A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic

Sarka Blazkova[1] and Keith Beven[2,3,4]

Received 3 December 2007; revised 5 February 2009; accepted 6 April 2009; published 27 June 2009.

[1] In this study continuous simulation flood frequency predictions on the Skalka catchment in the Czech Republic (672 km^2, range of altitudes from 460 to 1041 m above sea level), are compared against summary information of rainfall characteristics, the flow duration curve, and the frequency characteristics of flood discharges and snow water equivalent using the generalized likelihood uncertainty estimation limits of acceptability approach outlined by Beven (2006). Limits of acceptability have been defined, prior to running the Monte Carlo model realizations for subcatchment rainfalls, discharges (using rating data) at 5 sites within the catchment, and snow water equivalent in 13 snow zones, 4 of which have observed data. Flood frequency and flow duration data at the outlet of the whole catchment are not used in the evaluation but are used to test the predictions. In order to get sufficient behavioral models to assess adequately the prediction uncertainty it was necessary to refine the model structure, sample the model space more densely, and, in the end, relax the limits of acceptability to allow for a strong realization effect in predicted flood frequencies. We use a procedure of scoring deviations relative to the limits of acceptability to identify the minimum extension of the limits across all criteria to obtain a sample of 4192 parameter sets that were accepted as potentially useful in prediction. Results show that individual model realizations, with the same parameter values, of similar length to the observations can vary significantly in acceptability. Long-term simulations of 10,000 years for retained models were used to obtain uncertain estimates of the 1000 year peak and associated flood hydrographs required for the assessment of dam safety at the catchment outlet.

Citation: Blazkova, S., and K. Beven (2009), 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.

1. Introduction

[2] The estimation of flood frequency by continuous simulation provides an alternative method to direct statistical estimation for catchments where there are limited historical records of flood peaks. It was first applied using long observed rainfall records to drive a hydrological model calibrated on observed discharges [e.g., *James*, 1965]. *Beven* [1986a, 1986b, 1987] first used a calibrated stochastic rainfall model as the input to a continuous simulation model as an extension of the analytical storm-based derived distribution approach of *Eagleson* [1972], and also first took account of the realization effect of reproducing the characteristics of a short record of flood peaks. Later studies using continuous simulation for flood frequency estimation include *Cameron et al.* [1999, 2000a], *Lamb* [1999], *Calver et al.* [1999], *Lamb and Kay* [2004], and *Cameron* [2006] in the UK and *Blazkova and Beven* [1995, 1997, 2002, 2004] in the Czech Republic.

[3] This approach has the potential to represent properly the way in which rainfall characteristics, antecedent conditions in a catchment, and flood runoff generation processes change with time and severity of an event. Both rainfall and runoff generation can also vary in space, particularly on larger catchments, in a way that might be important in predicting flood peak magnitudes. This potential, however, is dependent on being able to specify good model structures and the parameters for rainfall and runoff models in the face of limited data availability. Thus, there will inevitably be some uncertainty associated with both the identification of parameter values [e.g., *Cameron et al.*, 1999; *Blazkova and Beven*, 2002, 2004; *Blazkova et al.*, 2002], particularly for ungauged catchments [e.g., *Blazkova and Beven*, 2002; *Lamb and Kay*, 2004; *Jones and Kay*, 2007], and in the prediction of future rainfall forcing [e.g., *Cameron et al.*, 2000a; *Kay et al.*, 2006].

2. Novelty of the New Methodology

[4] In previous applications of this methodology in the Czech Republic these uncertainties have been evaluated by

[1]T. G. Masaryk Water Research Institute, Prague, Czech Republic.
[2]Lancaster Environment Centre, Lancaster University, Lancaster, UK.
[3]Geocentrum, Uppsala Universitet, Uppsala, Sweden.
[4]Laboratoire d'Ecohydrologie, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.

0043-1397/09/2007WR006726$09.00

considering the residuals between observed and predicted data within the Monte Carlo realization based generalized likelihood uncertainty estimation (GLUE) methodology [*Beven and Binley*, 1992; *Beven and Freer*, 2001; *Beven*, 2006, 2009]. This includes a past study of a large catchment [*Blazkova and Beven*, 2004] with a lower elevation range and more uniform spatial pattern of inputs than in the Skalka catchment studied here. Here, this is extended in a number of novel ways, relative to earlier studies of estimating flood frequency by continuous simulation. In particular, (1) the evaluation of model realizations makes use of the GLUE limits of acceptability approach outlined by *Beven* [2006], with model realizations evaluated against 114 different limits of acceptability; (2) the evaluation of discharge predictions takes account of the observational errors in discharges derived from rating curve interpolation and extrapolation using fuzzy regression; (3) the stochastic model of inputs attempts to reproduce and evaluate the strong spatial patterns of rainfall, snow depths and temperatures across a larger catchment; and (4) both the hydrological model and stochastic inputs models were modified as a result of total model rejections early in the study. Within this version of GLUE, all models that survive the multicriteria tests of acceptability are then used in prediction of the frequency characteristics for the catchment, weighted by a likelihood measure that depends on performance of the model within the limits of acceptability.

[5] Following a brief description of the catchment studied and the semidistributed version of the rainfall-runoff model Topography-Based Model of Catchment Hydrology (TOPMODEL) [e.g., *Beven*, 1986a, 1986b, 1987, 2001] used in the simulations, the extended GLUE multiple limits of acceptability calibration strategy is described. In this strategy, models are treated as hypotheses about system response, to be rejected if the predictions fall outside of the limits of acceptability [*Beven*, 2006]. In the initial phase of this study, all the models tried could be rejected under the defined limits of acceptability. This led to modifications of both the stochastic rainfall model and the hydrological model to try and improve the description of the catchment system. There are many details of the study that are not easily summarized in a short paper, but more details of the component models have been included in the auxiliary material to this paper.[1]

3. Skalka Catchment

[6] The Skalka catchment in the headwaters of the Eger River lies across the border of the Czech Republic and Germany. Eger (Ohre in Czech) is a tributary of the River Elbe. The Eger catchment down to the Skalka dam has an area 671.7 km^2, an average altitude of 592 m above sea level, with a range from 460 m to 1041 m (Figure 1). Past observations of discharges are available at a number of subcatchment gauging stations (see Table 1). The common period of observation for the available precipitation gauges is 1971–1999, i.e., 29 years, from which the annual average precipitation has been computed. The subcatchments have been divided into 13 elevation zones for snow accumulation and melt computation (Table S2). The gauging station at Cheb, downstream of the Eger and Roslau subcatchments, for which data were available from an earlier 60 year period prior to the construction of the Skalka dam, has not been used in the model calibration process but has been retained for later evaluation of the model predictions.

[7] Observed floods at Hohenberg are markedly smaller than on the Roslau in spite of the fact that the Eger upstream of Hohenberg is about the same area as the Roslau upstream of Arzberg. This can be explained, at least in part, by the geologies of the Eger and Roslau catchments which are rather different. The Eger subcatchment is much less geologically variable and consists mostly of granites, with some phyllites, slates, and quartz porphyry. There are also tertiary and quaternary sediments associated with the Eger graben and some local storage in weathered granite. The Roslau subcatchment has a variety of relatively impermeable rocks, but also some limestones and dolomite, with local karst development (www.geologie.bayern.de). We therefore expect that different model parameterizations will be needed in representing the hydrology of the different subcatchments.

[8] There are also significant spatial differences in mean annual rainfalls across the subcatchments (Table S1). In the lower part of the Eger (subcatchment 2) there are some ponds and some small hydropower schemes, which should not have a significant effect on floods but which might affect the flow duration curves.

4. Modifications to the Flood Frequency Version of TOPMODEL

[9] A continuous simulation flood frequency version of TOPMODEL [*Beven and Kirkby*, 1979; *Beven*, 1987, 2001] has been used in this study, with multiple subcatchments the outputs of which are routed using a constant wave velocity, map-derived, width function algorithm [*Beven*, 1979]. Snowmelt is an important driver for floods in this catchment and a multiple elevation snow accumulation melt component has been introduced. A number of other important changes in the model from the version used by *Blazkova and Beven* [2004] have been implemented as a result of initial runs in the study that revealed deficiencies in the simulations for this catchment, and to take advantage of having more data available for the estimation of the *m* storage depletion parameter. This parameter controls the shape of the recession. With only one value of *m* (as in the original TOPMODEL structure) it was not possible to model low flows correctly, particularly at the Hohenberg site on the Eger. The modeled recession discharges were too low which was obvious on the modeled hydrographs and in the flow duration curve. Water balance calculations suggested that this might in part be due to an underestimation of rainfall inputs (see next section). There are also a number of small detention ponds in this subcatchment which might also have had an effect. However, consideration of the geology also suggested that the original exponential transmissivity profile assumption of this catchment was not adequate in this catchment under drier conditions.

