Ensemble predictions of runoff in ungauged catchments

Neil McIntyre, Hyosang Lee, and Howard Wheater
Department of Civil and Environmental Engineering, Imperial College London, London, UK

Andy Young
Centre for Ecology and Hydrology, Wallingford, UK

Thorsten Wagener
Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA

Received 31 May 2005; revised 9 September 2005; accepted 22 September 2005; published 29 December 2005.

[1] A new approach to regionalization of conceptual rainfall-runoff models is presented on the basis of ensemble modeling and model averaging. It is argued that in principle, this approach represents an improvement on the established procedure of regressing parameter values against numeric catchment descriptors. Using daily data from 127 catchments in the United Kingdom, alternative schemes for defining prior and posterior likelihoods of candidate models are tested in terms of accuracy of ungauged catchment predictions. A probability distributed model structure is used, and alternative parameter sets are identified using data from each of a number of gauged catchments. Using the models of the 10 gauged catchments most similar to the ungauged catchment provides generally the best results and performs significantly better than the regression method, especially for predicting low flows. The ensemble of candidate models provides an indication of uncertainty in ungauged catchment predictions, although this is not a robust estimate of possible flow ranges, and frequently fails to encompass flow peaks. Options for developing the new method to resolve these problems are discussed.

Citation: McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, *Water Resour. Res.*, *41*, W12434, doi:10.1029/2005WR004289.

1. Introduction

[2] The range of approaches to continuous time rainfall-runoff modeling is well known and the merits of alternative model types are well documented [e.g., *Jakeman and Hornberger*, 1993; *Wheater et al.*, 1993; *O'Connell and Todini*, 1996; *Wheater*, 2002]. The most common approach uses conceptual models, in which the hydrological storages and associated losses and routing processes are represented by a series of conceptual stores, which explicitly or implicitly represent one or more components of the real hydrological system (e.g., soil water and groundwater storages) and their interactions. A large number of conceptual models have been devised, to suit individual catchments and modelers' perceptions of important components [e.g., *Singh and Frevert*, 2002a, 2002b]. A key aspect of the conceptual model type is that the model parameters (MPs) do not have physically based associations with measurable catchment descriptors (CDs), although inherent correlation between MPs and CDs is often assumed [*Wagener et al.*, 2004]. Hence, prior to conditioning on observations, there is a range of parameter sets that might be used to model the response of a catchment. The uncertainty becomes larger when it is recognized that a range of conceptual model structures might be equally considered for any application [e.g., *Neuman*, 2003].

[3] If observations of relevant system inputs and outputs (e.g., precipitation, potential evapotranspiration and streamflow) are available, then the MP space is generally constrained by calibration. The model structure may then be adjusted if supported by analysis of model residuals or parameter uncertainty [e.g., *Lee et al.*, 2004]. However, unknown errors in input and output data may make it impossible to distinguish the performance of one model structure from another, and one MP set from another [*Beven and Binley*, 1992; *Andreassian et al.*, 2001; *Perrin et al.*, 2001; *Lee et al.*, 2005a]. The problem is compounded by model overparameterization, so that a multitude of significantly different MP sets lead to near-equivalent results (the equifinality problem [*Beven and Freer*, 2001]). Much research has been directed toward these issues, through development of improved model identification procedures and supporting tools for model optimization and error analysis (see reviews of *Wheater* [2002] and *Gupta et al.* [2005]). Despite considerable progress in this area, one major outcome is the recognition that model uncertainty is inherent and unavoidable, and therefore best practice is to employ an ensemble of feasible models to give a corresponding ensemble of predictions [*Beven and Freer*, 2001; *Wheater*, 2002].

[4] In the common case that the catchment of interest is ungauged or poorly gauged (e.g., long periods of data are

0043-1397/05/2005WR004289$09.00

missing, or large gauging errors exist), the conceptual model may be identified through a process of regionalization. The regionalization problem is relatively long standing, with notable early studies by *Nash* [1960] and, for continuous time models, by *Manley* [1978]. Recent improvements in the availability of spatial data, together with improved computational resources and new interest in continuous simulation, have led to new research activity in this field [*Sivapalan*, 2003]. One topical research area is how to best estimate and integrate all sources of uncertainty into the regional model and the subsequent ungauged catchment predictions. In this paper, we argue that established methods of regionalization may not be well suited to meet this challenge, and propose an alternative method founded on ensemble modeling and weighted averaging. Using a case study of 127 UK catchments, comparisons between this new method and more established methods are reported.

2. Regionalization Using Regression

[5] Regression of optimized MPs against observed (or modeled) numeric catchment descriptors (CDs) is the most commonly used approach to conceptual rainfall-runoff model regionalization. The established regression procedure may be summarized as: selection of a suitable pool of well-gauged catchments with known (or estimated) CDs, selection of a model structure deemed suitable for the relevant catchments, identification of the optimal MP set for each of the gauged "donor" catchments, multiple regression of these MP estimates against the CDs (where MPs are usually but not always treated as independent of each other), and use of the regression equation to estimate the MP values for the ungauged "target" catchment. A series of experiments have been conducted over the past 15 years by various investigators, with the aims of identifying relationships between conceptual MPs and CDs, and testing the predictive ability of the estimated ungauged catchment models (see the reviews and results of *Seibert* [1999], *Young* [2000], *Kokkonen et al.* [2003], *Wagener et al.* [2004], and *Merz and Blöschl* [2004]).

[6] There are a number of reasons which have limited the applicability of regression in this context. The significance of MPs varies from one catchment to the next, depending on how they are able to compensate for local data and model structure errors [e.g., *Andreassian et al.*, 2001]. This means that relationships between CDs and MPs may be weak or absent, especially when a wide range of catchment types has been included. For example, using an 11-parameter model of daily flows applied to 308 Austrian catchments, *Merz and Blöschl* [2004] found that for five MPs the strength of regression relationships with CDs was consistently less than or equal to $R^2 = 0.1$. *Seibert* [1999] found that only 6 of 13 MPs were related to CDs, for a conceptual model of monthly water balance applied to 11 catchments in Sweden. Using a relatively parsimonious, five-parameter model on a sample of nine UK catchments, *Wagener et al.* [2004] achieved relatively strong CD-MP regression relationships, with R^2 values of up to 0.94, for selected parameters. Using a wider range of 131 UK catchments and testing alternative MP data sets, *Lee et al.* [2005b] found that only the model's fast residence time had a strong relationship (i.e., with $R^2 > 0.4$) with the CDs, and that the slow residence time was statistically independent of CDs in more permeable catchment types.

[7] Some investigators have attempted to propagate the uncertainty in the regional relationships to ungauged catchment predictions. *Yeh et al.* [1997] applied Rosenblueth's and Harr's methods to propagate MP uncertainty through a unit hydrograph. In the continuous simulation context, *McIntyre et al.* [2004], *Lamb and Kay* [2004] and *Wagener and Wheater* [2005] propagated the standard errors in the regression model to uncertainty using Monte Carlo simulation. *Wagener and Wheater* [2005] used weighted regression to propagate the calibrated MP uncertainty into the regression equations. *Merz and Blöschl* [2004] obtained two estimates of the optimum MP sets by splitting the calibration period, and tested the sensitivity of the regression to this uncertainty.

[8] Implicit to almost all this research is the assumption that errors in MP data are normally distributed, and that the MPs are independent of each other, so that standard multiple univariate regression techniques may be reasonably applied. However, it is well known that conceptual MPs tend to interact with each other during calibration and produce complex response surfaces that are not well described by either independent or correlated normal distributions [*Sorooshian and Gupta*, 1995]. Although multivariate regression methods were employed with some success by *Tung et al.* [1997] for a two-parameter event-based model, these procedures would become complex if applied to a model with several parameters. *Kokkonen et al.* [2003] tested a simpler approach, in which a strong correlation between two MPs was used to estimate the value of one of them, having first regionalized the other, and found this to work better than univariate regression. Canonical correlation analysis [e.g., *Young*, 2000] uses linear combinations of the each catchment's CDs as the independent variables and linear combinations of the corresponding MP sets as the dependent variables, thus allowing for the linear interdependencies within and between both. However, both multivariate regression and canonical correlation analysis only account for the linear elements of the MP interdependency. *Lamb and Kay* [2004] handled the MP interdependencies using sequential regression, where univariate regression is applied to the optimum estimates of the most influential MP, its values are fixed for all the donor catchments using the regressed estimates, then all the other MPs are recalibrated. This continues sequentially through all the MPs, removing the issue of interdependency by making MP estimates conditional on regionalized values of more influential MPs. *Wagener et al.* [2004] found this approach difficult to apply without introducing bias into the regionalized parameters.

[9] The structural form of the regression equation is a further issue. Various transforms of the CDs and MPs have been made to introduce nonlinearities; however these transforms are either based on previous reports or on a limited amount of trial and error. Subjective decisions may also be used to decide which CDs to include as independent variables, for example based on prior expectations of important CDs and/or to avoid inclusion of strongly intercorrelated CDs [e.g., *Wagener et al.*, 2004; *Merz and Blöschl*, 2005]. More objective alternatives to selecting

CDs are principal component analysis [e.g., *Sefton and Howarth*, 1998] or stepwise regression [e.g., *Wagener et al.*, 2004; *Lee et al.*, 2005b]. Another issue, which is intuitively important, but which apparently has received no investigation in this context, is the errors in the CD estimates. MPs estimated via a regression model are conditional on the accuracy of the CDs. While we might assume that CDs based on topography and easily surveyed surface features have little error, CDs which describe the subsurface are generally themselves the output of a model linking them to land cover or to point observations, for example of soil properties (e.g., the HOST classification of *Boorman et al.* [1995]).

