODEL

In the Brue catchment the results were largely similar to the Paso La Ceiba catchment (Fig. 10). The $R_{\mathrm{FDC-V}}$ criterion also constrained the low-flow part of the FDC which the other criteria did not. Here, however, the behavioural simulations did not cover the entire low-flow range which could indicate that some of the observed behaviour could not be reproduced by the model. The majority of the flows at the low-flow EPs were overestimated for R_{eff} and $R_{\mathrm{FDC-Q}}$ in this catchment. Again, the number of increments used in the determination of $R_{\mathrm{FDC-Q}}$ allows significant freedom amongst behavioural parameter-value sets in the prediction of lower flows and a similar pattern is seen for R_{eff}.

5.5 Posterior analysis of simulated and observed discharges

The measures of overlap (OP and COP) between the simulated and observed uncertain discharge bounds were generally higher for the $R_{\mathrm{FDC-V}}$ measure compared to the other measures (Fig. 11). As the COP measure accounted for overestimated predictive uncertainty a high value of this measure was more important than for OP. The results for the time-shift calibration using the FDC from another time period gave results similar to that of the normal FDC calibration. The best R_{eff} simulations ($R_{\mathrm{eff}} > 0.8$) resulted in a similar number of behavioural simulations as $R_{\mathrm{FDC-V}}$ at Brue, but gave much lower overlap than for $R_{\mathrm{FDC-V}}$, which was largely because of the poorer low-flow performance. The $R_{\mathrm{FDC-Q}}$ measure resulted in better results in the Brue catchment compared to Paso La Ceiba. This might relate to the fact that there was more baseflow at Brue wherefore the EPs for the discharge-interval-selection method covered the low-flow part of the FDC better than at Paso La Ceiba.

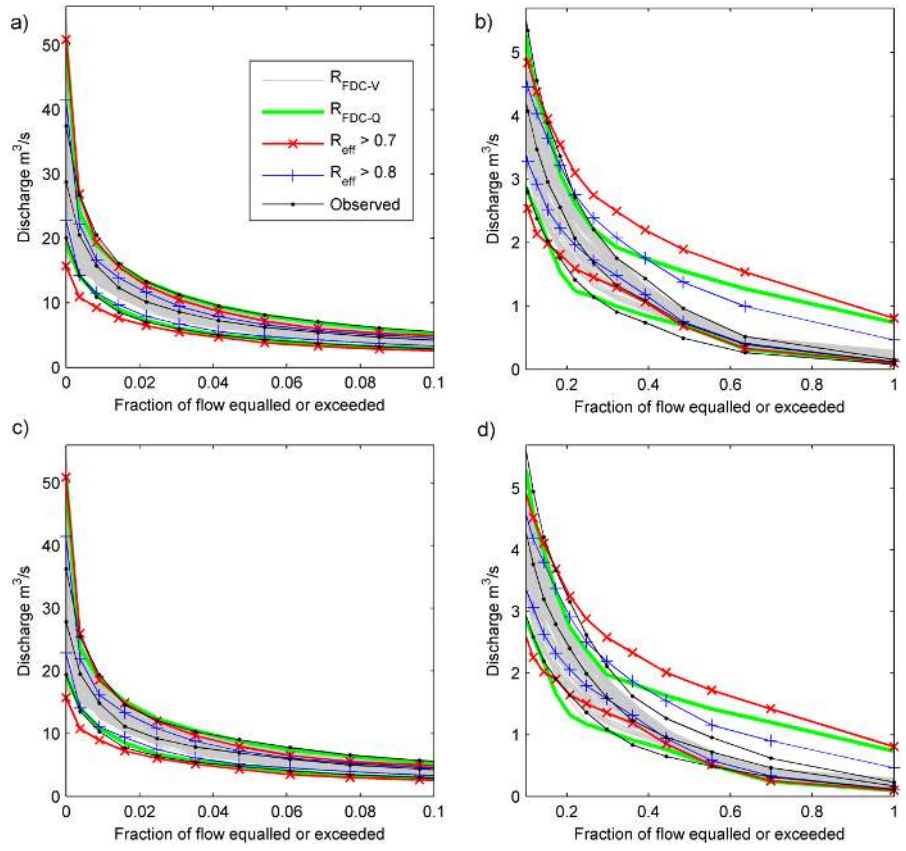

Fig. 10. **(a)** and **(b)** FDCs for behavioural parameter-value sets for Dynamic TOPMODEL in the Brue catchment for calibration in 1995–1996 using $R_{\mathrm{FDC-V}}$ (all FDCs plotted as grey/shaded lines), R_{eff}, and $R_{\mathrm{FDC-Q}}$ (maximum and minimum FDC values plotted as lines) and observed crisp, upper and lower discharge; **(c)** and **(d)** FDCs for prediction in 1997–1998 using the behavioural parameter-value sets from 1995–1996. The FDCs are split in two plots (left – high flows and right – low flows) at 10 % exceedance. All FDCs are plotted for the volume interval EPs.

5.5.1 The Paso La Ceiba catchment – WASMOD

The simulated discharge for the Paso La Ceiba catchment was in general in good agreement with the observed discharge (Fig. 12). During the low-flow periods of some years the discharge was underestimated for all performance measures, indicating a possible model-structural error in simulating a slower/deeper ground-water response or errors in the input data.

The posterior analysis of the mean scaled scores for different parts of the hydrograph (Fig. 13) for the prediction in 1989–1997 showed that when using the $R_{\mathrm{FDC-V}}$ calibration compared to R_{eff}: (1) the distributions of scaled scores were more centred on zero, (2) there were fewer base flows that were underestimated, and (3) the largest difference was seen for the troughs, falling limbs and base flows that are controlled by the slow-flow and evaporation parameters. The same results were seen in all the other calibration/prediction periods. Events where the predicted discharge was underestimated did not generate as large scaled scores as if the predicted discharge was overestimated, as the uncertainty bounds were wider in absolute terms for high flows compared to low flows, this explains the skew in the histograms in Fig. 13. The distributions of the scaled scores for R_{eff} and $R_{\mathrm{FDC-Q}}$ were always centred on negative scaled scores for all flow types.