[10] Unlike the study of *Blazkova and Beven* [2004] in this study discharge data were available for a recession curve analysis to estimate the recession characteristics for the different subcatchments using both hourly and daily data (Table S3). These data have been analyzed with the recession analysis program of *Lamb and Beven* [1997] (MRCtool) using various recession lengths in the program. This analysis allowed the range of the *m* parameters for the subcatchments

[1] Auxiliary materials are available in the HTML. doi:10.1029/2007WR006726.

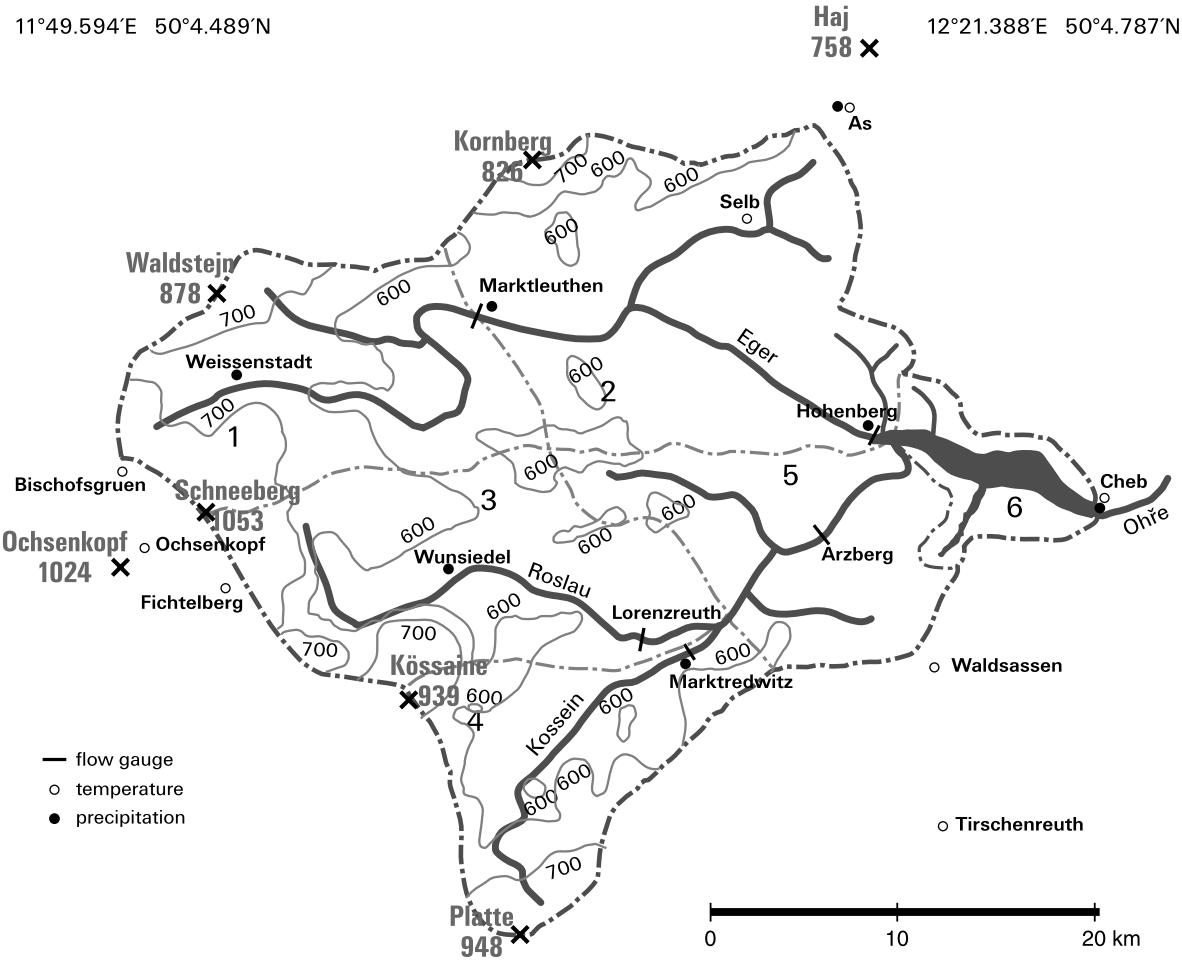

Figure 1. The Skalka catchment on the Ohre (Eger) River with the numbers of subcatchments, flow gauges, temperature, and precipitation stations; dashed-dotted lines are subcatchment boundaries.

to be constrained but also suggested that the base flow component of TOPMODEL should be modified to allow the possibility of a minimum base flow (Q_{min}) maintained by long-term storage in each subcatchment. Q_{min} is sampled in such a way that discharge continuity will be maintained if all subcatchments are at their local Q_{min} value.

[11] Since the flow duration curve at Hohenberg was still not adequately matched in the initial GLUE sampling, a further modification for the lower Eger subcatchment has been introduced such that an additional deficit (*SDref*) has to be satisfied before fast surface runoff occurs. More complete details of these modifications to the hydrological model are given in the auxiliary material, including a full table of the model parameters varied in the Monte Carlo sampling for each subcatchment (Tables S4–S6). This process of modification of the model in different subcatchments is the type of learning process suggested by *Beven* [2007].

5. Precipitation and Temperature Model for a Large Catchment With a Strong Elevation Gradient

[12] The hourly time step stochastic precipitation model used in the study is based on that used in past studies of this type in the Czech Republic [*Blazkova and Beven*, 1995, 1997, 2002, 2004]. The features of this model include the

Table 1. Flood Statistics for Gauged Discharge Stations in the Skalka Catchment[a]

Subcatchment or Interbasin	Area (km^2)	Cumulative Area (km^2)	Water Gauge Station (Figure 1)	Catchment Area of Station (km^2)	Number of Years of Observation	Mean Flood (m^3/s)	Estimated 100 Year Return Period Flood[b] (m^3/s)
Upper Eger	114.420	114.42	Marktleuthen on Eger	114.42	67 (1937–2004)	20.28	50.37
Lower Eger	209.376	323.796	Hohenberg on Eger	298.8	37 (1967–2004)	29.94	74.38
Upper Roslau	130.65	130.65	Lorenzreuth on Roslau	122.2	39 (1966–2004)	22.64	56.25
Kossein	94.82	94.82	Marktredwitz on Kossein	72.2	34 (1971–2004)	15.91	39.52
Lower Roslau	91.11	316.58	Arzberg on Roslau	290.04	28 (1977–2004)	53.86	133.80
Ohre	31.3240	671.7	Cheb on Ohre	683.34	60 (1887–1959)	89.98	223.52

[a]Data from Landesamt fuer Wasserwirtschaft Muenchen and CHMI.
[b]Estimated on the basis of *Hosking and Wallis* [1997] using the software of *Hosking* [1997].

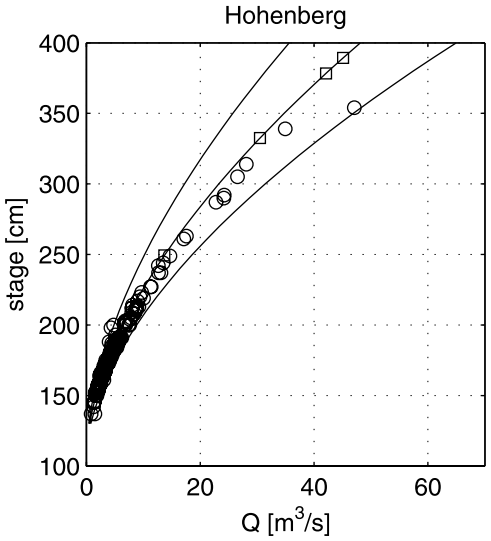

Figure 2a. Rating curve at Hohenberg; fuzzy estimate and bounds using HBS1 method. Here circles are observed points of current metering, squares are selected plotting positions used as criteria, and Q (m^3/s) is the discharge.

separate treatment of low-intensity and high-intensity events. In this larger catchment there is a greater range of elevation than in previous applications of the approach and so the precipitation model has been extended to take account of the gradient of precipitation in this catchment, the accumulation and melting of snow in 13 different elevation bands in the catchment, the potential for storms to move across the catchment from different directions on the basis of meteorological classification [*Hydrometeorological Institute*, 1967; *Kakos*, 1977, 1978], the significant input of fog drip in the upper parts of the catchment [*Tesař et al.*, 1995], and the potential for some events to produce rainfalls only over the lower part of the catchment. Full details are given in the auxiliary material. These changes improved the performance of the model in fitting the flow duration curves. Complete details of the semidistributed rainfall and temperature models are given in the auxiliary material.