[10] There has been mixed success in making predictions of flow in ungauged catchments, using regression approaches. For example, *Kokkonen et al.* [2003] applied a regression scheme to the six-parameter IHACRES model on thirteen Australian catchments, and reported an average loss in performance (Nash-Sutcliffe efficiency (NSE)) of only 0.06 when using a regionalized MP set instead of a locally optimized MP set. They also note that using a single, similar gauged catchment may be preferable if such a catchment exists. *Lee et al.* [2005b] applied a five-parameter probability distributed model to 66 low-permeability non-urban UK catchments, and reported a loss in NSE of less than 0.04 in 90% of catchments when using regressed MP values instead of a locally optimized MP set. However loss in performance was generally larger for a low-flow objective function and for more permeable catchments. *Merz and Blöschl* [2004] found that the average loss in NSE performance moving from a locally calibrated model to a regionalized model was 0.1, twice the loss associated with moving from the calibration to validation period, applying an 11-parameter model to 308 Austrian catchments. They also found that methods based on spatial proximity (kriging and using MPs from nearby gauged catchments) were generally able to produce more reliable predictions in these catchments. This result is, however, in contrast to the results of *Vandewiele and Elias* [1995], who found geographical proximity of little use in predicting ungauged flows. *Calver et al.* [1999] regionalized a probability distributed model using hourly data from 36 UK catchments, to estimate floods over a range of return periods. Moving from locally calibrated model parameters to those derived using univariate regression caused the average error in flood quantiles to increase from 9 to 39%, although this reduced to 24% when sequential regression was used.

[11] It may easily be argued that regression has proven to be a useful tool for making predictions of runoff in ungauged catchments. However, the fundamental limitations of regression analysis mentioned above, especially the need to neglect or greatly simplify the interdependencies between MPs and the neglect of errors in the CD values, leads to the view that further effort at refining the application of regression techniques may not be the optimum way forward.

3. Ensemble Modeling and Weighted Averaging

[12] The starting principle for the ensemble modeling and our similarity weighted averaging (SWA) method is that the CD-MP relationship should be treated as a response surface of likelihood. This response surface might be integrated across a representative range of CDs (i.e., representative of the uncertainty in the properties of the target ungauged catchment) to produce a continuous joint distribution function of MPs, which would then be applied to the relevant rainfall time series to generate a distribution function of runoff time series. In practice, defining a continuous CD-MP distribution function would be problematic due mainly to the issues of the complex parameter interdependencies and response surface discontinuities [*Duan et al.*, 1992]. Instead, we would hope to have a large enough number of CD-MP samples within the relevant CD range so that they could all be applied individually to the ungauged catchment rainfall, providing an ensemble time series of flow forecasts and allowing a weighted average best estimate of flow to be calculated.

[13] Developing the idea further, each donor gauged catchment can be described by a number of candidate models with corresponding prior likelihoods (i.e., prior to considering the properties of the target ungauged catchment). This allows prior uncertainty in calibrated models, for example that associated with equifinality, to be represented. The prior likelihoods may reflect knowledge about model performance on the gauged catchment data so that the influence of models that have performed relatively badly (due to model structural error, data errors or suboptimal approximations of MPs) will be weighted down. These prior likelihoods may then be conditioned using a measure of the similarity (of the donor catchment with the target catchment) into a posterior likelihood, and all models with nonzero posterior likelihood would be applied. The main theoretical advantage of this method over regression is that the MP sets are neither averaged or interpolated; instead, the full information content of the locally calibrated MP sets is retained at all stages.

[14] Ensemble modeling and using multiple weighted models are central to the generalized likelihood uncertainty estimation (GLUE) framework of *Beven and Binley* [1992], which has been widely applied to estimating uncertainty in gauged catchment predictions [*Beven and Freer*, 2001]. GLUE involves updating model likelihoods by relating likelihood to a fit statistic. The approach outlined above can be viewed as an extension of GLUE to the ungauged catchment problem, where candidate models and associated likelihoods are drawn from all gauged catchments, and are subject to a further stage of conditioning based on similarity to the target ungauged catchment. Unless the fit statistic used within GLUE is a formal likelihood function then the derived likelihoods are not frequential probabilities, but are subjective weights used to distinguish between the relative believability of models. This subjective element is a major criticism of GLUE; however because the subjectivities are expressed mathematically (as objective functions) they can be audited and revised, and alternatives can be compared and integrated together. The same can be said in defense of the subjectivities which will inevitably arise in selecting catchment similarity measures. Potentially, the variables used to define similarity may be optimized against predictive performance within a particular set of gauged catchments. Another anticipated problem is the outcome of a posterior likelihood equal to zero for all models. If

Table 1. Prior Ranges of Model Parameters

Parameter	Description	Range
c_{max}	maximum storage capacity (mm)	0–500
b	shape of Pareto distribution	0–2
T_q	time constant for quick flow reservoir (days)	0–50
T_s	time constant for slow flow reservoir (days)	50–500
f	fraction of flow through quick flow reservoir	0–1

this were the case, it would be an indication of a paucity of gauged catchments to support the approach and that another approach (e.g., regression, or a more physically based approach) might be better in that case. The proposed regionalization method also has the potential to generate prior distributions of MP sets for input to a conventional application of GLUE to a gauged catchment.

[15] Although ensemble modeling and weighted averaging have been widely used to represent MP uncertainty within the GLUE framework, and also for propagating climate change uncertainty to hydrological forecasts [e.g., *Anderson et al.*, 2001; *Wetherald and Manabe*, 2002; *Christensen et al.*, 2004], representation of alternative rainfall-runoff model structures is rare. *Neuman* [2003] argues that integrating the results of a number of candidate model structures will reduce the prediction bias; however a potential problem is the difficulty in assigning prior likelihoods, particularly when the candidate structures are structurally similar. In practice, choice of model structures may be limited to those easily available to the modelers, and which use a common set of available input data, and this will limit the degree to which structural bias can be averaged out. In the context of lumped conceptual rainfall-runoff modeling, this is seen in the model structure intercomparisons of *Perrin et al.* [2001] and *Lee et al.* [2005a], where optimized results of candidate model structures were often indistinguishable. *Shamseldin et al.* [1997] also noted the limitation of using a restricted range of model structures, when averaging storm runoff predictions from five rainfall-runoff model structures. *Georgakakos et al.* [2004] used a wide range of lumped and distributed model structures, for constructing flow ensembles in six catchments in central USA, focusing on reliability of flood forecasts. They found that generally more reliable results were achieved when using the ensembles rather than single models, with or without prior model calibration. In cases, model calibration reduced the ability of the ensemble to encompass flood peaks.

[16] Previous application of SWA to prediction in ungauged catchments is very limited. The statistical approach to flood estimation used by *Institute of Hydrology* [1999] is comparable. For a target catchment, similar well-gauged donor catchments are identified and their annual maxima flood (AMF) data are integrated, to provide an estimate of the AMF statistics in the target catchment. This method has also been applied to estimation of low-flow statistics [*Holmes et al.*, 2002]. *Young* [2000] extended the method to continuous time modeling. He calibrated a model on data from donor catchments within a threshold of similarity and applied each donor MP set to generate a realization of flow in the ungauged catchment. The flow time series were averaged (without weighting) over all realizations. However, this was found to perform generally worse than using a regression-based method.

4. Model and Data

[17] While the proposed application of SWA to predicting runoff in ungauged catchments clearly has the potential to integrate the results of more than one model structure, we limit the current investigation to a single conceptual model structure and focus on representing effects of MP uncertainty. The chosen model structure is a version of the probability distribution model (PDM) of *Moore* [1985], which represents the spatial variability of the catchment's storage capacity by a Pareto distribution. The soil moisture accounting model has two parameters. c_{max} is the maximum storage capacity in the catchment, and b is a Pareto distribution parameter which controls the spatial variability of c (for example $b = 1$ gives a very variable degree of saturation, while b close to 0 gives a relatively little spatial variation). Runoff $Q(t)$ is equal to the saturation excess integrated over the area of the catchment. The evapotranspiration rate is assumed equal to the potential evapotranspiration multiplied by the relative moisture state of the model. The routing component of the models is formed of two linear reservoirs in parallel, one representing a relatively quick catchment response (residence time T_q) and the other for a slower response (T_s). All the runoff is split between these two reservoirs, defined by parameter f (proportion of the total effective rainfall going to the quick response reservoir). The two components of outflow are aggregated into total streamflow. This five-parameter model structure has previously been shown to be at least as good as alternative lumped models of similar complexity over a wide range of UK catchment types, when using daily data [*Lee et al.*, 2005a]. The ranges in Table 1 are used to define plausible precalibration values of the five MPs.