A plot of the mean scaled scores and the discharge for 1989–1990 revealed the difference in low-flow performance (Fig. 14). A large scaled deviation can be seen for all performance measures in the end of 1990 where there is a peak in the predicted discharge but not in the observed. This is a type of epistemic error that could be a result of erroneous discharge data, influence of upstream dams or unrepresentative precipitation data. This type of event had a large effect on the R_{eff} calibration where it generated a large sum-of-squares error and a reduction in overall performance. A similar deviation is seen in the end of 1989. The maximum scaled scores for all the calibration and prediction periods at Paso La Ceiba were consistently larger for the FDC-based measures compared to R_{eff} which might indicate that the FDC criteria are not as sensitive to such disinformative events.

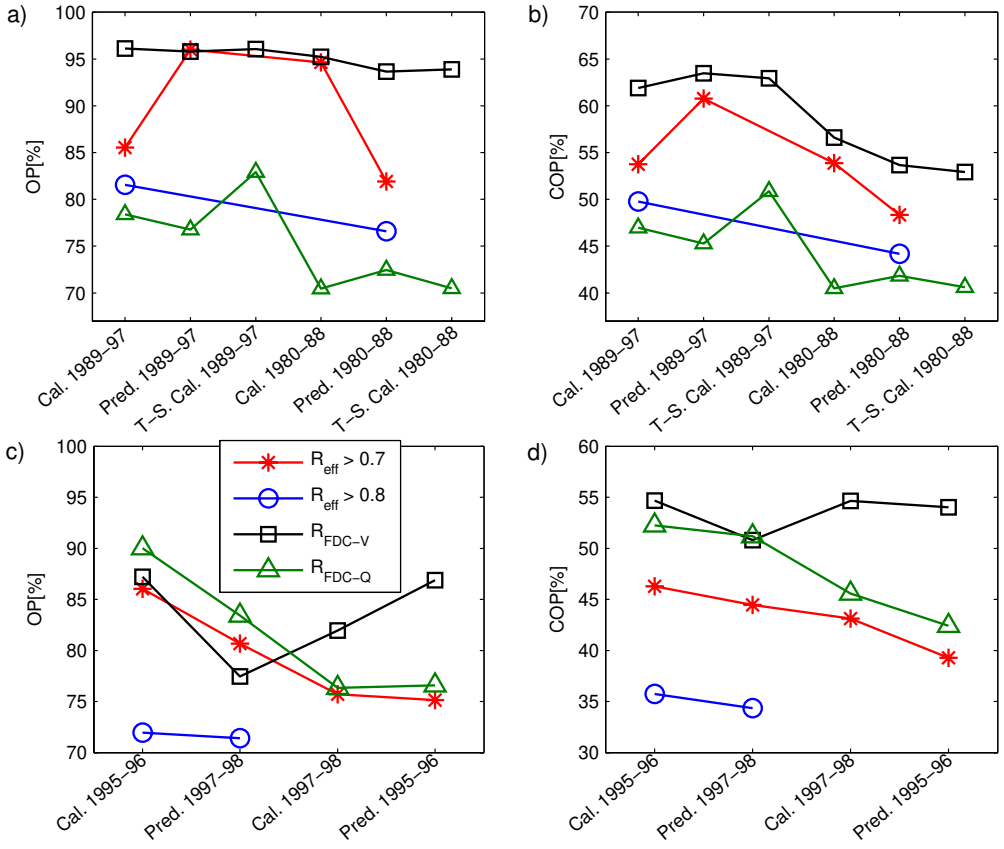

Fig. 11. Percentage of time that the simulated and observed uncertain discharges overlap (OP) and the combined overlap percentage (COP) for the calibration (Cal.), time-shift calibration (T-S. Cal.) and prediction (Pred.) using WASMOD in the Paso La Ceiba catchment **(a–b)** and calibration and prediction using Dynamic TOPMODEL in the Brue catchment **(c–d)**.

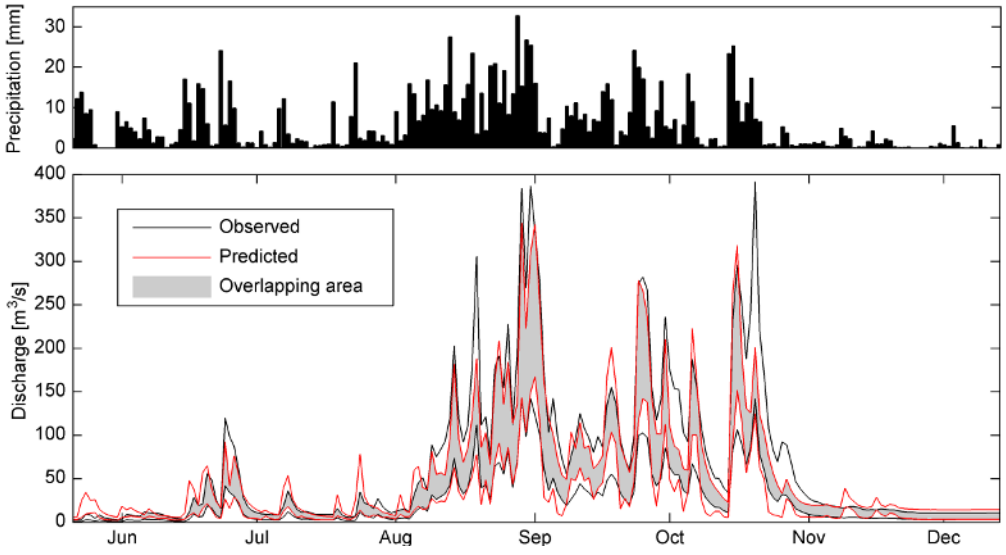

Fig. 12. Uncertainty limits for observed discharge and predicted discharge (5 % and 95 % percentiles of the predicted discharge of all behavioural parameter-value sets) in the rainy season 1995 with WASMOD parameters calibrated 1980–1988 using the $R_{\mathrm{FDC-V}}$ performance measure in the Paso La Ceiba catchment. The overlapping area between the two uncertain intervals is plotted in grey.