6. Model Evaluation Using Limits of Acceptability

[13] In the manifesto for the equifinality thesis [*Beven*, 2006] an approach to model evaluation on the basis of limits of acceptability for use within the GLUE methodology was outlined. In this approach it was suggested that the limits of acceptability be defined prior to making runs of a model on the basis of a careful assessment of the potential effects of observation error and input errors to define a range of "effective observational error." Behavioral models are then those that provide predicted variables that fall within the limits of acceptability. The performance of each model can still be associated with a likelihood weight that summarizes how close the predictions of the model are to the original observations. Here we are comparing model outputs against summary information of the flow duration curve and the frequency characteristics of flood discharges, snow water equivalent and hourly and daily rainfall frequencies. To ensure that modeled floods maintain an appropriate seasonal distribution, an evaluation of the proportion of winter floods is also made. Details of how the limits of acceptability were defined across the multiple evaluation criteria are as follows.

6.1. Rainfalls

[14] In the case of rainfall there is a commensurability or representational error in going from the point raingauge information (see Figure 1) on hourly and daily rainfall frequencies to the subcatchment values. We also found it necessary to modify the stochastic rainfall model to include a component of fog drip and a correction to the subcatchment annual average precipitation due to elevation. Thus the subcatchment rainfalls generated by the rainfall model component cannot easily be directly compared with the gauged data. Limits of acceptability were defined by putting bounds on the subcatchment hourly and daily rainfall frequency information estimated from the point raingauge sites by bootstrapping from the distributions of estimates of the average subcatchment rainfalls from the point measurements. Excluding one measurement point at a time and reestimating the subcatchment rainfalls allows a range of possible subcatchment rainfalls to be defined directly from the measurements. The largest percentage of model rejections (about 3 per cent) was on the lower bound for the daily averages in the headwater subcatchments.

6.2. Flow

[15] In the case of flow, the model realizations are compared against flood frequency characteristics at the gauging sites shown in Table 1 and Figure 1, allowing for error in the observed flood peak discharges. This involves the estimation error associated with going from the direct measurements of water level to a discharge through a rating curve. The errors involved have been assessed by obtaining the original flow measurements used to determine the rating curve that are available at the Marktleuthen, Hohenberg and Arzberg sites. Limits of acceptability were defined as fuzzy bounds for a log-log fuzzy regression derived using the HBS1 algorithm of *Hojati et al.* [2005]. Figure 2a shows the resulting observation uncertainties at Hohenberg. Rating curves for Marktleuthen and Arzberg are shown in the auxiliary material (Figure S2a and S3a). It is worth noting that for these sites the 95% prediction bounds of a log-log statistical regression were only slightly narrower. The advantage of fuzzy regression in this case is that it can take account of uncertainty in the individual rating curve measurements, which is known to be significant for higher discharges. The uncertainties in discharge measurement have been used in setting limits of acceptability for the flow duration curves at these sites. For the evaluation of the predicted flow duration curves, nine different quantiles have been used in model evaluation (from about 25 to 90% exceedence).

[16] The rating curve uncertainties are also used in setting limits of acceptability for flood frequency. In model evaluation for each site only the plotting positions for events close to the 1.07, 2, 5 and 10 year return periods (ev1 reduced variate values of -1, 0.37, 1.5, and 2 to 2.25) have been used to reduce the additional effect of uncertainty in estimating exceedence probabilities for longer return periods given only relatively short periods of observations. Uncertainty in the plotting position has been allowed for by fitting a fuzzy linear regression through the observations, with plotting position expressed in terms of the ev1 reduced variate (Figure 2b, also

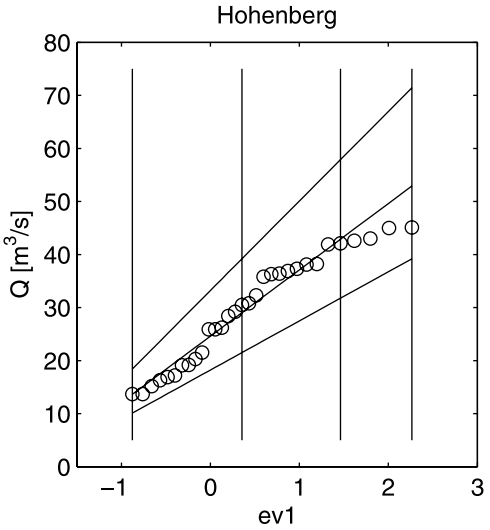

Figure 2b. Hohenberg fuzzy regression with uncertainty of Q on ev1 reduced variate in the linear range, i.e., approximately between ev1 reduced variate = −1 and 2. Vertical lines show plotting positions which have been used as criteria for computing scores. Q is the discharge, ev1 is the axis of the Gumbel distribution reduced variate, and circles are the observed annual maxima used for computing the linear relationship.

Figures S2b and S3b). Again the HBS1 algorithm has been used (for details, see the auxiliary material), with the observation uncertainties from the rating curve input to the analysis. In Figures 2b, S2b, and S3b the estimated discharge peaks are also plotted at their estimated plotting positions.

6.3. Snow

[17] In the case of snow water equivalent, there is also an issue of commensurability or representational error [see *Beven*, 2006] in that the estimates of frequency quantiles are derived from the measurements at particular locations but the model is predicting average snow water equivalents over different elevation zones in a way that tries to take account of the general increase of precipitation with elevation. Clearly only some of the 13 elevation zones contain one of the 4 measurement sites, and even within those zones there is the possibility of heterogeneity of snow water equivalents in space and time due to land cover, aspect, wind drift and all the other processes affecting the snowpack.

[18] The starting point for setting limits of acceptability in this case was the estimates of maximum annual snow water equivalent for different probabilities of exceedence at each of the four observation sites (on the basis of 10, 10, 35, and 39 years of record at Hohenberg, Weissenstadt, Cheb, and As, respectively, Table S7). Because of the relatively short lengths of observed records available only the sample estimates of return period of 2 years were used, again to reduce the effect of the additional sampling uncertainty for exceedence probabilities for longer return periods. A fuzzy regression using the HBS1 method was then carried out for the 4 points against elevation as a way of extrapolating from the measurement sites to all 13 elevation zones. Uncertainty in estimating the snow water equivalent frequency in each zone was computed by bootstrapping the observed data in each of the four stations as an input to the regression (Figure 3). Given that there are only four observed points, fuzzy regression here has the advantage of allowing some constraint over the variability in snow accumulation in the different elevation zones since here a statistical regression produces unrealistically wide prediction limits. On the basis of the fuzzy regression, limits of acceptability were set for each zone.

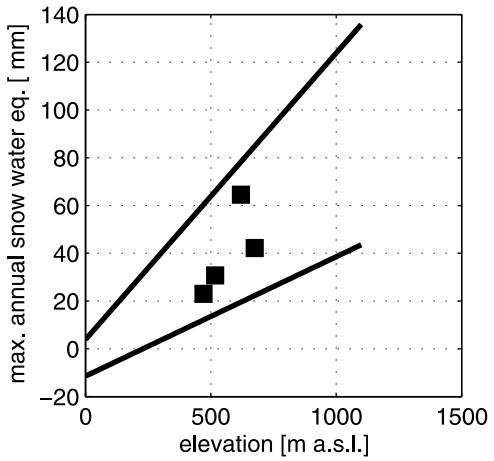

Figure 3. Fuzzy bounds for the dependence of maximum annual snow water equivalents on elevation for median; squares are the medians of observed data.

6.4. Season of the Annual Peak

[19] In the Skalka catchment the annual peak can come in any season but the vast majority occur in wintertime. Table 2 shows the percentage of annual peaks in winter for each observed series. On the basis of these data, limits of acceptability and a trapezoidal weighting function have been constructed to allow an additional fuzzy constraint on the model realizations (Figure 4).