[18] The data set is composed of daily precipitation, streamflow and potential evaporation, and a set of 17 CDs for 127 UK catchments. The CDs which we refer to in this paper are listed in Table 2. The 127 catchments are hydrologically quite varied, ranging from wet catchments with high annual rainfall (e.g., 2860 mm/year) in Scotland to dry catchments with small annual rainfall (e.g., 602 mm/year) in southeast England. Catchment average elevations range from 557.2 m to 37.1 m (above UK ordnance datum). The area of the catchments ranges from 1 km^2 to 1700 km^2 with a mean of 248 km^2. All catchments may be considered nonurbanized (URB_EXT < 0.0125). Full details of the data

Table 2. A Sample of the Available Catchment Characteristics[a]

Abbreviations	Definitions
BFIHOST	base flow index derived using the HOST classification [0–1]
AREA	catchment drainage area (km^2)
SAAR	1941–1970 standard- period average annual rainfall (mm)
ALTBAR	mean catchment altitude (m above sea level)
URB_EXT	index of fractional urban extent in 1990
SPRHOST	SPR (standard percentage runoff) derived using the host classification

[a]Definitions from *Institute of Hydrology* [1999].

sources and quality control are given by *Young* [2000]. The period 1 October 1986 to 30 September 1996 is used for model calibration except in 12 catchments where there was more than 30 days of missing flow data in that period; in these cases another continuous 10-year period was selected. Independent 5-year validation periods were used to test the ability of the models to make local predictions, except in six catchments where the period was less than 5 years. When measuring model performance in the calibration and validation periods, the first 20% of the data time series was neglected, to reduce sensitivity to initial conditions.

5. Method Description

[19] The general equation describing the averaging of simulations of a runoff time series $Q(t)$ at an ungauged catchment is

$$\overline{Q}(t) = \sum_{j=1}^{S} \sum_{i=1}^{N} Q_{i,j}(t) \times W_{i,j} \tag{1}$$

where $W_{i,j}$ is the posterior likelihood given to the ith candidate model originating from the jth gauged donor catchment, normalized so that all W sum to unity; N is the number of candidate models from each gauged donor catchment (here constant for all j); S is the number of donor catchments; and $Q_{i,j}(t)$ is the runoff obtained by applying the respective model to the ungauged catchment's rainfall and other time series inputs. In the present application, W is defined by the product of a prior likelihood of a model (P) and the relative likelihood of that model being applicable to the target gauged catchment (B). Various experiments are done here to assess alternative definitions of P and B in terms of the accuracy of the predicted $Q(t)$.

5.1. Prior Likelihoods

[20] The prior likelihoods P aim to represent the relative applicability of the candidate model prior to considering the nature (i.e., the CDs) of the target ungauged catchment. An objective approach is to define P by the success with which the model has previously simulated runoff at gauged catchments, based on a fit statistic. Then, P is linked to the quality of the model, and the quality of the data set upon which the model has been tested.

[21] This was applied using the Nash-Sutcliffe efficiency (NSE) as the fit statistic. For each catchment, 10,000 sets of MPs were randomly sampled from within the uniform distributions defined by the ranges in Table 1, and using each sample the PDM model was run giving 10,000 realizations of Q. Corresponding NSE values were calculated in the calibration periods. A fraction of the 10,000 sampled models for each gauged catchment were then considered as good enough, based on the NSE values, to take forward as "prior" candidate models. This fraction might be decided by using a threshold in the NSE value, as is the normal GLUE procedure. Here, however, the N parameter sets with the best (highest) NSE values for each catchment were retained, and we investigated the performance for different values of N. The NSE values were rescaled to P values as defined below.

$$P_{i,j} = \frac{\left(\mathrm{NSE}_{i,j} - \mathrm{NSE}_{\min,j}\right)/\left(1 - \mathrm{NSE}_{\min,j}\right)}{\sum_{i=1}^{N} \left(\mathrm{NSE}_{i,j} - \mathrm{NSE}_{\min,j}\right)\left(1 - \mathrm{NSE}_{\min,j}\right)} \tag{2}$$

where subscript i represents the ith of the N parameter sets, subscript j indexes the donor catchment, and $\mathrm{NSE}_{\min,j}$ is the lowest of the N corresponding Nash-Sutcliffe efficiency values from that gauged catchment. Rather than defining P as directly proportional to NSE, this formula accentuates the difference in performance between the best and worst of the N parameter sets.

[22] The same procedure was repeated but using a low-flow fit statistic, FSBM, as it is known that NSE is less suitable for low-flow applications [*Legates and McCabe*, 1999].

$$\mathrm{FSBM} = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n} (Qo_t - Q_t)^2}}{\frac{1}{n}\sum_{t=1}^{n} Qo_t} \tag{3}$$

where here subscript t represents the time step, Q and Qo are the calculated and observed flows for all t for which Qo is less than 75% of average Qo, during hydrograph recession periods (i.e., reducing flow), and n is the total number of time steps within such periods. FSBM values are then rescaled into P values in a similar manner as applied to NSE. The NSE and FSBM fit statistics are used here to represent all around model performance and to illustrate the possible significance of the performance measure. Alternative fit statistics, or other measures of operational performance of the model may be appropriate for specific applications of the method [*Wagener and McIntyre*, 2005].

5.2. Posterior Likelihoods

[23] The likelihood of a model of a donor catchment being applicable to a target catchment is defined by the similarity of the two catchments. The chosen measure of catchment similarity is that used by *Institute of Hydrology* [1999] which considers catchment area (*AREA*), standardized annual average rainfall (*SAAR*) and estimated base flow index (*BFIHOST*). These three CDs are considered to be the primary factors categorizing nonurban UK catchment types. Other potentially important CDs are well correlated with one of these [*Young*, 2000; *Wagener et al.*, 2004] hence need to be omitted from a similarity measure. For a target catchment with CD values $AREA'$, $SAAR'$ and $BFIHOST'$;

$$E_j = \sqrt{\frac{1}{2}\left(\frac{\ln AREA_j - \ln AREA'}{\sigma(\ln \mathbf{AREA})}\right)^2 + \left(\frac{\ln SAAR_j - \ln SAAR'}{\sigma(\ln \mathbf{SAAR})}\right)^2 + \left(\frac{BFIHOST_j - BFIHOST'}{\sigma(\mathbf{BFIHOST})}\right)^2} \tag{4}$$

where E_j is the dissimilarity of the jth candidate donor catchment to the target catchment; $AREA_j$, $SAAR_j$ and $BFIHOST_j$ are the donor catchment CDs; and **AREA**,

SAAR and **BFIHOST** are the sets of CD values from the donor catchments, and σ indicates standard deviation within each set. The natural logarithms of *AREA* and *SAAR* are used to avoid highly skewed distributions, i.e., to moderate the dissimilarity of the largest and wettest catchments. Donor catchments with E above a specified value ($=E_T$) are given zero posterior likelihood, $B = 0$. The other E values are rescaled into B values as defined below.

$$B_j = \frac{\left(1 - E_j/E_{\max}\right)}{\sum_{j=1}^{S}\left(1 - E_j/E_{\max}\right)} \quad \text{for } E_j \le E_T \tag{5}$$

where E is the measure of catchment dissimilarity, $E_{\max}$ is the maximum value of $E \le E_T$ for the target catchment, and S is the number of catchments for which $E \le E_T$. The posterior likelihood of the ith prior model from the jth donor catchment is then,

$$W_{i,j} = \frac{P_{i,j}B_j}{\sum_{j=1}^{S}\sum_{i=1}^{N}P_{i,j}B_j} \tag{6}$$

Equation (6) provides the W values for use in equation (1). For comparison with the similarity measure, the spatial proximity of catchments was also used to define posterior likelihood,

$$D_j = \sqrt{\left(NORTH' - NORTH_j\right)^2 + \left(EAST' - EAST_j\right)^2} \tag{7}$$

where $NORTH_j$, and $EAST_j$ are the national grid coordinates of the outlet of the jth gauged catchment, and $NORTH'$, and $EAST'$ are the same for the target gauged catchment. D then replaces E in equation (5).

5.3. Assessment

[24] Each of the 127 catchments in turn was considered to be ungauged, with the N best MP sets from the remaining 126 catchments used to define the prior candidate donor models. For each catchment's calibration period, this provided a locally calibrated model result (i.e., the best fit out of the 10,000 samples) and a regionalized result (i.e., using SWA). A comparison of the two fits using the NSE and FSBM values in each of the catchments provided a general assessment of the SWA scheme. As a truer test of the relative predictive power of the locally calibrated and regionalized models, the fits in the validation periods were also compared. Further to this assessment, the general ability of the ensemble of regionalized results to represent prediction uncertainty was assessed in each catchment; the percentage of observed data points outlying the ensemble and associated 90% confidence intervals is reported. Results are reported separately for more permeable catchments (BFIHOST > 0.5) and less permeable catchments (BFIHOST ≤ 0.5) so that the success of the regionalization in these different types of catchments can be reviewed. This categorization according to BFIHOST follows the analysis of *Lee et al.* [2005b], who found that catchments above the 0.5 threshold were generally more difficult to model using the regression approach to regionalization. The assessments based on NSE and FSBM are supported by illustrating the modeled and observed time series data for selected catchments.

6. Results

[25] Figure 1 plots the NSE values (represented as 1-NSE for minimization) achieved using locally calibrated MP sets, against the values achieved using the regionalized (SWA) model. Figures 1a–1d are all based on using $S = 1$ (i.e., only the most similar catchment based on equation (4) is used as a donor), and $N = 10$ (i.e., the 10 MP sets with the best NSE values from the donor catchment). Figures 1a and 1b show results in the 10-year calibration period for the less and more permeable catchments respectively, while Figures 1c and 1d show the same in the shorter validation period. Figures 1e–1h are the same format, but with $S = 10$. Figures 1i–1l are the same format as Figure 1a–1d, but the single nearest catchment (minimum D as defined by equation (7)) is used instead of the most similar. Figures 1m–1p show the same but where results are the weighted average of the results of all models with $D <$ 30 km. On the basis of this latter criterion, Figure 2a is a histogram showing the number of ungauged catchments associated with different values of S for the less permeable catchments; Figure 2b is the same for the more permeable catchments. The 22 ungauged catchments with zero donor catchments have been omitted from Figures 1m–1p. In all plots in Figure 1 the line of equivalent performance is marked as a solid 45° line, while the dotted line is the 90 percentile, above which only 10% of results lie. Therefore the closeness of this dotted diagonal to the solid diagonal indicates the general success of using regionalization relative to using local calibration. Various arbitrary decisions have been made to define values of N, S, and E_T in this analysis; sensitivity to these will be examined later.