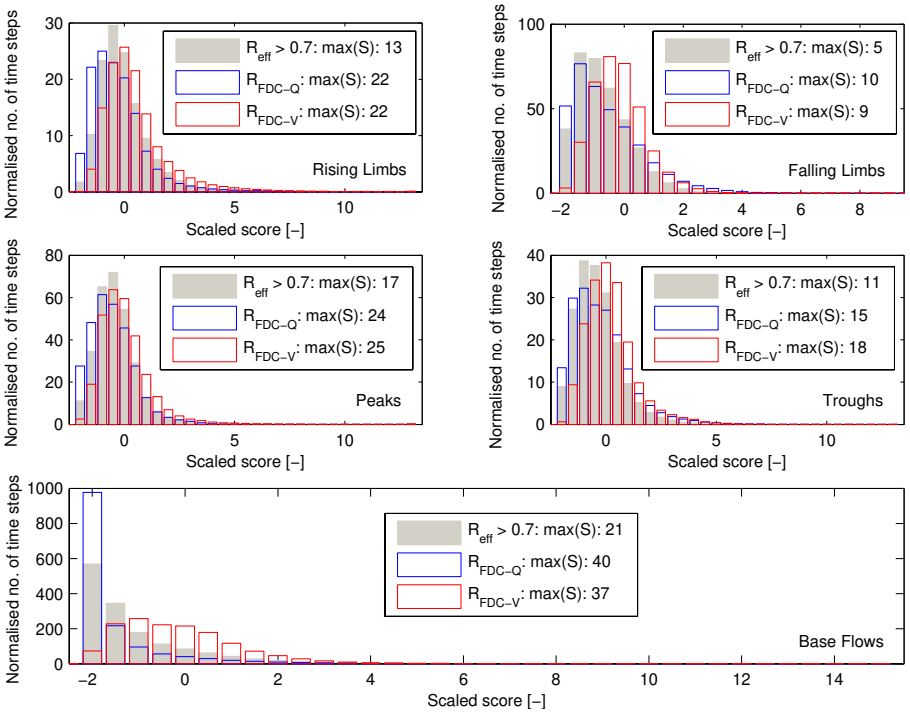

Fig. 13. Scaled scores to limits of acceptability for different parts of the hydrograph at Paso La Ceiba for prediction in 1989–1997 with behavioural parameter-value sets for 1980–1988 for WASMOD. For each performance measure the histograms were normalised by the number of behavioural simulations, which means that the y-axis represents the number of time steps. The upper range of the histogram x-axis was limited to improve the visibility of the lower range, the maximum scaled scores, max(S), for each criterion are given in the legends and all scaled scores larger or equal to the last bin are plotted in the last bin.

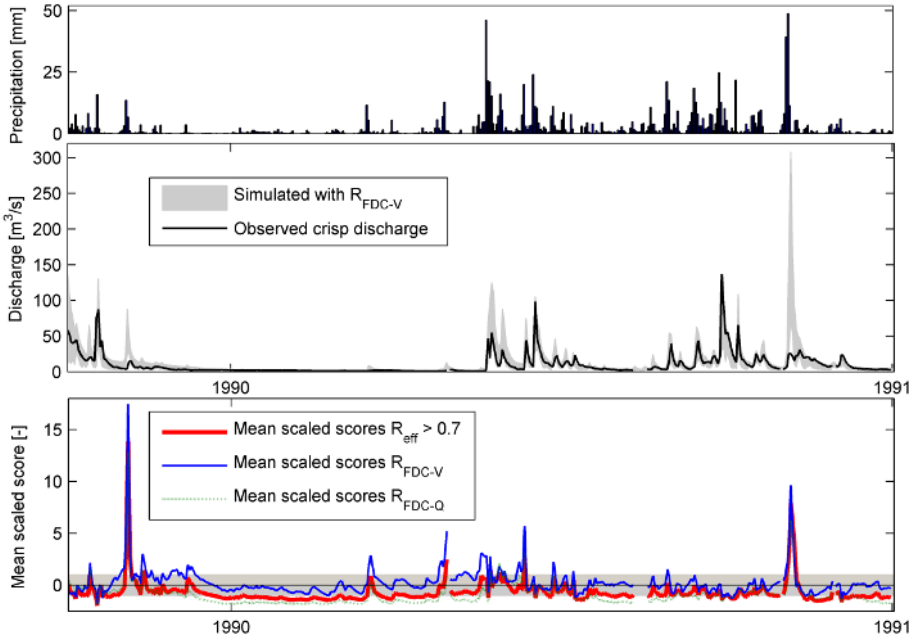

Fig. 14. Daily precipitation in 1989–1990 (top) and predicted and observed crisp daily discharge for behavioural parameter-value sets from using $R_{\mathrm{FDC-V}}$ for calibration of WASMOD in the Paso La Ceiba catchment in 1980–1988 (middle). The mean scaled scores for all performance measures are plotted in the bottom plot where the grey area represents a scaled score from −1 to 1, i.e. a simulated discharge with a score inside this range is inside the discharge uncertainty limits.

5.5.2 The Brue catchment – Dynamic TOPMODEL

The results for the Brue catchment were similar to Paso La Ceiba with generally better performance for base flows, falling limbs and troughs for $R_{\mathrm{FDC-V}}$. In contrast to Paso La Ceiba the results were poorer for peaks and rising limbs compared to R_{eff} (Fig. 15), this difference was less pronounced in 1995–1996 where the calibration worked better. Also in contrast to the Paso La Ceiba catchment, the R_{eff} and $R_{\mathrm{FDC-Q}}$ measures resulted in more overestimation of low flows here, which is also seen in Fig. 10. The maximum scaled scores were in general larger for the FDC-based criteria but not for all flow types as was the case at Paso La Ceiba. Some periods of plausible model-structural errors were visible for the base flows where there were many time steps with overprediction with a scaled score around 5. These periods did indeed seem to be a result of model-structural error in July–October 1997 as shown by a plot (Fig. 16) of the mean scaled scores for the calibration during the same years; all of the performance measures gave simulations that overpredicted in this period. Another period of probable model-structural error could be seen where the simulated discharge was underestimated in the wetting-up period for the prediction in 1997–1998 (Fig. 17).