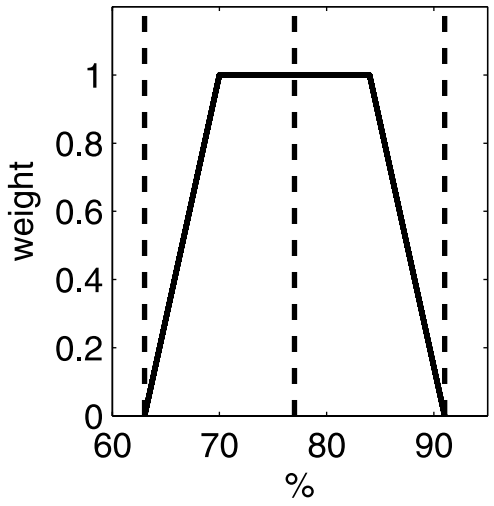

Figure 4. Weighting function for evaluating the percentage of annual floods occurring in winter; black dashed lines are an estimate from observations with the acceptability limits and the solid line is the trapezoidal weighting function.

Table 2. Percentages of Annual Floods Occurring in Winter

Subcatchment or Interbasin	Water Gauge Station (Figure 1)	Number of Years of Observation	Season 1	Season 2	Season 3	Season 4	Winter in Percent of Years
Upper Eger	Marktleuthen on Eger	67	56	3	4	4	83.6
Lower Eger	Hohenberg on Eger	37	31	2	2	2	83.8
Upper Roslau	Lorenzreuth on Roslau	39	32	2	3	2	82.1
Kossein	Marktredwitz on Kossein	34	24	0	8	2	70.6
Lower Roslau	Arzberg on Roslau	28	25	0	1	2	89.3
Ohre	Cheb on Ohre	60	43	4	9	4	71.7

6.5. Initial Model Evaluations

[20] Each model run requires parameter values to be defined for the precipitation, temperature and hydrological model components. These are generated randomly by taking independent samples from within specified ranges (see Tables S4 and S5), since little information is available to be able to specify how these parameters might interact. Initial runs were used to assess sensitivity of individual parameters and suggested that some parameter values could be fixed (see Table S4). Each model run is then assessed as to whether the results are within the specified limits of acceptability on the following measures: (1) four selected plotting positions (at approximately ev1 reduced variate = −1, 0.37, 1.5 and 2 to 2.25) in subcatchments 1 to 5; (2) nine quantiles of flow duration curve in subcatchments 2 and 5; (3) the median (2 years return period) of maximum annual snow water equivalent in 13 snow zones; and (4) percentage of annual floods occurring in winter in 6 subcatchments.

[21] Taking account of both upper and lower bounds, this gives a total of 114 limits of acceptability that we would like acceptable models to satisfy. The predam construction flood frequency curve and flow duration curves at the dam site at Cheb (Figure 1) are not used in evaluation and are left for validation.

[22] In evaluating the model predictions, we would wish that all the predictions of a behavioral model should lie within the limits of acceptability for all the evaluation criteria. Within the GLUE methodology, the set of behavioral models is then used in prediction, each weighted by some likelihood measure that summarizes performance over all the evaluation criteria. Here a trapezoidal weighting function has been used for each criterion. Figures 5a and 5b show, for the estimate of the approximately 10 year flood at the Marktleuthen site, the limits of acceptability (point b–d) and the trapezoid used in defining the weighting function. The trapezoid allows a central area of maximum weight (halfway between the points b–c and c–d, respectively) that can be used to reflect an estimate of the error in the original discharge measurements used to generate the rating curves.

[23] There is an important realization effect in producing model outputs for flood peaks, flow duration curves and snow water equivalent estimates, depending on the length of the simulation. In comparing model outputs with observations

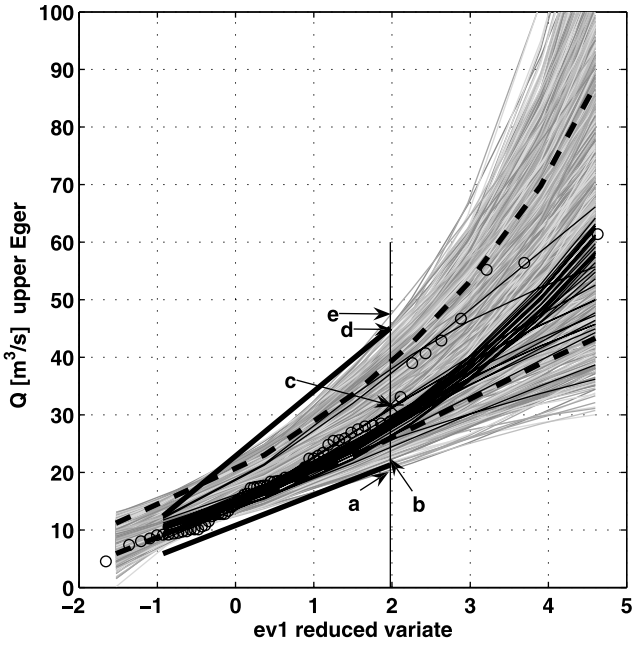

Figure 5a. Flood frequency curve at Marktleuthen (subcatchment 1); thick solid black lines are the initial acceptability limits, circles are the observed annual floods, gray lines are the 4192 simulations, dashed lines are the 5 and 95% uncertainty bounds from the trapezoidal weighting, and thin solid black lines are the behavioral simulations with scores on all criteria <1. Points a to e corresponding to Figure 5b.

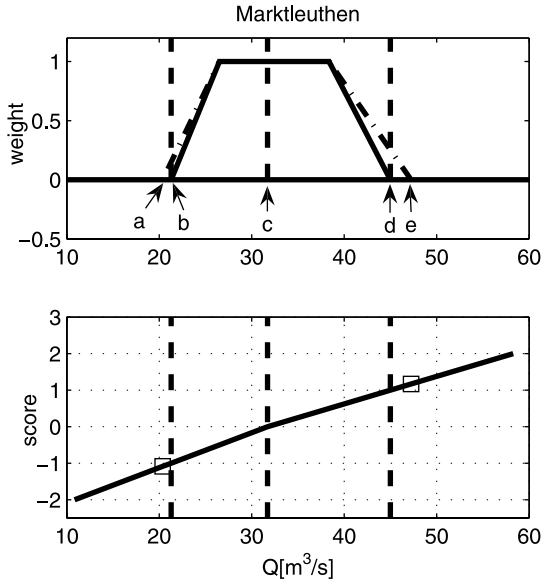

Figure 5b. (top) Trapezoidal weighting function; full line trapezoid is the original limits of acceptability, dashed lines are the estimate and acceptability bounds, and dashed–dotted are the lines of the expanded trapezoid. Points a–e correspond to the Figure 5a. (bottom) Scores; squares are points to which the bounds have to be expanded.

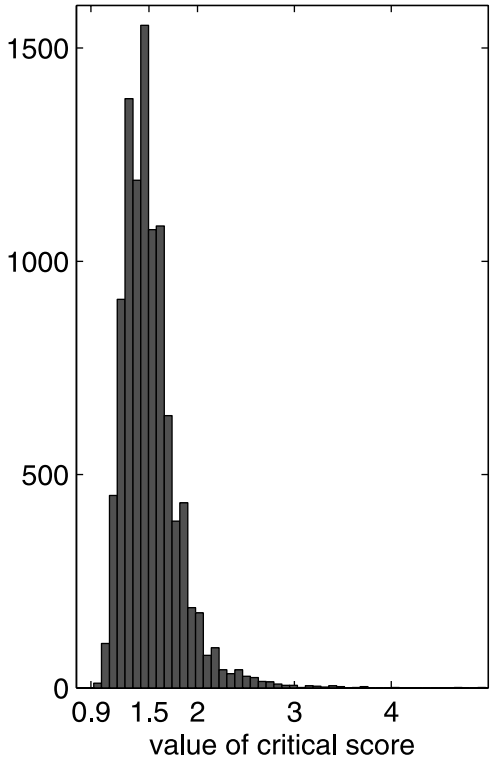

Figure 6a. Distribution of critical normalized scores for a single parameter set with all absolute scores <1 in original simulation over 10,000 input realizations. The normalized score is the scaled deviation within the range from the minimum (set to −1) to maximum (set to +1) original limits of acceptability. The critical score of a simulation is the largest absolute value of the score from all the 114 limits of acceptability when normalized in this way. Values >1 represent realizations with critical scores outside the original limits of acceptability on one or more criteria.

it is then important to use the same record length, since the observations are only one possible realization of all possible sequences of floods for a particular site. Thus the length of simulation for the initial model evaluations is 67 years (longest observed series) but the evaluations for each observation site are based on the same length of record as the observations in each case. Tests of running multiple simulations for a fixed parameter set showed that this realization effect could be significant, even when comparing with the 67 years of record at Markleuthen (see discussion of the results below). This realization effect will also affect the evaluation of the results at the Cheb gauging station that is used as the ''validation'' site. This 60 year record is for a period mostly before the observations at the other sites.