[26] Using $N = 10$ and $S = 10$, with similarity measure E, gives generally more accurate predictions in terms of NSE than the other weighting schemes. For example, Figures 1e and 1f show 90 percentile losses in performance of 0.06 and 0.21 respectively, compared, for example, to 0.10 and 0.36 for Figures 1a and 1b (where only the most similar catchment is used as a donor rather than $S = 10$), and 0.12 and 0.23 for Figures 1m and 1n (where a number of nearby catchments are used). Using the single nearest catchment gives a significantly worse performance overall, reflecting UK geology which often changes markedly between neighboring catchments. The especially poor performance of the regionalized models shown in Figures 1j and 1l implies that the NSE measure is especially sensitive to using an inappropriate MP set in more permeable catchments, and/or that the similarity measure is more likely to be inappropriate for more permeable catchments. Comparing, for example, Figures 1e and 1f illustrates the general difficulty in predicting flow in more permeable ungauged catchments; while their locally calibrated NSE values are generally better than in the less permeable catchments, the values achieved using the best regionalization scheme are generally worse. The validation period results show that in a significant proportion of the catchments, the regionalized models achieve more accurate predictions than the locally calibrated models (e.g., in Figures 1g and 1h, 33 and 26% of the data are below the solid diagonal). The regionalized NSE values in the validation periods are, overall, slightly

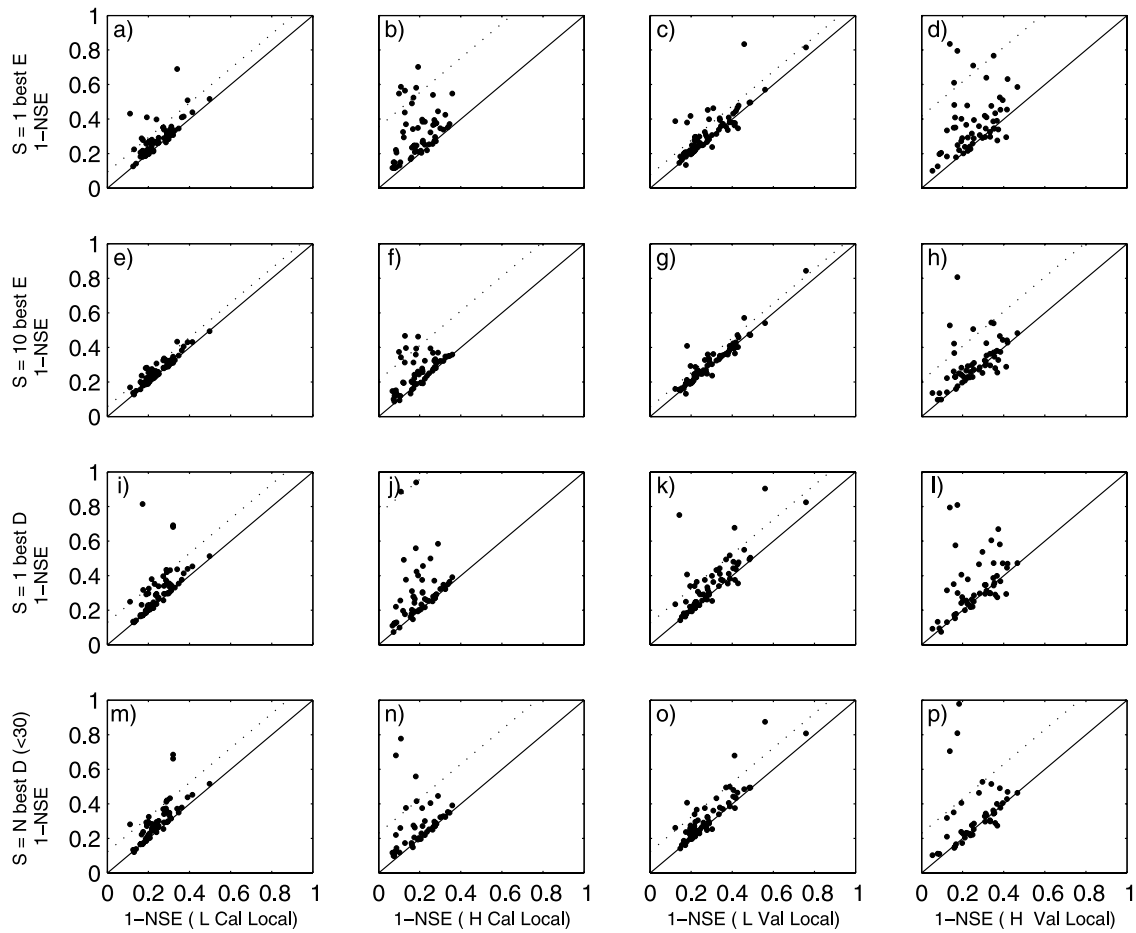

Figure 1. (a–p) Performance (1-NSE) achieved using a locally calibrated model (x axes) plotted against performance achieved using the regionalized model (y axes, showing four alternative SWA schemes that are described in text), in calibration (Cal) and validation (Val) periods, for lower-permeability (L) and higher-permeability (H) catchments. Note that all axes are curtailed at (1-NSE) = 1. Outliers not plotted have (x,y) coordinates of [0.27, 1.5] (Figure 1b); [15.2, 11.2], [0.29, 1.5], [0.23, 2.1] (Figure 1d); [0.27, 1.7] (Figure 1f); [15.2, 15.3], [0.29, 1.1], [0.23, 2.5] (Figure 1h); [0.13, 1.5], [0.27, 9.1], [0.08, 1.1], [0.26, 1.8], [0.10, 7.5] (Figure 1i); [15.2, 22.2], [0.23, 13.8], [0.18, 1.5], [0.16, 6.7] (Figure 1l); [0.27, 1.7] (Figure 1n); and [15.2, 22.2], [0.32, 1.9] (Figure 1p). In the catchment with an x coordinate of 15.2, only 1 year was available for the validation period, and the poor performance is due to errors in initial conditions.

worse than the equivalent values in the calibration periods. This loss in performance may be due to the reduced length of period used for the validation which causes the result to be slightly more sensitive to errors in the model's initial conditions. There was limited evidence to suggest that higher similarity between the target and donor catchments led to improved ungauged catchment predictions; the correlation between average E (i.e., the measure of dissimilarity averaged over the 10 donor catchments) and NSE was −0.13.

[27] Various other values of S and N were tested, as were alternative measures of catchment similarity. Using only one MP set (N = 1) rather than 10 did not make a large difference; for example, the 90% loss in performances in Figures 1e–1h dropped from (0.06, 0.21, 0.05, 0.20) to (0.05, 0.22, 0.05, 0.26). Raising N to 100 (1% of the total number of samples) showed a similar lack of sensitivity. This implies that averaging the results of the best 1% of the sampled MP sets is equivalent to using the single best set, i.e., MP equifinality in local calibrations is not relevant to the single best flow prediction, insofar as can be measured by NSE and FSBM. However, we did not test sensitivity to values of $N > 100$. The sensitivity to the definition of S was tested; rather than fixing S = 10, S was defined by the number of catchments within a threshold value of E (E_T in equation (5)). Figures 2c and 2d are histograms showing the number of target catchments associated with different values of S using $E_T = 0.5$; Figures 2e and 2f show the same for E_T = 1. Both these schemes resulted in generally poorer ungauged catchment predictions than fixing S = 10. The similarity measure was also changed to one defined by *SPRHOST* and *ALTBAR* (see Table 2 for definitions), which were identified to be the two most consistently important CDs in the UK by *Lee et al.* [2005b] as covariants of *SAAR*, *BFIHOST* and *AREA*. Results were comparable with, or slightly worse than, those in Figures 1a–1h; for example, the 90% loss in performances in Figures 1e–1h changed from (0.06, 0.21, 0.05, 0.20) to (0.08, 0.38, 0.08, 0.39).