6 Discussion and conclusions

This paper has explored a calibration method that addresses four particular problems that arise in calibration with traditional performance measures: (1) uncertain discharge data, (2) variable sensitivity of different performance measures to different flow magnitudes, (3) influence of input/output errors of an epistemic nature and (4) inability to evaluate model performance when observation time periods for discharge and model input data do not overlap. The method was evaluated in two catchments with contrasting flow regimes where two different models were applied at two different time scales. The results showed that when the exceedance percentages (EPs) of the FDC were chosen based on volume intervals, this calibration method resulted in more constrained low-flow parameters and a better overlap with the observed data compared to a "traditional" calibration using the Nash-Sufcliffe model efficiency.

FDCs have been used previously in model calibration and evaluation (Blazkova and Beven, 2009; Son and Sivapalan, 2007; Sugawara, 1979; Yu and Yang, 2000). The novel aspect of our use of the FDC is that it takes account of uncertainty in the discharge data and at the same time shows that the FDC can work surprisingly well as a single criterion in some cases. Here discharge uncertainty was calculated using a fuzzy linear regression for the rating curve based on estimations of the uncertainty in both stage and discharge measurements. Other methods could also be considered to do this (e.g. Pappenberger et al., 2006), but the non-stationarity of the stage-discharge relationship at Paso La Ceiba (Westerberg et al., 2011) constrained the number of feasible methods for that site. Our construction of the uncertain FDC implies an interpretation of the discharge uncertainty as an epistemic error with an expectation of non-stationary bias rather than a random error, which would lead to averaging of individual errors. There might be many reasons for such epistemic errors including current meters that have not been re-calibrated and base levels subject to erosion and deposition (Westerberg et al., 2011). Correlation in fitting successive EPs is handled naturally in the limits-of-acceptability approach, since only models that satisfy all limits are retained in prediction, and simulations with consistent bias relative to the best-estimate discharge are given a low weight.

The choice of the evaluation points at which the limits of acceptability for the FDC are set is an important consideration in the FDC calibration and the selection could be made in different ways. The important point is that the choice should be informed by the perceptual understanding of the uncertainties in the hydro-meteorological data and made with the aims of the modelling study and the characteristics of the FDC in mind. For example, if high or low-flow performance is of special importance then additional points could be chosen for these flow ranges. The shape of the FDC will influence how the EPs are spaced for a given selection method (e.g. the Brue catchment had higher base flow and therefore for $R_{\mathrm{FDC-Q}}$ the lowest EP occurred at a higher exceedance percentage than at Paso La Ceiba). In both catchments in this study the volume weighting gave the best overall results as it constrained the model also for the low flows and recession periods. At the daily time scale it also resulted in better simulations for peak flows, while at the sub-daily time scale there was greater uncertainty in peak-flow timing compared to R_{eff}. The volume-based EP-selection method would be especially suitable for water-balance studies where the correct volume of water for different flow ranges is of specific concern, but exact timing is not as critical. The low sensitivity to timing errors will have a limited effect as long as runoff coefficients are represented correctly. At sub-daily time steps and where peak-flow timing is of greater concern, additional criteria could be enforced to constrain this aspect of the simulations. In doing so, the epistemic uncertainties associated with estimates of the higher discharges, particularly resulting from rating-curve extrapolation, should be taken into account. The FDC-calibration approach allows different weightings by including different EPs and one could also consider giving different weights to different EPs in the calculation of the likelihood measure. In other catchments than those studied here, other factors may come into play, such as the effects of the timing of snowmelt in snow-dominated catchments. Using FDC calibration, the exact timing of the melt would not be as important as for a Nash-Sutcliffe measure (see the example in Ambroise et al., 1996), but the distribution of the melt over time would still be important and would likely require additional constraints. The posterior