7. Results

7.1. Calibration of Parameter Ranges

[24] Some preliminary calibration of parameter ranges was done to find out if there were any simulations, which could fulfill all the criteria. It became clear that the lower Eger subcatchment needs different ranges than the other subcatchments for a number of the parameters. This, however, was not enough to bring the model realizations acceptably close to the observed frequency and duration estimates. Thus the improvements to the model structure described above were gradually introduced. Using the parameter ranges of Tables S4 and S5, parameter sets were sampled randomly. With all the improvements to the model components implemented, a total of 610,000 short simulations of 67 years with an hourly time step were run. A parallel system of 22 PCs (most of them with 2 processors) was used running the open Mosix system which distributes the model runs automatically across the available processors. 50 tasks of 50 short simulations each are handled at the same time. Computation of the 67 year realizations required about 90 days of computation on the parallel system.

[25] Only 39 behavioral simulations were found that fulfilled all the acceptability criteria (though we note that no other study, to our knowledge, has attempted to evaluate a rainfall-runoff model on so many separate, albeit related, criteria). There are a number of options at this point. One would be to sample the model space more densely to check that areas of behavioral models are not being missed. 610,000 samples does not represent a dense sample when sampling a model space of 46 dimensions, especially given the realization effect of short simulations noted above. We could also try to refine the prior estimates of model parameter ranges, distributions and interactions, should information be available to do so (but normally the only real information is an analysis of the posterior behavioral parameter sets). Another option would be to add a statistical model inadequacy component (following, for example, *Kennedy and O'Hagan* [2001]) though it is not clear that it would be possible to formulate simple statistical error models across all sites and types of measures in this case. A further option, that has been taken here, is to relax the limits of acceptability to obtain a wider sample of models that will be treated as behavioral, and then to evaluate whether that sample is fit for purpose. One justification for this is the effect of different input realizations on acceptability. Taking just one of the behavioral parameter sets and generating 10,000 input sequences of the same length as the observed flood series results in a range of critical values for acceptability (Figure 6a) across a range of evaluation criteria (Figure 6b; see Figure S8 for a more complete demonstration of the realization effect). It follows that other

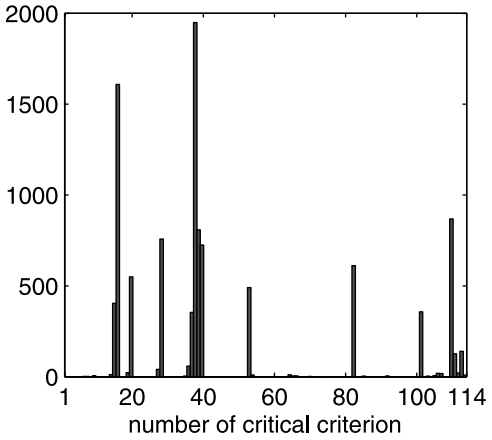

Figure 6b. Histogram of occurrences of the individual critical scores; on the *x* axis are the 114 criteria of acceptability. For the simulations of the Figure 6a the criterion number 38 (underprediction of low-magnitude floods at Arzberg) was critical most often.

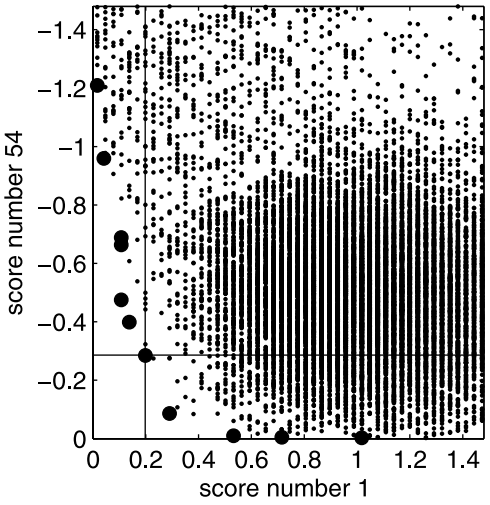

Figure 7. A simplified case of Pareto front constructed from two contradictory criteria: the overprediction of smaller floods at Marktleuthen and the underprediction of flow duration at Hohenberg. Here 19,460 simulations (small circles) that can be treated as acceptable on those two criteria with critical scores of 1.48 are shown; big circles are the simulations on the Pareto front.

parameter sets with relatively low critical values for acceptability might be fully behavioral given other input realizations. Computing constraints mean that this realization effect cannot be explored fully. It should also be noted that this reveals a constraint on how well a particular realization of observations is representative of possible future probabilities of extremes and how well we might expect a model conditioned on data that was mostly collected after the construction of the Skalka dam might reproduce the predam data at the Cheb site that is being used as verification data in this study. We feel that the demonstrated realization effect is therefore an adequate reason to relax the limits of acceptability, but will return to this issue in the discussion of the results.

7.2. Relaxing the Limits of Acceptability

[26] Relaxing the limits can then be considered as a multicriteria minimization problem. We want to minimize the degree to which we must expand the limits to obtain a given number of models that are acceptable across all the criteria. This requires putting the different evaluation measures on a common scale. This is easily achieved within the limits of acceptability approach by treating each evaluation in terms of a normalized score that has the value −1 at the lower limit, 0 at the observed value and +1 at the upper limit, noting that the scaling can then be nonsymmetric for under- and overprediction limits (see Figure 5b). Values outside the predefined limits of acceptability will have normalized score values either less than −1 or greater than +1. The model runs can then be ranked in terms of the maximum excursions away from the initial limits for each of the 114 measures being considered. The minimization can then be treated as a Pareto set problem (details are given in the auxiliary material) to retain models that are not dominated on any single criteria by any other model as the absolute value of all the scaled limits are increased together. The degree of relaxation can be examined in obtaining sets of acceptable models of different size. In this way a set of best models (in terms of the normalized scores) can be defined. Note that this is not the same as the multicriteria Pareto optimal set used by *Gupta et al.* [1998] and others, since the methodology used here will include any models that satisfy the extended limits but which are behind the Pareto front (as shown in Figure 7). Table 3 shows how more simulations are included in the acceptable set as the allowable scores are gradually increased.

[27] After computing 610,000 simulations, the 4192 best simulations have been selected on the maximum score measure (absolute score values up to the value of 1.48, i.e., requiring expansion by a factor of 1.48). Some criteria, however, are easier to match than others. The limits did not have to be relaxed at all on 44 out of the 114 criteria. These are, mostly, for overprediction (positive scores) in the flow duration quantiles and for either overprediction or underprediction in different elevation zones of the median of annual maximum snow water equivalent.

[28] Table 4 shows the limiting criteria as the critical normalized score is gradually increased for the ranked models in the set. Difficult criteria to match are maximum snow water equivalent in the three highest-elevation zones where there is no direct observation and the estimates have been extrapolated, albeit with uncertainty. These also occupy relatively small areas (12.5, 6.25 and 8.16% of the subcatchment area, respectively). The percentage of winter floods can be both over and underpredicted. The most difficult criteria to match are the underprediction limits on the flow duration curves at Hohenberg (Figure 8) and Arzberg (Figure S5).

7.3. Defining a Likelihood Measure After Relaxing the Limits of Acceptability

[29] Having defined a set of 4192 acceptable models by expanding the scores in this way within the GLUE methodology it is also possible to assign different likelihood weights to each model according to how well it has performed in the evaluations. Here, a trapezoidal function has been used for each of the criteria, such that after expansion of the limits, all models will have a positive weight on all the evaluation criteria (Figure 5b). The central part of the trapezoid corresponding to observation error is given a maximum weight and the weight for that criterion reduces to zero at the maximum positive and negative scores after expansion. The individual weights were then combined by taking the sum of the individual weights over all the criteria and rescaling such that the sum of weights for all models in the behavioral set is unity.