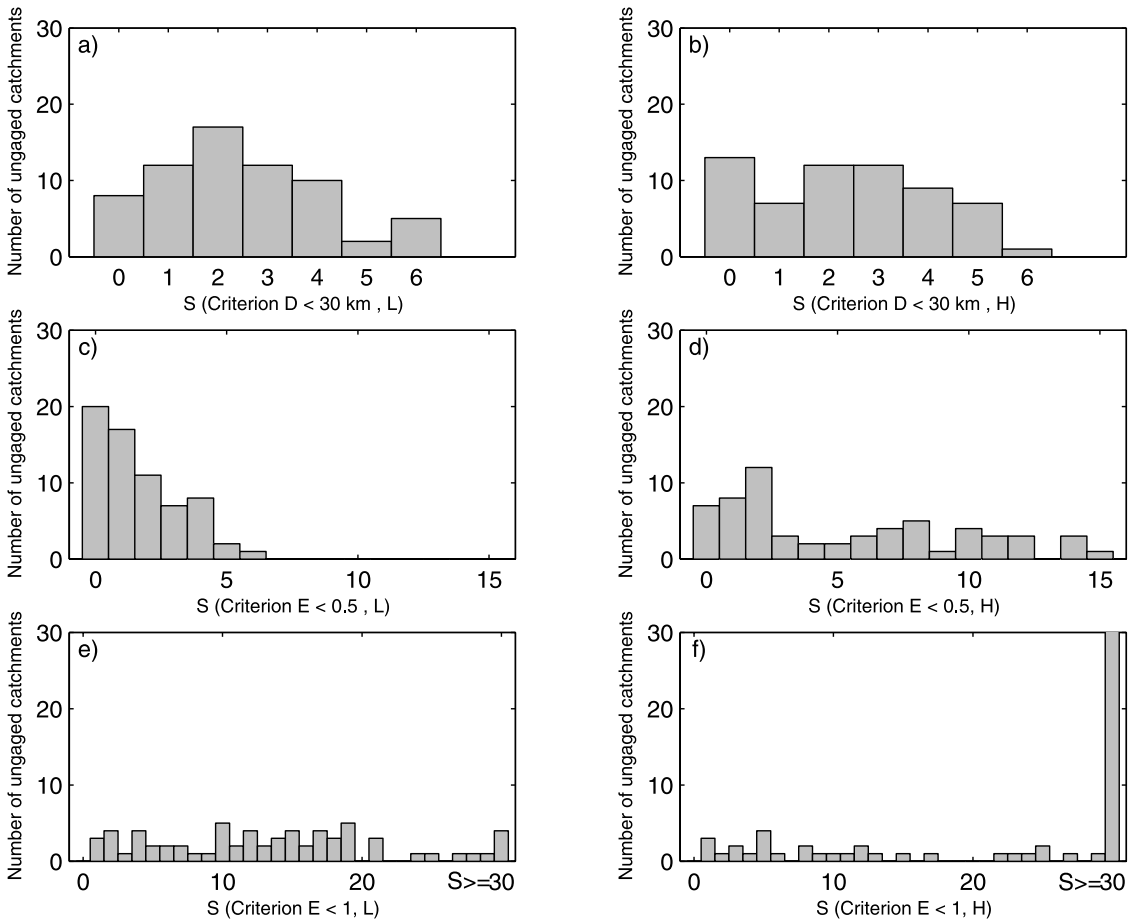

Figure 2. Histograms showing the number of ungauged catchments having different numbers of donor catchments (S). Lower-permeability (L) and higher-permeability (H) catchments are shown separately. (a, b) S defined by D < 30 km, (c, d) S defined by E < 0.5, and (e, f) S defined by criteria E < 1.

Using $N = 1$ and all the catchments as equally weighted donors ($S = 126$, $B = 1/S$), the corresponding statistics were (0.14, 1.02, 0.14, 1.24), which illustrates the importance of conditioning the models on catchment similarity.

[28] Figure 3 shows the same results as Figure 1 but where prior likelihoods and performances have been quantified using the low-flow objective function, FSBM (equation (3)). This experiment allows assessment of whether a different regionalization scheme is warranted for low-flow studies. However, results reinforce the conclusion that using $S = 10$ with E defined by equation (4) is the best of all the tested schemes. Compared to NSE, there was slightly more evidence that higher similarity between the target and donor catchments led to improved ungauged catchment predictions; the correlation between average E (i.e., averaged over the 10 donor catchments) and FSBM was 0.27.

[29] The NSE and FSBM performances using SWA were also compared with the performances achieved by *Lee et al.* [2005b], who tested various MP-CD regression models as a basis for predicting flows in ungauged UK catchments, using the same data set as is used here, but excluded one catchment which produced an outlying MP set. They began with a set of 17 CDs plus their log transforms and applied univariate stepwise regression to identify the CDs significantly related to each of the MPs. Each of the 126 catchments in turn was considered to be ungauged and left out of the regression data, so that the regionalized MP values were truly independent of the local flow data. They also tested various options for prespecifying the significant CDs based on previous studies and experience. The comparisons of their best regression-based NSE values and the current SWA results are in Figures 4a–4d separately for low and high BFIHOST catchments and for the defined "calibration" and "validation" periods. In these four plots respectively, 62, 59, 62, and 66% of the results lie below the diagonal, i.e., SWA was better than regression in these percentages of catchments. Figures 4e–4h show the FSBM results in the same format, for which the corresponding statistics are 85, 67, 83, and 66%. The degree of improvement evident from Figure 4 is arguably marginal in the case of NSE, but is clearly substantial for FSBM.

[30] Figure 5 illustrates the time series of validation results for the Irvine River catchment (coordinates {0.18, 0.21} in Figure 1g, BFIHOST = 0.4, AREA = 75km^2, SAAR = 1352 mm) based on $N = 10$ (using NSE as prior likelihood criterion), $S = 10$ (using the E similarity measure). Figure 5a shows the ensemble of time series obtained using the optimum MP sets from all 126 prior candidate models, representing the uncertainty prior to considering the nature of the Irvine catchment. The high uncertainty is particularly notable in the base flow periods. However, the upper limit of the prior ensemble gives quite a robust

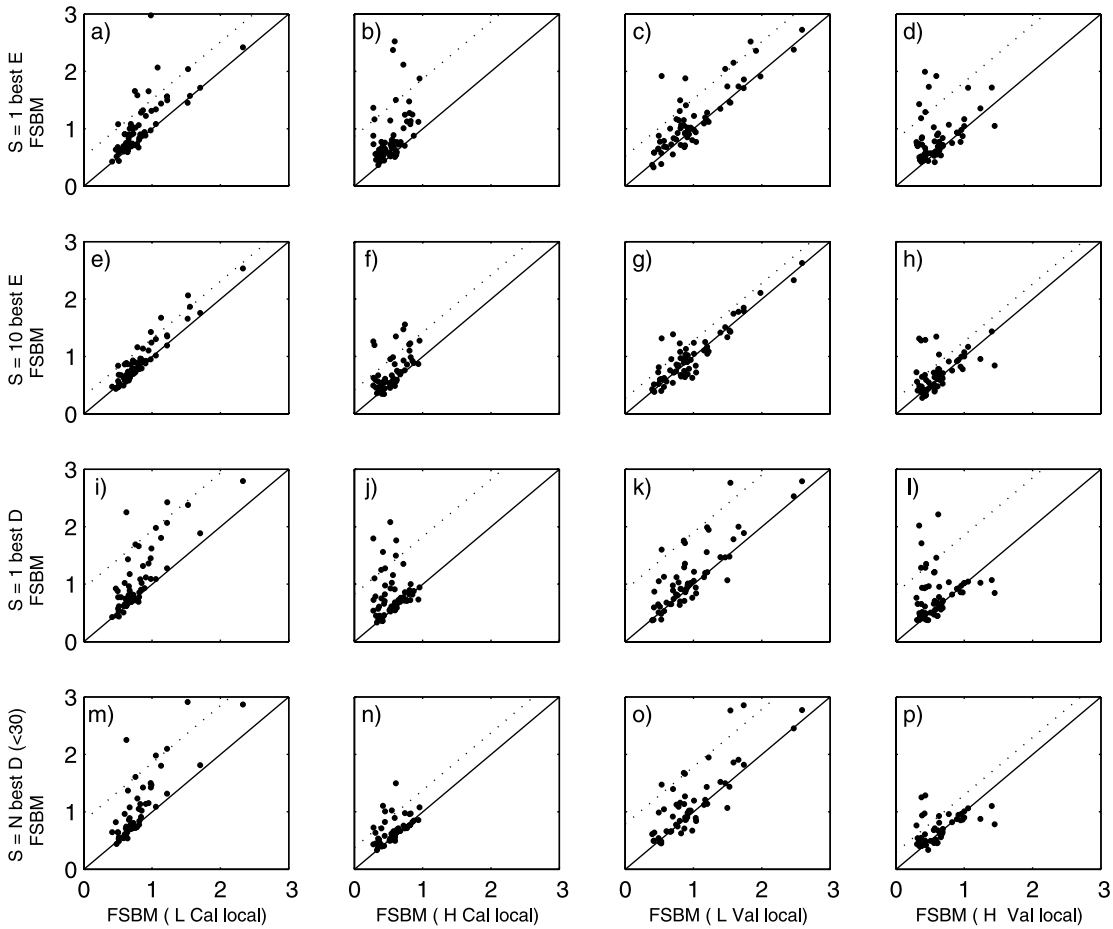

Figure 3. (a–p) Low-flow performance (FSBM) achieved using a locally calibrated model (x axes) plotted against performance achieved using the regionalized model (y axes, showing four alternative SWA schemes that are described in text), in calibration (Cal) and validation (Val) periods, for lower-permeability (L) and higher-permeability (H) catchments. Note that all axes are curtailed at FSBM = 3. Outliers not plotted have (x,y) coordinates of [2.1, 4.9] (Figure 3a); [0.7, 3.2] (Figure 3c); [5.0, 6.0] (Figure 3d); [2.1, 6.5], [1.1, 3.3] (Figure 3e); [1.9, 3.4], [1.8, 3.4] (Figure 3g); [5.0, 4.2] (Figure 3h); [1.5, 3.3], [2.9, 14.1], [0.90, 3.6], [1.6, 3.9] (Figure 3i); [1.7, 3.1], [1.9, 8.4], [1.8,3.6], [0.81, 3.3], [2.0, 3.6] (Figure 3k); [5.0, 3.9] (Figure 3l); [2.1,13.6], [1.1,3.4] (Figure 3m); [1.9, 8.1], [1.8, 3.6] (Figure 3o); and [5.0, 3.9] (Figure 3p).

representation of the flood peaks; few peaks are significantly overestimated and only isolated peaks are underestimated (this, presumably, being due to input rainfall error and/or model structural error). This indicates that the MP sets from only the flashiest catchments in the data set provide simulations which encompass the observed flood peaks in the Irvine catchment (suggesting a problem of flood peak underestimation in these flashiest catchments). Figure 5b shows the posterior ensemble using the 10 best MP sets from 10 most "similar" catchments. This maintains the robust representation of the flood peaks, and provides a much better ensemble of the overall flow regime. Figure 5c shows the 90% confidence limits derived from the posterior ensemble, showing that in this case the use of the percentiles (and hence the consideration of different likelihoods for candidate models) may not be desirable because the more extreme results in the ensemble, beyond the upper 90 percentile, are needed to encompass the flood peaks. Arguably, there may not be enough significantly different candidate models in this study to support meaningful calculation of 90% confidence limits. Figure 5d shows the weighted average of the posterior ensemble, and Figure 5e is the result obtained by *Lee et al*. [2005b] using their best regression scheme. Visual comparison shows that SWA is only marginally better than regression in this case.