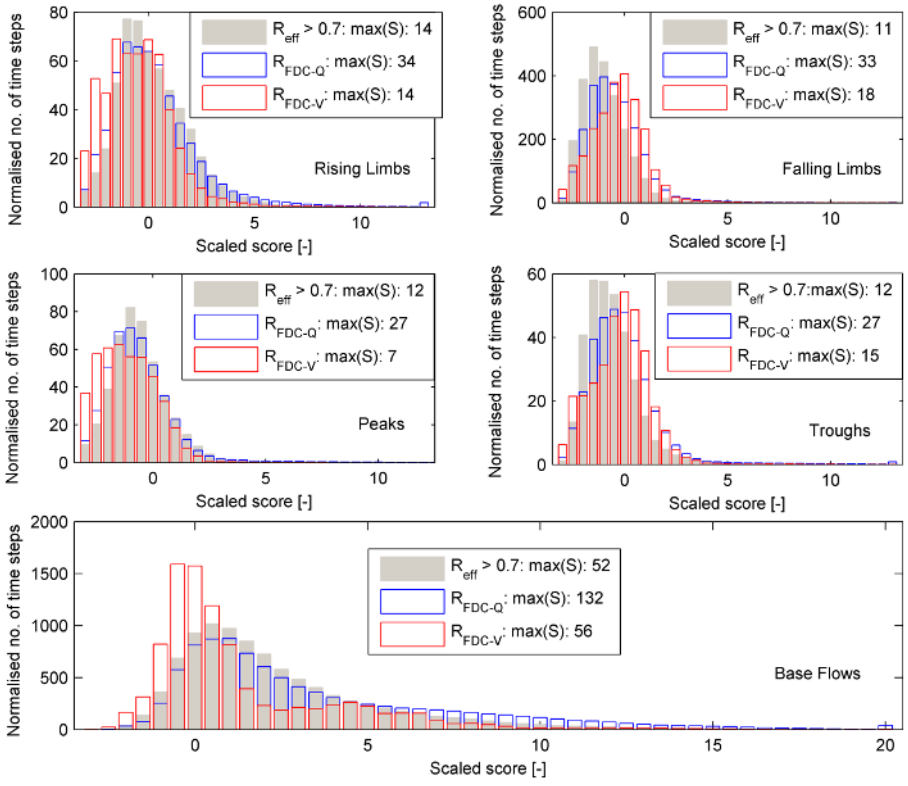

Fig. 15. Scaled scores to limits of acceptability for different parts of the hydrograph at Brue for calibration in 1997–1998 using Dynamic TOPMODEL. For each performance measure the histograms were normalised by the number of behavioural simulations, so the y-axis represents the normalised number of time steps. The upper range of the histogram x-axis was limited to improve the visibility of the lower range, the maximum scaled scores, max(S), for each criterion are given in the legends and all scaled scores larger or equal to the last bin are plotted in the last bin.

analysis of the simulated time series employed here can be useful in deciding whether additional criteria are necessary.

In calibration to "hydrological signatures" such as an FDC calculated from the discharge series, the simulated uncertainty bounds have a direct interpretation relative to the uncertainty in the observed discharge data. This is an advantage compared to say a behavioural threshold-value of R_{eff} of 0.7 that is not easily interpretable (Legates and McCabe, 1999; Seibert, 2001). Winsemius et al. (2009) set limits of acceptability in GLUE (for different types of signatures such as recession curves) based on inter-annual variability but took no explicit account of the uncertainty in the observed discharge data.

It is interesting to note that the 19 EPs used for the $R_{\text{FDC-V}}$ criterion provided better information for the calibration of the model than the 3288 days or 17544 hours for the first years of calibration/prediction used for R_{eff}. Limited information content in discharge time series was also demonstrated by Juston et al. (2009) and Seibert and Beven (2009), who found that calibration using a small fraction of data points chosen at hydrologically informed times was comparable to when the whole time series was used. We chose R_{eff} for comparison with the FDC-calibration as it is sensitive to timing errors, well-known and commonly used. Other approaches such as multi-criteria calibration or the calculation of R_{eff} on transformed discharge can of course also be used to constrain simulations. We also tested log and square-root transformed discharge in the calculation of R_{eff}. This resulted in good simulations for low flows whereas the simulation for the highest flows was poorer constrained compared to R_{eff} and the FDC-calibration. A multi-criteria calibration could constrain different aspects simultaneously, but the problems of deciding on a behavioural threshold value and accounting for discharge-data uncertainty remain in such approaches.

When the FDC-method was first developed it was tested with inconsistent satellite-derived precipitation in a Honduran basin which resulted in that no simulations were found that were consistent with the observed FDC. In such cases a traditional calibration will result in low values for the performance measure and not point as strongly to where the inconsistencies in the simulated flow regime occur. This is therefore an advantage of using constraints based on signatures (such as a FDC) calculated from the flow data, as suggested elsewhere for use in regionalisation methods for estimating the response of ungauged basins (e.g. Yadav et al., 2007).

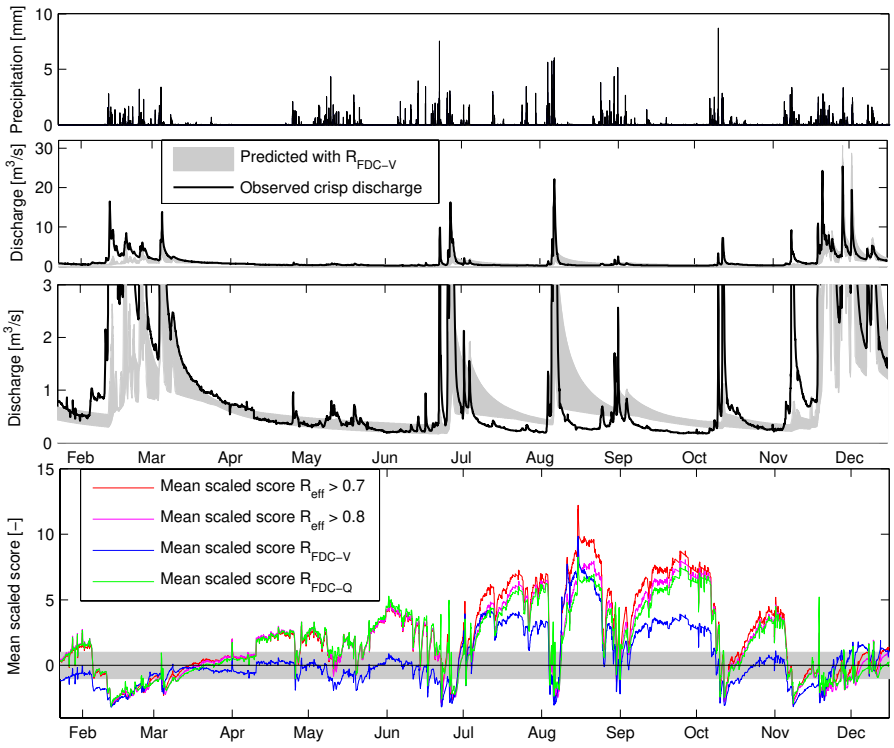

Fig. 16. Predicted and observed crisp discharge for 1997–1998 for behavioural parameter-value sets for $R_{\mathrm{FDC-V}}$ from calibration using Dynamic TOPMODEL in 1995–1996 for the Brue catchment (upper plot shows the whole flow range, middle the low flows). The mean scaled scores for all performance measures are plotted in the bottom plot where the grey area represents a scaled score from −1 to 1, i.e. a simulated discharge with a score inside this range is inside the discharge uncertainty limits. The $R_{\mathrm{FDC-V}}$ criterion gave simulations with less overprediction in the summer. In July–October 1997 there was a period of consistent overprediction at low flows for all performance measures where the model could not reproduce the observations.