[30] The resulting prediction quantiles for flood frequency estimates for each of the gauging sites are shown in Figures 5a, 9, S5, S6, and S7 in comparison with the observed

Table 3. Number of Behavioral Simulations (From a Total of 610,000) as Critical Score for Acceptance Over All Criteria is Increased

Limits Relaxed by	Number of Behavioral Simulations	Maximum Score
0.00	39	1.00
0.25	597	1.25
0.30	1037	1.30
0.35	1603	1.35
0.40	2311	1.40
0.45	3219	1.45
0.48	4192	1.48

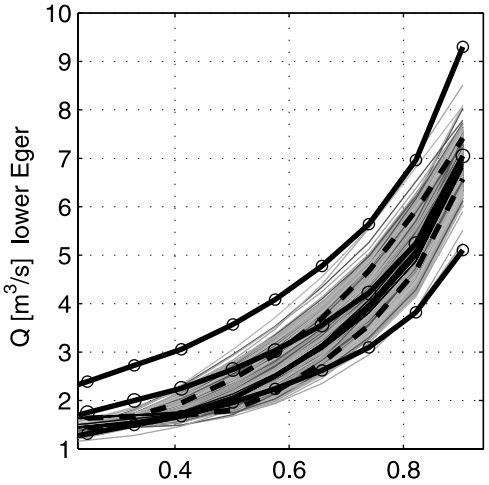

Figure 8. Flow duration curve at Hohenberg; thick solid black lines are the flow duration curve from observed data with acceptability limits, circles are the quantiles, gray lines are the 4192 simulations, thin solid black lines are the behavioral simulations with scores on all criteria <1, and dashed lines are the 5 and 95% uncertainty bounds from the trapezoidal weighting.

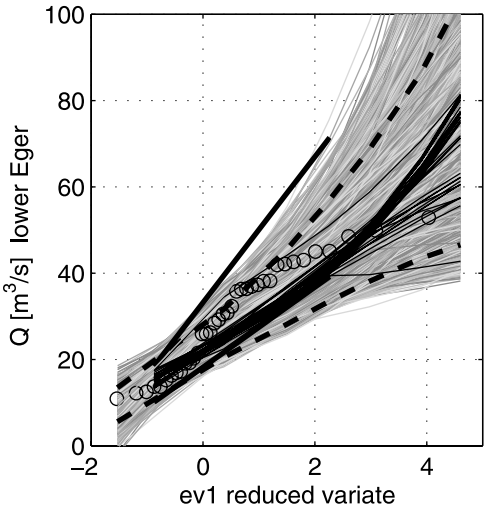

Figure 9. Flood frequency curve at Hohenberg; thick solid black lines are the acceptability limits, circles are the observed floods, gray lines are the 4192 simulations with scores on all critera <1.48, dashed lines are the 5 and 95% uncertainty bounds from the trapezoidal weighting, and thin solid black lines are the behavioral simulations with scores on all criteria <1.

series and the acceptability limits (simulations are shown as fitted Wakeby distributions). At Marktleuthen (Figure 5a) the prediction limits span the (uncertain) observations well, while at Hohenberg (Figure 9), despite the difficulty of reproducing the full flow duration curve (Figure 8), the observed flood frequency curve lies within the 5 and 95% bounds of the simulations, albeit close to the upper bound. The uncertainty in the predicted discharges at the 100 year return period (ev1 reduced variate = 4.61) is very high for these short simulations.

Table 4. Critical Scores for Acceptance of Ranked Simulations With Description of Limiting Criterion

Rank	Score	Number of Criterion	Description of Criterion
1	0.91	101	Arzberg snow lower bound
8	0.93	65	Arzberg duration upper bound
9	0.94	15	Hohenberg flood 5 years lower bound
11	0.95	66	Arzberg duration upper bound
12	0.95	52	Hohenberg duration lower bound
14	0.96	20	Lorenzreuth flood 10 years upper bound
15	0.97	40	Arzberg flood 10 years lower bound
25	0.98	16	Hohenberg flood 10 years lower bound
33	1.00	53	Hohenberg duration lower bound
34	1.00	38	Arzberg flood 2 years lower bound
42	1.03	39	Arzberg flood 5 years lower bound
43	1.03	28	Markredwitz flood 10 years upper bound
48	1.05	1	Marktleuthen flood 1 year upper bound
51	1.06	110	Hohenberg winter floods lower bound
52	1.06	37	Arzberg flood 1 year lower bound
70	1.09	112	Markredwitz winter floods lower bound
78	1.10	85	Markredwitz snow highest zone upper bound
79	1.10	82	Lorenzreuth snow highest zone upper bound
83	1.10	111	Lorenzreuth winter floods lower bound
85	1.11	27	Markredwitz flood 5 years upper bound
93	1.11	14	Hohenberg flood 2 years lower bound
507	1.24	70	Arzberg duration lower bound
1052	1.30	64	Arzberg duration upper bound
4192	1.48	19	Lorenzreuth flood 5 years upper bound

7.4. Predictions at the Predam Gauging Site at Cheb

[31] It will be recalled that the data from one site was not used in the model evaluations but was retained for testing the model predictions. This was the site at Cheb, downstream of the junction between the Eger and Roslau, where data were collected for 60 years prior to the construction of the Skalka dam. This site has the second largest sample of annual maximum discharges, but for a different period to the periods at the other sites used in defining the behavioral model set. Figures 10 and 11 show the prediction quantiles for flood frequency and flow duration at the dam site in comparison with observed series of annual floods and Czech Hydrometeorological Institute (CHMI) regional estimate of flow duration (on the basis of regression of statistical characteristics on physical-geographic characteristics which is checked and if needed modified on the basis of relations within the river network [see *Novický et al.*, 1993]). The observed points of the 60 years fall mostly within the range of the simulated prediction limits for this site, except for the very lowest return periods (Figure 10), remembering that this flood record (1887–1959) overlaps only with the record at Markleuthen (1937–2004) and not with any of the other sites used in conditioning the model realizations in GLUE. For the 7 most extreme floods at Markleuthen (above ev1 reduced variate = 2), only one is common with the Cheb period (in 1954), and only 2 are common with all the other sites (1987 and 1999). This is evidence of a realization effect in the observations that affects the shape of the apparent flood frequency curve. From the normalized plot of observed flood frequencies of all sites it is obvious that the earlier period had lower small floods and generally higher floods in the range of ev1 reduced variate about 1 and 2 (Figure S9). Figure 11 also shows a flow duration curve for the Cheb site constructed from 7 years of observation 1931–1938, i.e., also before the period of measurement of most of the other stations.

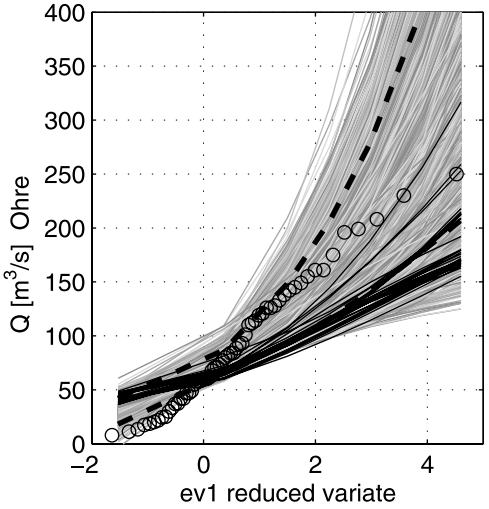

Figure 10. Flood frequency curve at the Skalka dam site (flood series of Cheb station) which was not used in the evaluation; circles are the observed annual floods, gray lines are the 4192 simulations with scores on all critera <1.48, dashed lines are the 5 and 95% uncertainty bounds from the trapezoidal weighting, and thin solid black lines are the behavioral simulations with scores on all criteria <1.

7.5. Long Simulations

[32] Regulations for the assessment of dam safety generally require estimates of high return period discharges. In the past, estimates of the 1000 year event on the basis of statistical analysis of rather short records, or regionalizations based on short records at multiple sites have been used to estimate the 1000 year event. This has then normally been combined with two design hydrographs (from rain and from snowmelt) for evaluating the performance of the dam and spillway under such extreme conditions.

[33] Frequency estimation by continuous simulation allows this restriction to be relaxed. Under the assumption that the models identified above provide an acceptable representation of the hydrology of the catchment, long simulations can be run to get more precise estimates of the frequency characteristics for such long return periods, subject to the uncertainty in representing the current hydrology by the set of 4192 models that have been considered most acceptable. At least 10 times the record length is required to reduce the sampling variability of the required event frequency. Thus, simulations of 10,000 years have been computed with the 4192 models considered acceptable.