[31] Figure 6 repeats the format of Figure 5 for the more permeable Evenlode catchment (coordinates {0.18, 0.25} in Figure 1h, BFIHOST = 0.7, AREA = 427 km^2, SAAR = 730 mm). In this case, the value of conditioning on similar catchments is emphatic for both low and high flows, and the use of 90 percentiles to remove the influence of the most extreme realizations has improved the visual representation of the data. Comparing Figures 6d and 6e shows that the SWA has significantly improved on the regression method, but high flows continue to be a challenge.

[32] Wider review of the time series fits leads to the conclusion that there is a general problem in fitting flow peaks, with the highest peaks almost always being underestimated by SWA's best estimate (i.e., the weighted average) prediction. However, this is also true for the locally

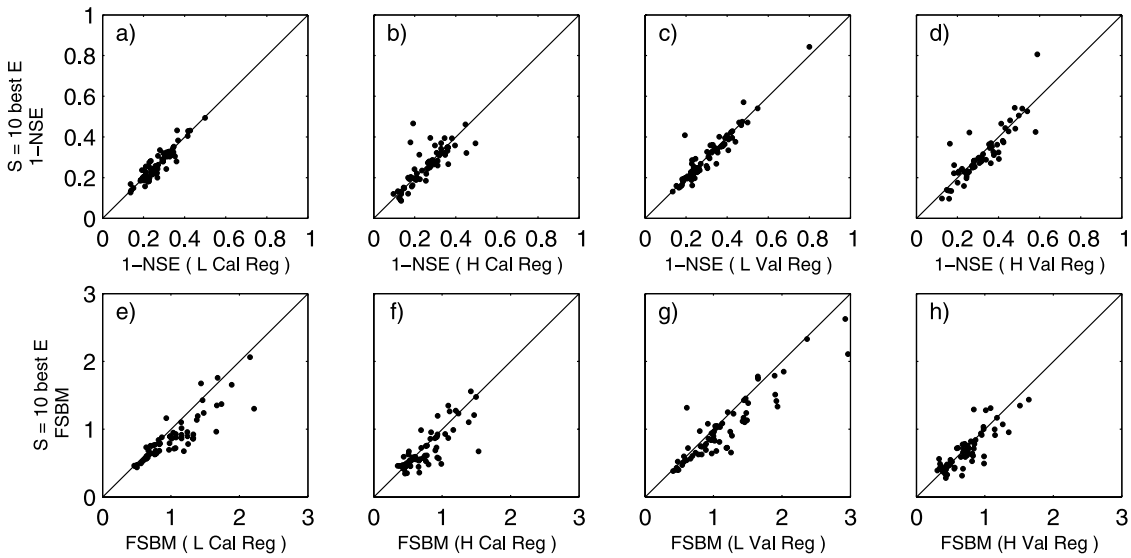

Figure 4. (a–h) Performances achieved using SWA (y axes) plotted against the best performances achieved by *Lee et al.* [2005b] using regression models on the same data set (x axes). NSE and FSBM performances are shown in calibration (Cal) and validation (Val) periods for lower-permeability (L) and higher-permeability (H) catchments separately. Note that outliers not plotted have (x,y) coordinates of [1.5, 1.7] (Figure 4b); [13.7, 15.3], [0.9, 1.1], [2.3, 2.5] (Figure 4d); [8.6,6.5], [4.2, 3.3] (Figure 4e); [4.1, 3.4], [4.6, 3.4] (Figure 4g); and [7.3, 4.2] (Figure 4h). In the catchment with an x coordinate of 13.57, only 1 year was available for the validation period, and the poor performance is due to errors in initial conditions.

calibrated model and is thought to be a fundamental limitation of using daily rainfall data (because peak rainfalls are averaged out), and neglecting the infiltration excess process, as well as due to rainfall errors. The SWA does generally very well at simulating low flows, frequently better than the locally calibrated model (evident on Figures 3g and 3h). The worst FSBM values relative to the locally calibrated model results arise from problems in fitting the hydrograph recessions; some simulated recessions are too smooth, others include flow variations which they should not, while others have poor recession response time. We speculate that this is due to effects of CDs not included in the employed measure of catchment similarity, for example land use, and due to local geology which may not be well represented by the BFIHOST values.

[33] The percentage of observed flow data within the ensemble limits indicates how successfully the prediction error is represented. Figures 7a–7d shows the number of target catchments for which different values of this percentage have been achieved, based on $N = 10$ (based on NSE), $S = 10$ (based on similarity E) calculated for the validation period. The more permeable and less permeable catchment groups are shown separately. Figures 7e–7h are the same format but for the 90% confidence limits. The general conclusion from Figure 7 is that the ensemble limits and associated percentiles are not effective at representing the possible errors in ungauged catchment predictions. Using just one model structure and only one realization of input rainfall in each calibration, it is arguably inevitable that the employed ensemble of MP sets will fail to encompass the actual prediction bias. Although using a larger number of prior candidate MP sets (e.g., $N = 10\%$ of the available sample), or changing the definition of weights P and B, may be used to increase the estimated prediction uncertainty, this seems unlikely to solve the flow peak estimation problem; instead, we may need to integrate input rainfall error and a wider range of model structures, potentially including distributed models [e.g., *Carpenter and Georgakakos*, 2004; *Georgakakos et al.*, 2004].

7. Concluding Discussion

[34] This paper has argued that a similarity weighted averaging (SWA) may be a constructive way forward toward improved flow prediction in ungauged catchments. The proposed procedure builds upon established Bayesian methods [*Beven and Freer*, 2001; *Neuman*, 2003] and preliminary applications of model averaging to the ungauged catchment problem [*Young*, 2000]. A large sample of models of gauged donor catchments are taken, and the prior likelihoods of these models are updated based on the similarity of the respective donor catchment to the target ungauged catchment. All models with nonzero posterior likelihood are applied, allowing a weighted average of the ensemble to be calculated as the best estimate, and confidence limits to be estimated. The main theoretical advantage of this method, as opposed to more traditionally applied methods based on univariate regression of MPs against CDs [e.g., *Sefton and Howarth*, 1998; *Seibert*, 1999; *Kokkonen et al.*, 2003; *Merz and Blöschl*, 2004; *Wagener et al.*, 2004], is that MP sets are maintained as sets, so that the MP interdependencies are not neglected or linearized to facilitate regression. The method has been applied to a database of observed daily rainfall-runoff from 127 UK catchments, which are all nonurban but otherwise widely varied in nature (in the UK context). Using a similarity