Disinformative data can lead to biased parameter estimates in calibration if the model is forced to compensate for such errors. We expect the FDC-calibration method to be more robust to disinformation in many cases, especially pure timing errors such as an isolated single precipitation event registered on the wrong day or single events with inconsistent inputs and outputs which might lead to rejection of all models in a limits-of-acceptability evaluation based on individual time steps (e.g. Liu et al., 2009). The extent to which it is robust needs to be assessed in future studies. It would likely be most sensitive to disinformation that affects the tails of the simulated and observed distributions, as that would lead to a greater effect on the shape of the simulated or observed FDC. In the absence of methods to identify and remove disinformative data prior to calibration, a posterior analysis like the one we employed here can be used to readily identify periods where the simulations from the behavioural parameter-value sets are failing. These periods can then be analysed to see whether the lack of fit can be attributed to disinformative data or to model-structural errors (which in that case could lead to learning from where the model is failing). In some cases it might be obvious where there are problems in the observations, for example where a discharge hydrograph is observed without significant rainfall. In the Paso La Ceiba catchment a large peak flow was simulated in 1990 without a peak in observed discharge (Fig. 14), which is not likely for that type of hydrological regime where there is a direct relationship between rainfall and runoff, and this event was therefore likely an epistemic error in the discharge data such as the effect of an upstream dam or wrongly digitised data. In the case of the Brue catchment, with 49 rain gauges in 135 km^2, significant departures between observed and predicted discharge (such as the large scaled scores for the low-flows in July–October 1997 in Fig. 16) might be inferred to be more a result of model deficiencies than input errors. These periods of probable model failure at low flows could be readily seen in the analysis of the scaled scores for the different parts of the hydrograph.

Are these two models then acceptable hypothesis about the hydrological processes in the respective catchments or should they be rejected? As noted in the introduction this depends on the hydrological processes of interest and the aims of the modelling application. In the Paso La Ceiba catchment the simulated discharge overlapped with the observed discharge for around 95 % of the time steps for the $R_{\mathrm{FDC-V}}$ calibration and prediction in both periods. If the overall

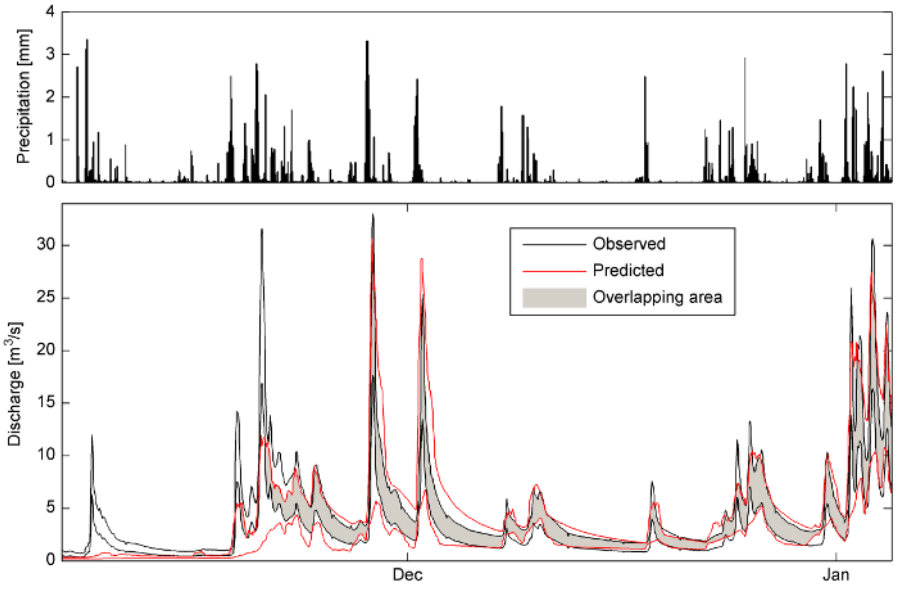

Fig. 17. Uncertainty limits for observed discharge in 1997–1998 and predicted discharge (5 % and 95 % percentiles of the predicted discharge of all behavioural parameter-value sets calibrated in 1995–1996 using the $R_{\mathrm{FDC-V}}$ performance measure) for the same period for Dynamic TOPMODEL in the Brue catchment. The overlapping area between the two uncertain intervals is plotted in grey. In the beginning of November there was a period where the model could not reproduce the observations.

water-balance is of interest then this would be an acceptable result, especially considering the likely time-variable uncertainty in the rainfall inputs because of the low and time-varying number of precipitation stations for this complex precipitation regime (Westerberg et al., 2010). Additional evaluation criteria might of course still reveal that we are not getting the right answers for the right reasons (Kirchner, 2006), a possibility that should be kept in mind if making predictions of changed future conditions. In the Brue catchment the overlap between simulated and observed discharge was much lower, between 75–90 % of the time for the $R_{\mathrm{FDC-V}}$ calibration and prediction in both periods. In combination with the analysis of the scaled scores this suggests that, given the number of rain gauges in this catchment, the model structure can be rejected as a good hypothesis for the hydrological processes in this catchment. The information about likely model-structural errors revealed in this posterior analysis could be investigated to see if some improvements might be implemented, such as in the representation of the storage-discharge function at low flows (which in Dynamic TOPMODEL is not restricted to any particular functional form).

Experiments using the FDC calibration with time-shifted data in the Honduran catchment resulted in similar parameter-value distributions and overlap with the observed discharge as the normal FDC calibration. It might therefore have potential for bridging temporal mismatch of data availability in regions such as Central America where there are few available discharge data in the last decades but more data for the 70–90's. The effect of climate variability and the stationarity of the flow regime in the longer term must be accounted for in such applications. If the flow regime is non-stationary or if the time-shifted period does not cover periods of climate variability (e.g. El Niño/La Niña years) to a sufficient extent, the extra uncertainty stemming from this realisation effect should be added to the FDC. The method might also be useful for studying the effect of modifications to the hydrological regime such as dams, where "pre-dam" data could be used for calibration to the natural flow regime. Another area of possible application is calibration to regional FDCs such as in the study by Yu and Yang (2000), but also taking uncertainties in the calibration of the hydrological model and the data into account. A major advantage of the FDC-calibration approach is the way in which it requires structured consideration of the uncertainties expected to affect the observed and simulated FDCs, not the least in the discharge estimates themselves but also other sources of uncertainties that affect model calibration.