[34] This does give rise to a problem in the generation of stochastic rainfalls. The distributions used in the stochastic rainfall models here are infinite tailed. In generating 10,000 years of rainfall events, there is thus a very small but finite probability of generating unrealistically large rainstorms (though see *Cameron et al.* [2000b] for a variation that imposes an asymptotic upper limit to the distribution). UFA (Institute of Atmospheric Physics, Prague) provides estimates of probable maximum precipitation (PMP) in the Czech Republic [*Rezacova et al.*, 2005]. Thus in carrying out the long simulation runs, an additional criterion was imposed as by *Blazkova and Beven* [2004] using 1 h, 1 day and 3 days PMP. There were 341 cases when PMP was exceeded by less than 10% and there was no penalization imposed for that. In 111 realizations rainfall peaks were simulated between 10 and 50% greater than PMP estimates and the weight of the simulations was reduced to one half. In 34 realizations the rainfall extremes were more than 50% larger than PMP estimates and, as a result, these realizations were rejected. In fact only 19 series have been rejected because in 15 series both 1 and 3 day PMP limits were exceeded. Details are in Table S8. The resulting likelihood weighted flood frequency quantile estimates at the dam site from the resulting set of 4173 long model runs are shown in Figure 12a, together with the cumulative likelihood weighted density for the 10, 100, and 1000 year return period peaks (Figure 12b). A frequency analysis of the 60 years of observations at this site again falls within the range of the long-term simulations, except at the very lowest return periods (see the comments above and Figure S9). The long-term simulations show how uncertainty in the flood discharges is high at return periods greater than 100 years, an uncertainty that might be important in assessing dam safety.

[35] In making such assessments, continuous simulation also allows a sample of flood hydrographs associated with the 1000 year events to be saved. These will also have a variety of volumes, which might be significant in assessing the potential for spillway failure.

8. Discussion and Conclusions

[36] The use of continuous simulation for flood frequency estimation is based on the assumption that both the stochastic weather generation and the runoff generation model components are an adequate representation of conditions in the catchment of interest. Here it was found that, even after making modifications to the model, using the limits of acceptability approach within GLUE, only 39 models (from a sample of 610,000 parameter sets) could be found that simultaneously satisfied all the 114 limits of acceptability

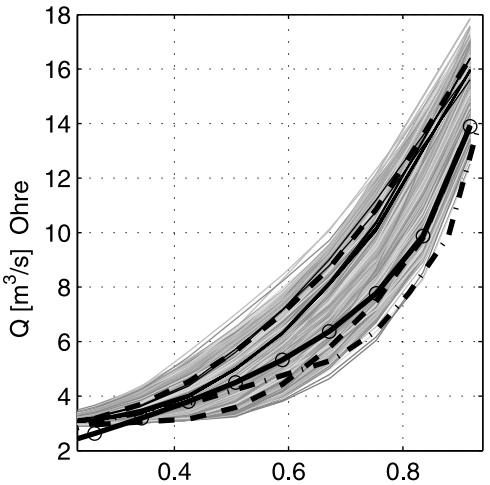

Figure 11. Flow duration curve at the Skalka dam site (series of the Cheb station) which was not used in the evaluation; thick solid black line is the regional estimate of CHMI, circles are the quantiles, dashed-dotted curve is the flow duration from 7 years of observed data at Cheb, gray lines are the 4192 simulations, dashed lines are the 5 and 95% uncertainty bounds from the trapezoidal weighting, and thin solid black lines are the behavioral simulations with scores on all criteria <1.

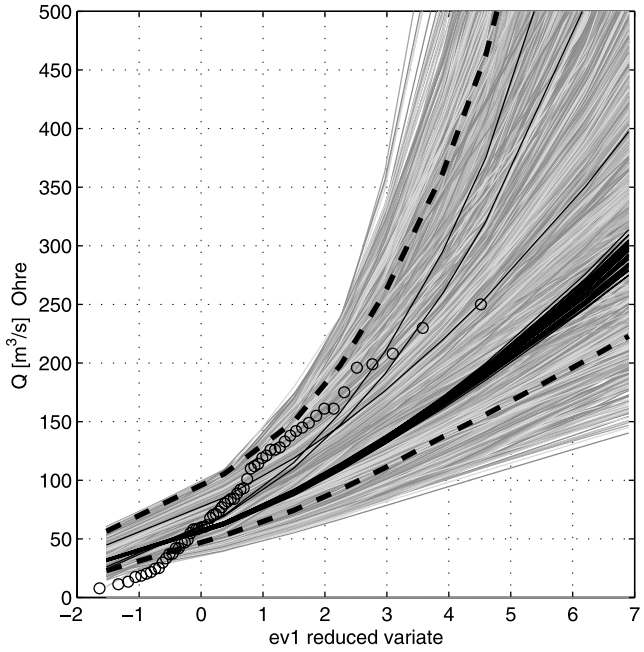

Figure 12a. Flood frequency curve at the Skalka dam site (Cheb station) from the long simulations; ev1 reduced variate is 6.91 for return period 1000 years, circles are the observed annual floods, gray lines are the 4173 simulations with scores on all critera <1.48, dashed lines are the 5 and 95% uncertainty bounds from the trapezoidal weighting, thin solid black lines are the behavioral simulations with scores on all criteria <1.

criteria for evaluation of flood frequency and flow duration at all the gauging sites, median annual maximum snow water equivalent in all elevation zones and proportion of winter floods.

[37] It has been noted that 610,000 runs was a very small sample for a parameter space of 46 dimensions. Thus, it is possible that more models might exist that would satisfy all the requirements, particularly because an examination of the realization effect showed that the critical score for acceptance of a model was dependent on the particular input realization over lengths of record equivalent to the observed flood series (Figures 6a and 6b). It was also the case that some of the ranges of acceptability were based on extrapolations from only a small number of measures that made it difficult to properly estimate the associated uncertainty (e.g., the maximum snow water equivalents in the uppermost elevation bands that proved difficult to satisfy).

[38] Thus, to obtain a set of models that might be useful in prediction, we allowed extension of the limits of acceptability in a way that, by normalizing the limits of acceptability for different evaluation criteria to a common scale, allowed the 4192 best models (in a Pareto sense over all criteria) to be identified with a minimal extension on each of 114 criteria. This is one way of avoiding the rejection of models that might be acceptable given a different input realization (a Type II error) at the expense of increasing the possibility of accepting a poor model (a Type I error). We note that these models might not all be at the Pareto optimal front (Figure 7) but are consistent with the equifinality concepts of *Beven* [2006]. This set of models was shown to provide reasonable estimates of flood frequencies at higher return periods at the gauging sites in comparison with statistical estimates and regionalized estimates. Reproduction of the flow duration characteristics for some of the more base flow dominated sites was less successful. The 4192 retained models were then used in long 10,000 year continuous simulations to provide uncertain estimates of the 1000 year return period peaks required to evaluate dam safety (we could equally analyze the results to obtain estimates of the 1000 year return period maximum flood volumes in a particular time period, or multiple time periods if volume rather than peak flow is more important in the potential for failure of a particular dam). As a product of these continuous simulations it is also possible to provide a sample of hydrographs associated with the 1000 year flood peaks.

[39] Are these predictions then useful, relative to standard methods for estimating the 1000 year event for dam safety evaluation? The process of continuous simulation does allow them to be consistent in mass balance terms, and does allow the nonlinear effects of antecedent conditions on runoff generation to be represented. It also provides a selection of predicted hydrographs with which to test dam safety under extreme conditions. However, the small number of models able to match all the 114 evaluation criteria might suggest that either the stochastic weather model or the runoff generation and routing models (or, in fact, the prior estimates of the limits of acceptability) might need some reconsideration. In particular, it would appear that the low-flow regime on the more permeable lower Eger subcatchment to Hohenberg might need to be improved, while some measurements of snow accumulation on the upper elevation zones would be valuable in improving model evaluation. Such improvements might be useful in constraining the very high prediction uncertainties associated with return periods of greater than 100 years, but at the present time it is important that dam safety assessments recognize the uncertainty in the estimates of both peaks and associated flood volumes.

[40] The value of the extended GLUE limits of acceptability approach in forcing a more rigorous and thoughtful

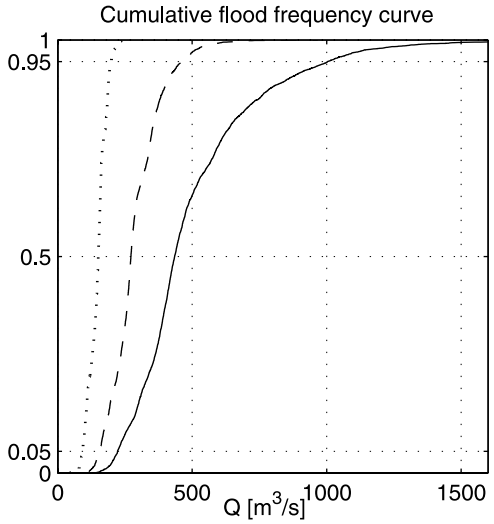

Figure 12b. Cumulative likelihood weighted density for the 10 (dotted line), 100 (dashed line), and 1000 (solid line) year return periods peaks based on 10,000 year runs of 4173 models with critical scores <1.48.

examination of both the observational data and model performance has been demonstrated. A formal procedure for extending the limits of acceptability across multiple criteria to find models that might provide useful predictions has also been demonstrated. Model failures, relative to the original limits of acceptability, provide evidence on which model improvements, reconsideration of observation uncertainties, or user assessments of fit for purpose of a particular model structure and its predictions, might be based.