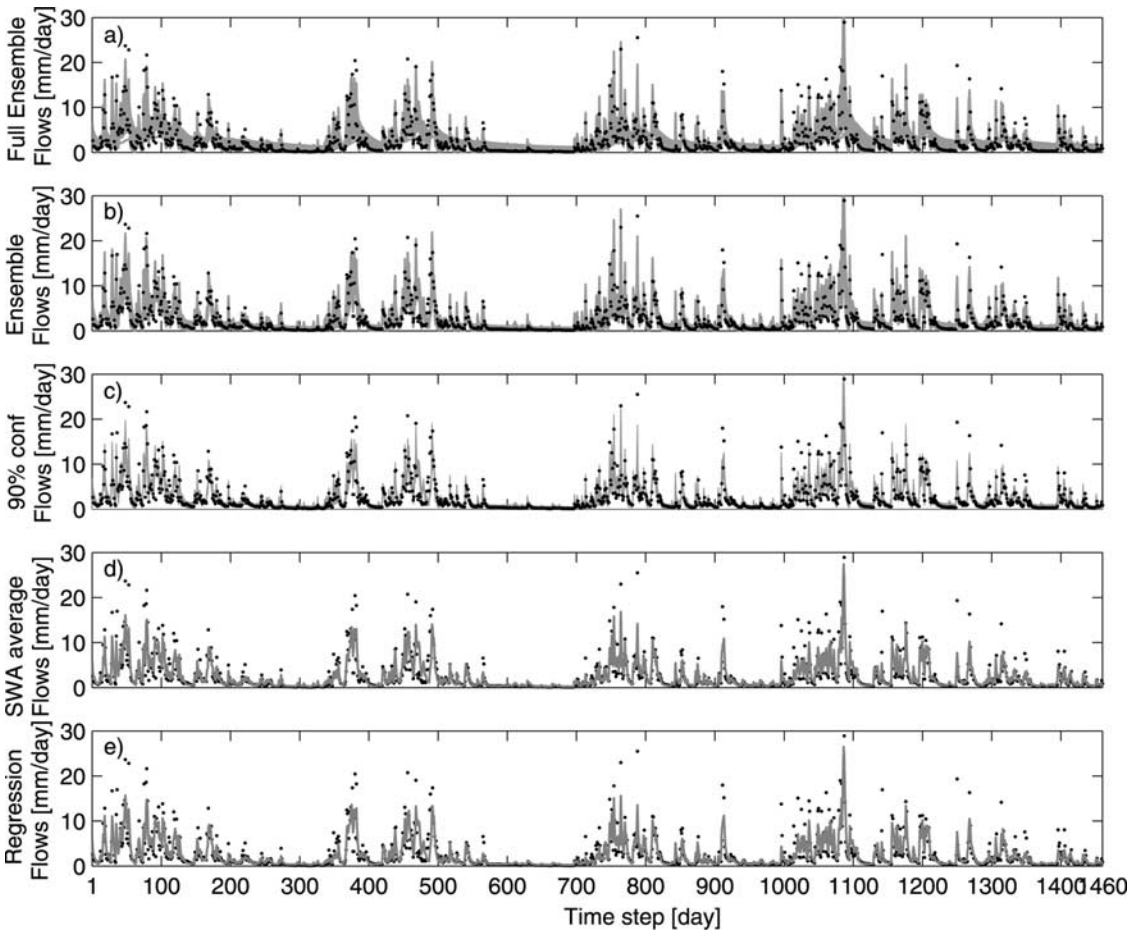

Figure 5. Time series results for River Irvine in the validation period. (a) Outline of the ensemble using optimum MP sets from all 126 prior candidate models. (b) Outline of the posterior ensemble using the 10 best MP sets from the 10 most similar catchments. (c) The 90% confidence limits derived from the posterior ensemble. (d) Weighted average of the posterior ensemble. (e) Best result obtained by *Lee et al.* [2005b] using MP-CD regression.

measure based on catchment size, permeability and rainfall, SWA generally out performed models based on regression, most notably within the low-flow part of the hydrograph. A persistent problem is underestimation of flow peaks; in many cases, even the full ensemble of prior models (i.e., models of all catchment types) failed to encompass the observed flow peaks, although to some extent, this is an expected consequence of the daily time step used in the modeling.

[35] Although the results achieved within this research are very promising, a number of issues remain for discussion and future attention. It may be argued that the primary source of prediction bias comes from model structure error (the limited NSE performance of the locally calibrated models is notable in Figures 1a–1d, for example). Although SWA is well suited to integrating more than one model structure, we have chosen to not do so here. This choice was based on previous research which showed that there was very little difference between the optimized results of various alternative lumped conceptual models of similar complexity [*Lee et al.*, 2004, 2005a]. Therefore a priority for further work is to integrate the results of a wider range of model types [*Neuman*, 2003; *Georgakakos et al.*, 2004; *Butts et al.*, 2004]. The method is also well suited to allowing for different realizations of input-output errors during the calibration process, which would result in a more objective, and potentially more useful, set of prior MPs for each gauged catchment. In this paper, the chosen similarity measures are based on previous reports, and they do not explicitly account for CD uncertainty. Further research might reveal, for example, that individual CDs should be given less weight than others due to their high uncertainty, although objective estimation of this uncertainty may be a problem in itself. Furthermore, while arbitrary values of S or E_T (see equation (5)) have been used here and subject to limited sensitivity analysis, it may in future be possible to use more objectively founded values, or to empirically optimize their values for a given set of catchments.

[36] Behind the SWA method is the premise that enough models of nonzero posterior likelihood can be sampled to sufficiently represent the continuous MP-CD response surface. The sample numbers actually used were relatively small (several hundred at most), giving a sparse sample from the eight-dimensional MP-CD space. As results were not sensitive to the number of equifinal models derived from each donor catchment, it seems that the limitation in the number of similar gauged catchments is likely to be more important, at least within the case study used here. It is reasonable to speculate that a larger database of gauged

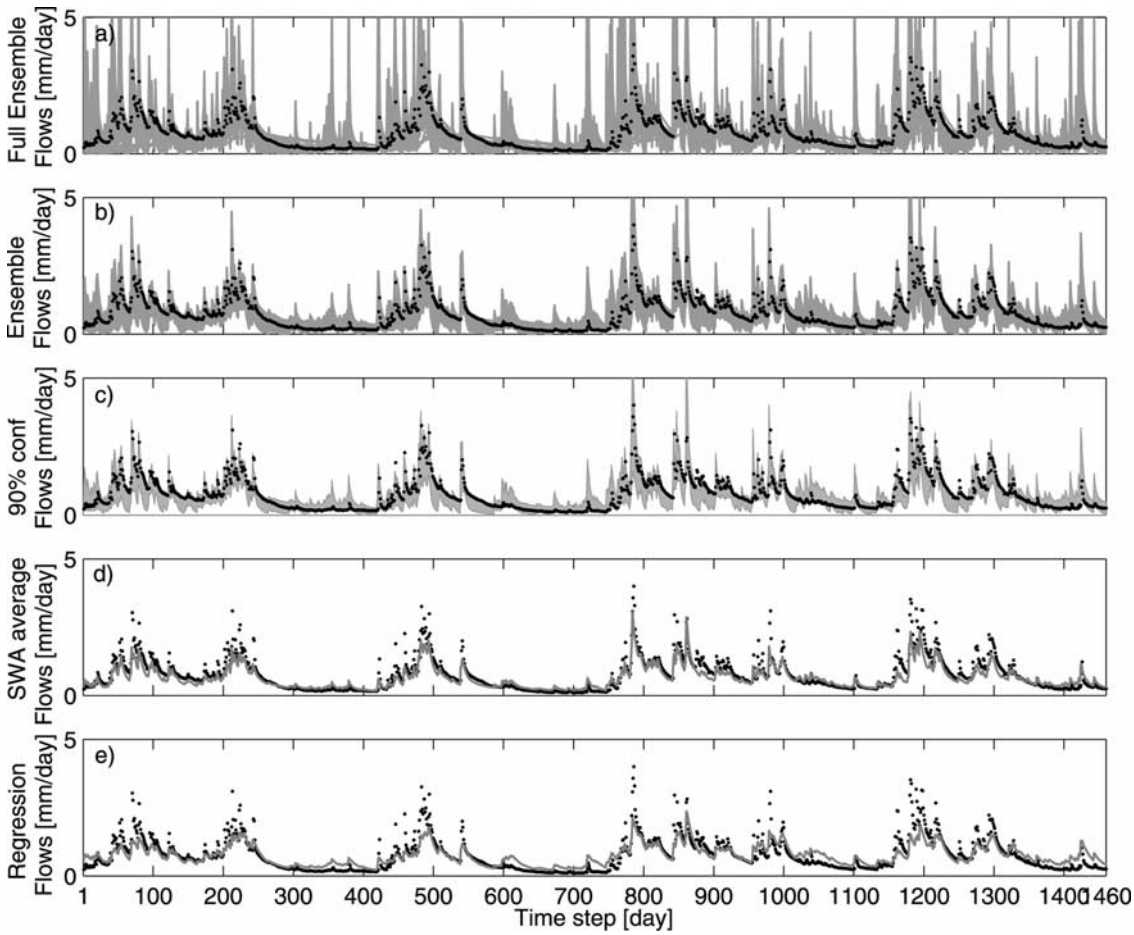

Figure 6. Time series results for Evenlode River in the validation period. (a) Outline of the ensemble using optimum MP sets from all 126 prior candidate models. (b) Outline of the posterior ensemble using the 10 best MP sets from the 10 most ''similar'' catchments. (c) The 90% confidence limits derived from the posterior ensemble. (d) Weighted average of the posterior ensemble. (e) Best result obtained by *Lee et al.* [2005b] using MP-CD regression.