Acknowledgements. This work was funded by the Swedish International Development Cooperation Agency grant number 75007349 and SWE-2005-296. The authors thank the staff at SANAA, SERNA, UNAH and SMN in Honduras for their kind assistance in providing data for the study. Freer's time on this paper was in part made possible with funding from the UK Natural Environment Research Council (NERC), Flood Risk from Extreme Events (FREE) programme (grant number NE/E002242/1).

Edited by: J. Vrugt

References

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration – guidelines for computing crop water requirements, FAO, 300 p., 1998.

Ambroise, B., Freer, J., and Beven, K.: Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France, Water Resour. Res., 32, 2147–2159, 1996.

Aronica, G. T., Candela, A., Viola, F., and Cannarozzo, M.: Influence of rating curve uncertainty on daily rainfall-runoff model predictions., in: Predictions in Ungauged Basins: Promise and Progress, edited by: Sivapalan, M., Wagener, T., Uhlenbrook, S., Liang, X., Lakshmi, V., Kumar, P., Zehe, E., and Tachikawa, Y., IAHS Publ 303, 116–124, 2006.

Bell, V. A. and Moore, R. J.: The sensitivity of catchment runoff models to rainfall data at different spatial scales, Hydrol. Earth Syst. Sci., 4, 653–667, doi:10.5194/hess-4-653-2000, 2000.

Beven, K. J.: Changing Ideas in Hydrology - the Case of Physically-Based Models, J. Hydrol., 105, 157–172, 1989.

Beven, K. J.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, 2006.

Beven, K. J.: Environmental Modelling: An Uncertain Future?, Routledge, London, 2009.

Beven, K. J.: Preferential flows and travel time distributions: defining adequate hypothesis tests for hydrological process models Preface, Hydrol. Process., 24, 1537–1547, doi:10.1002/Hyp.7718, 2010.

Beven, K. J. and Freer, J.: A dynamic TOPMODEL, Hydrol. Process., 15, 1993–2011, 2001.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrolog. Sci. B., 24, 43–69, 1979.

Beven, K. J. and Westerberg, I. K.: On red herrings and real herrings: disinformation and information in hydrological inference, Hydrol. Process., 25, 1676-1680, 2011

Beven, K. J., Smith, P. J., and Freer, J. E.: So just why would a modeller choose to be incoherent?, J. Hydrol., 354, 15–32, doi:10.1016/j.jhydrol.2008.02.007, 2008.

Beven, K., Smith, P. J., and Wood, A.: On the colour and spin of epistemic error (and what we might do about it), Hydrol. Earth Syst. Sci. Discuss., 8, 5355–5386, doi:10.5194/hessd-8-5355-2011, 2011

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.

Boyle, D. P., Gupta, H. V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., 36, 3663–3674, 2000.

Bulygina, N., McIntyre, N., and Wheater, H.: Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis, Hydrol. Earth Syst. Sci., 13, 893–904, doi:10.5194/hess-13-893-2009, 2009.

Criss, R. E. and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals, Hydrol. Process., 22, 2723–2725, doi:10.1002/Hyp.7072, 2008.

Di Baldassarre, G., and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol Earth Syst Sc, 13, 913-921, 2009.

Diaz, H. F., Hoerling, M. P., and Eischeid, J. K.: ENSO variability, teleconnections and climate change, Int. J. Climatol., 21, 1845–1862, 2001.

Enfield, D. B. and Alfaro, E. J.: The dependence of Caribbean rainfall on the interaction of the tropical Atlantic and Pacific oceans, J. Climate, 12, 2093–2103, 1999.

Freer, J., Beven, K., and Ambroise, B.: Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, Water Resour. Res., 32, 2161–2173, 1996.

Freer, J., Beven, K. J., and Peters, N.: Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure, in: Calibration of Watershed Models, edited by: Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A. N., and Turcotte, R., AGU Books, Washington, 69–87, 2003.

Freer, J. E., McMillan, H., McDonnell, J. J., and Beven, K. J.: Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures, J. Hydrol., 291, 254–277, doi:10.1016/j.jhydrol.2003.12.037, 2004.

Garrick, M., Cunnane, C., and Nash, J. E.: A Criterion of Efficiency for Rainfall-Runoff Models, J. Hydrol., 36, 375–381, 1978.

Grayson, R. B., Moore, I. D., and Mcmahon, T. A.: Physically Based Hydrologic Modeling .2. Is the Concept Realistic, Water Resour. Res., 28, 2659–2666, 1992.

Global Runoff Data Centre: http://grdc.bafg.de, last access: 23 February 2010, 2010.

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., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Houghton-Carr, H. A.: Assessment criteria for simple conceptual daily rainfall-runoff models, Hydrol. Sci. J., 44, 237–261, 1999.

Huard, D., and Mailhot, A.: Calibration of hydrological model GR2M using Bayesian uncertainty analysis, Water Resour. Res., 44, W02424, doi:10.1029/2007wr005949, 2008.

Juston, J., Seibert, J., and Johansson, P. O.: Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment, Hydrol. Process., 23, 3093–3109, doi:10.1002/Hyp.7421, 2009.

Kavetski, D., Fenicia, F., and Clark, M.: Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, Water Resour. Res., 47, W05501, doi:10.1029/2010WR009525, 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, W03s04, doi:10.1029/2005wr004362, 2006.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, 2005,
http://www.adv-geosci.net/5/89/2005/.

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta,

G. S., Brazier, R. E., Butler, P., and Haygarth, P. M.: Ensemble evaluation of hydrological model hypotheses, Water Resour. Res., 46, W07516, doi:10.01029/02009WR00784, 2010.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, 1999.

Liu, Y. Q. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resour. Res., 43, W07401, doi:10.1029/2006WR005756, 2007.