[41] **Acknowledgments.** The study was supported by the Ministry of Environment of the Czech Republic under the grant SP/2e7/229/07 and its previous grants. Czech data was provided by the Czech Hydrometeorological Institute (CHMI). An important part of the simulations has been carried out at the Lancaster University parallel system. Data from Eger and Roslau have been provided by Landesamt fur Wasserwirtschaft Muenchen partly through the Commission for Boundary Streams. German meteorological data was provided by Deutscher Wetterdienst, Muenchen. The continued development of GLUE has been supported by NERC long-term grant NER/L/S/2001/00658. K.B. was supported by Konung Carl XVI Gustafs Gästprofessor i Miljövetenskap at Uppsala University in 2006/7. S.B. is grateful to Mehran Hojati for the help with the fuzzy regression implementation. The cooperation of the authors was partly within the framework of the UNESCO FRIEND project. We are grateful to Alberto Montanari and three anonymous referees whose comments have greatly improved the paper.

References

Beven, K. J. (1979), On the generalized kinematic routing method, *Water Resour. Res.*, *15*(5), 1238–1242.

Beven, K. J. (1986a), Hillslope runoff processes and flood frequency characteristics, in *Hillslope Processes*, edited by A. D. Abrahams, pp. 187–202, Allen and Unwin, Boston, Mass.

Beven, K. J. (1986b), Runoff production and flood frequency in catchments of order n: An alternative approach, in *Scale Problems in Hydrology*, edited by V. K. Gupta, I. Rodriguez-Iturbe, and E. F. Wood, pp. 107–131, Reidel, Dordrecht, Netherlands.

Beven, K. J. (1987), Towards the use of catchment geomorphology in flood frequency predictions, *Earth Surf. Processes Landforms*, *12*(1), 69–82, doi:10.1002/esp.3290120109.

Beven, K. J. (2001), *Rainfall-Runoff Modelling: The Primer*, John Wiley, Chichester, U. K.

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

Beven, K. J. (2007), Working towards integrated environmental models of everywhere: Uncertainty, data, and modelling as a learning process, *Hydrol. Earth Syst. Sci.*, *11*(1), 460–467.

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

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*, 279–298, doi:10.1002/hyp.3360060305.

Beven, K. J., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems, *J. Hydrol.*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

Beven, K. J., and M. J. Kirkby (1979), A physically-based variable contributing area model of basin hydrology, *Hydrol. Sci. Bull.*, *24*(1), 43–69.

Blazkova, S., and K. J. Beven (1995), Frequency version of TOPMODEL as a tool for assessing the impact of climate variability on flow sources and flood peaks, *J. Hydrol. Hydromech.*, *43*, 392–411.

Blazkova, S., and K. J. Beven (1997), Flood frequency prediction for data limited catchments in the Czech Republic using a stochastic rainfall model and TOPMODEL, *J. Hydrol.*, *195*, 256–278, doi:10.1016/S0022-1694(96)03238-6.

Blazkova, S., and K. J. Beven (2002), Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty), *Water Resour. Res.*, *38*(8), 1139, doi:10.1029/2001WR000500.

Blazkova, S., and K. J. Beven (2004), Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic, *J. Hydrol.*, *292*, 153–172, doi:10.1016/j.jhydrol.2003.12.025.

Blazkova, S., K. Beven, P. Tacheci, and A. Kulasova (2002), Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): The death of TOPMODEL?, *Water Resour. Res.*, *38*(11), 1257, doi:10.1029/2001WR000912.

Calver, A., R. Lamb, and S. E. Morris (1999), River flood frequency estimation using continuous runoff modelling, *Proc. Inst. Civ. Eng. Water Mar. Eng.*, *136*, 225–234.

Cameron, D. (2006), An application of the UKCIP02 climate change scenarios to flood estimation by continuous simulation for a gagued catchment in the northeast of Scotland, UK (with uncertainty), *J. Hydrol.*, *328*, 212–226, doi:10.1016/j.jhydrol.2005.12.024.

Cameron, D., K. J. Beven, J. Tawn, S. Blazkova, and P. Naden (1999), Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty), *J. Hydrol.*, *219*, 169–187, doi:10.1016/S0022-1694(99)00057-8.

Cameron, D., K. Beven, and P. Naden (2000a), Flood frequency estimation under climate change (with uncertainty), *Hydrol. Earth Syst. Sci.*, *4*(3), 393–405.

Cameron, D., K. J. Beven, and J. Tawn (2000b), Modelling extreme rainfalls using a modified random pulse Bartlett-Lewis stochastic rainfall model (with uncertainty), *Adv. Water Resour.*, *24*, 203–211, doi:10.1016/S0309-1708(00)00042-7.

Eagleson, P. S. (1972), Dynamics of flood frequency, *Water Resour. Res.*, *8*, 878–898, doi:10.1029/WR008i004p00878.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Towards improved calibration of hydrologic models: Multiple and incommensurable measures of information, *Water Resour. Res.*, *34*, 751–763, doi:10.1029/97WR03495.

Hydrometeorological Institute (1967), *Katalog povětrnostních situací pro území ČSSR*, 94 pp., Prague.

Hojati, M., C. R. Bector, and K. Smimou (2005), A simple method for computation of fuzzy linear regression, *Eur. J. Oper. Res.*, *166*, 172–184, doi:10.1016/j.ejor.2004.01.039.

Hosking, J. R. M. (1997), Fortran routines for use with the method of L-moments, version 3.02, *IBM Res. Rep. RC 20525(90933)*, IBM Res. Div., Almaden, N. Y.

Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments*, 224 pp., Cambridge Univ. Press, Cambridge, U. K.

James, L. D. (1965), Using a digital computer to estimate the effects of urban development on flood peaks, *Water Resour. Res.*, *1*(2), 223–233, doi:10.1029/WR001i002p00223.

Jones, D. A., and A. L. Kay (2007), Uncertainty analysis for estimating flood frequencies for ungauged catchments using rainfall-runoff models, *Adv. Water Resour.*, *30*, 1190–1204, doi:10.1016/j.advwatres.2006.10.009.

Kakos, V. (1977), Meteorologické příčiny povodní v oblasti Krušných hor, *Vodohospod. Technickoekon. Inf.*, *19*(9), 321–327.

Kakos, V. (1978), Hydrometeorologická charakteristika povodní na území ČSR, *Vodohospod. Technickoekon. Inf.*, *20*(4), 127–131.

Kay, A. L., R. G. Jones, and N. S. Reynard (2006), RCM rainfall for UK flood frequency estimation. Part II. Climate change results, *J. Hydrol.*, *318*, 163–172, doi:10.1016/j.jhydrol.2005.06.013.

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of mathematical models, *J. R. Stat. Soc. Ser. A*, *63*(3), 425–450.

Lamb, R. (1999), Calibration of a conceptual rainfall-runoff model for flood frequency estimation by continuous simulation, *Water Resour. Res.*, *35*(10), 3103–3114, doi:10.1029/1999WR900119.

Lamb, R., and K. J. Beven (1997), Using interactive recession curve analysis to specify a general catchment storage model, *Hydrol. Earth Syst. Sci.*, *1*(1), 101–113.

Lamb, R., and A. L. Kay (2004), Confidence intervals for a spatially generalized, continuous simulation flood frequency model for Great Britain, *Water Resour. Res.*, *40*, W07501, doi:10.1029/2003WR002428.

Novický, O., L. Kašpárek, and S. Kolářová (1993), Hydrological design data estimation techniques, *Rep. WCASP 26 TD 554*, World Meteorol. Organ., Geneva, Switzerland.

Rezacova, D., P. Pesice, and Z. Sokol (2005), An estimation of the probable maximum precipitation for river basins in the Czech Republic, *Atmos. Res.*, *77*, 407–421.

Tesař, M., V. Eliáš, and M. Šír (1995), Preliminary results of characterization of cloud and fog water in the mountains of southern and northern Bohemia, *J. Hydrol. Hydromech.*, *43*, 412–426.

K. Beven, Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ UK. (k.beven@lancaster.ac.uk)

S. Blazkova, T. G. Masaryk Water Research Institute, Podbabska 30, 160 00 Prague 6, Czech Republic. (sarka_blazkova@vuv.cz)