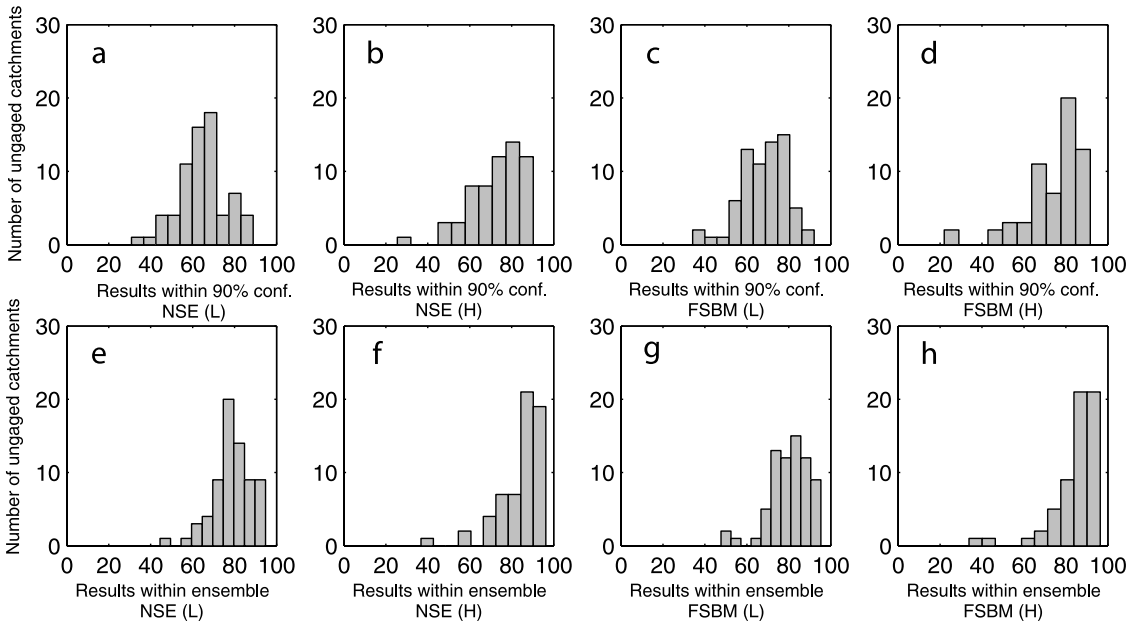

Figure 7. Frequency of ungauged catchments against proportion of observed data within the prediction ensemble limits and 90% confidence limits, in validation periods. Results are shown for the cases of using NSE and FSBM to define prior likelihood for both the more permeable (H) and less permeable (L) catchment groups.

catchments would lead to higher similarities between target and donor catchments, and better ungauged catchment predictions. There was some evidence of this, especially when looking at low-flow performance, FSBM. It might be the case that regression-based methods of MP estimation, or more physically based alternatives would be preferred only when a large degree of interpolation between catchments is required. However, this has not yet been investigated and in general more research is needed to support informed selection of suitable regionalization schemes.

Notation

c moisture storage capacity.
Q modeled runoff.
W posterior likelihood.
P prior likelihood.
B conditioning likelihood.
E catchment similarity measure.
D catchment outlet proximity.
S number of donor catchments.
N number of candidate models per donor catchment.
Qo observed runoff.

References

Anderson, M. L., M. L. Kavvas, and M. D. Mierzwa (2001), Probabilistic/ensemble forecasting: A case study using hydrologic response distributions associated with El Nino/Southern Oscillation (ENSO), *J. Hydrol.*, *249*, 134–147.

Andreassian, V., C. Perrin, C. Michel, I. Usart-Sanchez, and J. Lavabre (2001), Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, *J. Hydrol.*, *250*, 206–223.

Boorman, D. B., J. M. Hollis, and A. Lilly (1995), Hydrology of soil types: A hydrologically-based classification of the soils of the United Kingdom, *Rep. 126*, Inst. of Hydrol., Wallingford, U. K.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and predictive uncertainty, *Hydrol. Processes*, *6*, 279–298.

Beven, K. J., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266.

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

Carpenter, T. M., and K. P. Georgakakos (2004), Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model, *J. Hydrol.*, *298*, 202–221.

Christensen, N. S., A. W. Wood, N. Voisin, D. P. Lettenmaier, and R. N. Palmer (2004), The effects of climate change on the hydrology and water resources of the Colorado River basin, *Clim. Change*, *62*(1–3), 337–363.

Duan, Q., V. K. Gupta, and S. Sorooshian (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*, 1015–1031.

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, *298*, 222–241.

Gupta, H. V., K. J. Beven, and T. Wagener (2005), Model calibration and uncertainty estimation, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson et al., John Wiley, Hoboken, N. J., in press.

Holmes, M., A. Young, A. Gustard, and R. Grew (2002), A region of influence approach to predicting flow duration curves within ungauged catchments, *Hydrol. Earth Syst. Sci.*, *6*(4), 721–731.

Institute of Hydrology (1999), *Flood Estimation Handbook*, vol. 5, Wallingford, U. K.

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model, *Water Resour. Res.*, *29*, 2637–2649.

Kokkonen, T., A. Jakeman, P. Young, and H. Koivusalo (2003), Predicting daily flows in ungauged catchments: Model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina, *Hydrol. Processes*, *17*(11), 2219–2238.

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.

Lee, H., N. McIntyre, H. Wheater, A. Young, and T. Wagener (2004), Assessment of rainfall-runoff model structures for regionalisation purposes, in *Hydrology—Science and Practice for the 21st Century*, vol. 1, edited by B. Webb et al., pp. 302–308, Br. Hydrol. Soc., London.

Lee, H., N. R. McIntyre, H. S. Wheater, and A. R. Young (2005a), Selection of conceptual models for regionalisation of the rainfall-runoff relationship, *J. Hydrol.*, *312*, 125–147.

Lee, H., N. R. McIntyre, H. S. Wheater, and A. R. Young (2005b), Predicting runoff in ungauged UK catchments, *Proc. Inst. Civ. Eng. Water Manage.*, in press.

Legates, D. R., and G. J. McCabe Jr. (1999), Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, *35*(1), 233–242.

McIntyre, N. R., H. Lee, H. S. Wheater, and A. R. Young (2004), Tools and approaches for evaluating uncertainty in streamflow predictions in ungauged UK catchments, in *Complexity and Integrated Resources Management* [CD-ROM], edited by C. Pahl-Wostl, S. Schmidt, and T. Jakeman, Int. Environ. Modell. and Software Soc., Manno, Switzerland.

Merz, R., and G. Blöschl (2004), Regionalisation of catchment model parameters, *J. Hydrol.*, *287*, 95–123.

Merz, R., and G. Blöschl (2005), Flood frequency regionalisation-spatial proximity vs. catchment attributes, *J. Hydrol.*, *302*, 283–306.

Manley, R. E. (1978), Simulation of flows in ungauged basins, *Hydrol. Sci. Bull.*, *23*(1), 85–101.

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

Nash, J. E. (1960), A unit hydrograph study with particular reference to British catchments, *Proc. Inst. Civ. Eng.*, *17*, 249–282.

Neuman, S. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*, 291–305.

O'Connell, P., and E. Todini (1996), Modelling of rainfall, flow and mass transport in hydrological systems: An overview, *J. Hydrol.*, *175*, 3–16.

Perrin, C., C. Michel, and V. Andreassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, *242*, 275–301.

Sefton, C. E. M., and S. M. Howarth (1998), Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales, *J. Hydrol.*, *211*, 1–16.

Seibert, J. (1999), Regionalisation of parameters for a conceptual rainfall-runoff model, *Agric. For. Meteorol.*, *98-99*(31), 279–293.

Shamseldin, A., K. O'Connor, and G. Liang (1997), Methods for combining the output of different rainfall-runoff models, *J. Hydrol.*, *197*, 203–229.

Singh, V. P., and D. Frevert (2002a), *Mathematical Models of Large Watershed Hydrology*, vol. 1, Water Resour. Publ., Highlands Ranch, Colo.

Singh, V. P., and D. Frevert (2002b), *Mathematical Models of Small Watershed Hydrology*, vol. 2, Water Resour. Publ., Highlands Ranch, Colo.

Sivapalan, M. (2003), Prediction in ungauged basins: A grand challenge for theoretical hydrology, *Hydrol. Processes*, *17*(15), 3163–3170.

Sorooshian, S., and V. K. Gupta (1995), Model calibration, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 23–68, Water Resour. Publ., Highlands Ranch, Colo.

Tung, Y., K. Yeh, and J. Yang (1997), Regionalization of unit hydrograph parameters: Comparison of regression analysis techniques, *Stochastic Hydrol. Hydraul.*, *11*, 145–171.

Vandewiele, G. L., and A. Elias (1995), Monthly water balance of ungauged catchments obtained by geographical regionalization, *J. Hydrol.*, *170*, 277–291.

Wagener, T., and N. McIntyre (2005), Identification of rainfall-runoff models for operational applications, *Hydrol. Sci. J.*, *50*(5), 735–751.

Wagener, T., and H. S. Wheater (2005), Parameter estimation and regionalisation of continuous rainfall-runoff models, *J. Hydrol.*, in press.

Wagener, T., H. S. Wheater, and H. V. Gupta (2004), *Rainfall-Runoff Modelling in Gauged and Ungauged Catchments*, Imperial College Press, London.

Wetherald, R. T., and S. Manabe (2002), Simulation of hydrologic changes associated with global warming, *J. Geophys. Res.*, *107*(D19), 4379, doi:10.1029/2001JD001195.

Wheater, H. S. (2002), Progress in and prospects for fluvial flood modelling, *Philos. Trans. R. Soc. London, Ser. A*, *360*(1796), 1409–1431.

Wheater, H. S., A. J. Jakeman, and K. Beven (1993), Progress and destinations in rainfall-runoff modelling, in *Modelling Change in Environmental Systems*, edited by A. J. Jakeman, M. B. Beck, and M. J. McAleer, pp. 101–132, John Wiley, Hoboken, N. J.

Yeh, K. C., J. C. Yang, and Y. K. Tung (1997), Regionalization of unit hydrograph parameters: 2. Uncertainty analysis, *Stochastic Hydrol. Hydraul.*, *11*, 173–192.

Young, A. (2000), Regionalising a daily rainfall-runoff model within the United Kingdom, Ph.D. thesis, Univ. of Southampton, Southampton, U. K.

H. Lee, N. McIntyre, and H. Wheater, Department of Civil and Environmental Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. (hyo.lee@imperial.ac.uk; n.mcintyre@imperial.ac.uk; h.wheater@imperial.ac.uk)

T. Wagener, Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA 16802, USA. (thorsten@engr.psu.edu)

A. Young, Centre for Ecology and Hydrology, Wallingford, Oxfordshire OX10 8BB, UK. (ary@ceh.ac.uk)