Liu, Y., Freer, J., Beven, K. J., and Matgen, P.: Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error, J. Hydrol., 367, 93–103, doi:10.1016/j.jhydrol.2009.01.016, 2009.

Magaña, V., Amador, J. A., and Medina, S.: The midsummer drought over Mexico and Central America, J. Climate, 12, 1577–1588, 1999.

McDonnell, J. J.: Where does water go when it rains? Moving beyond the variable source area concept of rainfall-runoff response, Hydrol. Process., 17, 1869–1875, 2003.

McMillan, H. and Clark, M.: Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme, Water Resour. Res., 45, W04418, doi:10.1029/2008wr007288, 2009.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrol. Process., 24, 1270–1284, doi:10.1002/Hyp.7587, 2010.

Montanari, A. and Toth, E.: Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins?, Water Resour. Res., 43, W05434, doi:10.1029/2006wr005184, 2007.

Monteith, J. L.: Evaporation and the Environment, in: The State and Movement of Water in Living Organisms., edited by: Fogg, G. E., Cambridge University Press, 205–234, 1965.

Moore, R. J., Jones, D. A., Cox, D. R., and Isham, V. S.: Design of the HYREX raingauge network, Hydrol. Earth Syst. Sci., 4, 521–530, doi:10.5194/hess-4-521-2000, 2000.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models 1. A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Page, T., Beven, K. J., Freer, J., and Neal, C.: Modelling the chloride signal at Plynlimon, Wales, using a modified dynamic TOPMODEL incorporating conservative chemical mixing (with uncertainty), Hydrol. Process., 21, 292–307, doi:10.1022/Hyp.6186, 2007.

Pappenberger, F., Matgen, P., Beven, K. J., Henry, J. B., Pfister, L., and Fraipont de, P.: Influence of uncertain boundary conditions and model structure on flood inundation predictions, Adv. Water Resour., 29, 1430–1449, 2006.

Pappenberger, F., Frodsham, K., Beven, K., Romanowicz, R., and Matgen, P.: Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations, Hydrol. Earth Syst. Sci., 11, 739–752, doi:10.5194/hess-11-739-2007, 2007.

Pelletier, P.: Uncertainties in the single determination of river discharge: a literature review, Can. J. Civil Eng., 15, 834–850, 1988.

Petersen-Overleir, A., Soot, A., and Reitan, T.: Bayesian Rating Curve Inference as a Streamflow Data Quality Assessment Tool, Water Resour. Manag., 23, 1835–1842, doi:10.1007/s11269-008-9354-5, 2009.

Portig, W. H.: The climate of Central America, in: World Survey of Climatology, edited by: Schwerdtfeger, W., Elsevier, New York, 405–464, 1976.

Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32, 2189–2202, 1996.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, doi:10.1029/2009wr008328, 2010.

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

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, W10531, doi:10.1029/2009WR008933, 2010.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrol. Process., 15, 1063–1064, doi:10.1002/hyp.446, 2001.

Seibert, J. and Beven, K. J.: Gauging the ungauged basin: how many discharge measurements are needed?, Hydrol. Earth Syst. Sci., 13, 883–892, doi:10.5194/hess-13-883-2009, 2009.

Smith, P., Beven, K. J., and Tawn, J. A.: Informal likelihood measures in model assessment: Theoretic development and investigation, Adv Water Resour, 31, 1087–1100, doi:10.1016/j.advwatres.2008.04.012, 2008.

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:01410.01029/02006wr005032, 2007.

Sugawara, M.: Automatic calibration of the tank model, Hydrol. Sci. B., 24, 375–388, 1979.

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.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13–26, doi:10.5194/hess-5-13-2001, 2001.

Westerberg, I., Walther, A., Guerrero, J.-L., Coello, Z., Halldin, S., Xu, C. Y., Chen, D., and Lundin, L.-C.: Precipitation data in a mountainous catchment in Honduras: quality assessment and spatiotemporal characteristics, J. Theor. Appl. Clim., 101, 381–396, doi:10.1007/s00704-009-0222-x, 2010.

Westerberg, I., Guerrero, J.-L., Seibert, J., Beven, K. J., and Halldin, S.: Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras, Hydrol. Process., 25, 603–613, doi:10.1002/hyp.7848, 2011.

Widen-Nilsson, E., Halldin, S., and Xu, C. Y.: Global water-balance modelling with WASMOD-M: Parameter estimation and regionalisation, J. Hydrol., 340, 105–118, 2007.

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:12410.11029/12009wr007706, 2009.

Vogel, R. M., and Fennessey, N. M.: Flow-Duration Curves. 1: New Interpretation and Confidence-Intervals, J Water Res Pl-Asce, 120, 485-504, 1994.

Wood, S. J., Jones, D. A., and Moore, R. J.: Accuracy of rainfall measurement for scales of hydrological interest, Hydrol. Earth Syst. Sci., 4, 531–543, doi:10.5194/hess-4-531-2000, 2000.

Xu, C.-Y.: WASMOD – The water and snow balance modeling system, in: Mathematical Models of Small Watershed Hydrology and Applications, edited by: Singh, V. J. a. F., D.K., Water Resources Publications LLC, Highlands Ranch, Colorado, US, 555–590, 2002.

Xu, C.-Y. and Halldin, S.: The effect of climate change on river flow and snow cover in the NOPEX area simulated by a simple water balance model, Nord Hydrol., 28, 273–282, 1997.

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.

Younger, P. M., Freer, J. E., and Beven, K. J.: Detecting the effects of spatial variability of rainfall on hydrological modelling within an uncertainty analysis framework, Hydrol Process, 23, 1988–2003, doi:10.1002/Hyp.7341, 2009.

Younger, P. M., Beven, K. J., and Freer, J. E.: Limits of acceptability and complex error reconstruction in a rainfall-runoff simulation, J Hydrol, in review, 2011.

Yu, P. S. and Yang, T. C.: Using synthetic flow duration curves for rainfall-runoff model calibration at ungauged sites, Hydrol. Process., 14, 117–133, 2000.