Author/s:
Reichl, JPC;Western, AW;McIntyre, NR;Chiew, FHS

Title:
Optimization of a similarity measure for estimating ungauged streamflow

Date:
2009-10-01

Citation:
Reichl, J. P. C., Western, A. W., McIntyre, N. R. & Chiew, F. H. S. (2009). Optimization of a similarity measure for estimating ungauged streamflow. Water Resources Research, 45 (10), https://doi.org/10.1029/2008WR007248.

Persistent Link:
https://hdl.handle.net/11343/297411

Optimization of a similarity measure for estimating ungauged streamflow

J. P. C. Reichl,[1,2] A. W. Western,[1,2] N. R. McIntyre,[3] and F. H. S. Chiew[2,4]

Received 27 June 2008; revised 1 July 2009; accepted 9 July 2009; published 17 October 2009.

[1] One approach to predicting streamflow in an ungauged catchment is to select an ensemble of hydrological models previously identified for similar gauged catchments, where the similarity is based on some combination of important physical catchment attributes. The focus of this paper is the identification of catchment attributes and optimization of a similarity measure to produce the best possible ungauged streamflow predictions given a data set and a conceptual model structure. As a case study, the SimHyd rainfall-runoff model is applied to simulate monthly streamflow in 184 Australian catchments. Initial results show that none of 27 catchment attributes can be safely said to consistently give a better ensemble of models than random selection when used independently of other attributes. This is contrary to prior expectations and indicates the sparseness of information within our database of catchments, the importance in this case of prior knowledge for defining important attributes, and the potential importance of combining multiple attributes in order to usefully gauge similarity. Seven relatively independent attributes are then selected on the basis of prior knowledge. The weight with which each of these attributes contributes to the similarity measure is optimized to maximize streamflow prediction performance across a set of 95 catchments. The other 89 catchments are used to independently test the accuracy of streamflow predictions. Using the optimal set of weights led to marked improvement in the accuracy of predictions, showing that the method, while inferior to local calibration, is superior to alternative methods of model regionalization based on regression and spatial proximity. However, there is evidence of nonuniqueness in the optimal solution and the possibility that the attribute weights are somewhat dependent on the catchments used.

Citation: Reichl, J. P. C., A. W. Western, N. R. McIntyre, and F. H. S. Chiew (2009), Optimization of a similarity measure for estimating ungauged streamflow, *Water Resour. Res.*, *45*, W10423, doi:10.1029/2008WR007248.

1. Introduction

[2] When studying natural systems, hydrologists employ simplified models. Although often built upon the modeler's conceptual understanding of the system, the components of the model rarely represent true components of the system. This means that the model parameters cannot be measured and must be conditioned to reproduce observed data. In the common case that we lack sufficient data to condition the model (the "ungauged" case), the model parameters or hydrological behavior must be inferred somehow, a process historically referred to as "regionalization" [*Vogel*, 2005].

[3] Of course, we cannot hope to perfectly predict even locally gauged streamflow when utilizing erroneous data that are coarsely sampled both spatially and temporally. The large errors and simplifications involved in using surrogate information for inference of local behavior will likely hinder current and future attempts at regionalization. However, seeking to improve upon our thus far limited skill remains an important task, and should certainly be pursued. Increased understanding will certainly help in this regard, as is seen in the improvements in predictive ability that have been made over recent decades.

[4] Issues surrounding regionalization are well known and have been discussed at length, and as such a thorough review will not be undertaken here; the reader is referred instead to an excellent and thorough historical review of regionalization techniques by *Vogel* [2005]. Briefly, however, there are two common approaches to regionalizing hydrological models. The first is to use calibrated model parameter sets from the nearest gauged catchment, and apply them to the ungauged catchment. This is referred to as the "spatial proximity" approach herein. This approach assumes that the major controllers of the catchment processes vary smoothly in space, and requires a certain degree of gauging density. The second common approach to regionalization is to develop a regional relationship between individual calibrated model parameters (from a set of gauged catchments) and catchment attributes, referred to as the "regression" approach herein. The relationship can then be used to estimate the model parameters for the

[1]Department of Civil and Environmental Engineering, University of Melbourne, Melbourne, Victoria, Australia.
[2]eWater CRC, University of Canberra, Canberra, ACT, Australia.
[3]Department of Civil and Environmental Engineering, Imperial College, London, UK.
[4]CSIRO Land and Water, Canberra, ACT, Australia.

0043-1397/09/2008WR007248

ungauged catchment. This approach, while intuitive, suffers from our inability to uniquely identify model parameters (the result of both overly complex models and errors in forcing and response data) and from the need to neglect the complex parameter interdependencies inherent to calibration results [*McIntyre et al.*, 2005]. These parameter interactions can result in multiple equivalent parameter sets: the equifinality problem of *Beven and Freer* [2001]. The key failing of regression approaches is the neglect or simplification of these interactions.

[5] In the regionalization context, the problem of parameter identifiability can be partially overcome by estimating the standard error of the regression, and hence estimating confidence intervals on parameters and flow predictions [*Wagener and Wheater*, 2006]. However, this does not overcome the fundamental problem that regression does not conserve the parameter interdependencies associated with conceptual models. A method that better captures these interdependencies is required.

[6] *Fernandez et al.* [2000] and *Vogel* [2005], and, more recently, *Bastola et al.* [2008] attempt to strengthen regional regression relationships, while at the same time mitigating the effect of strong model parameter covariance, by optimizing both concurrently. *Fernandez et al.* [2000] and *Vogel* [2005] found that the strong regional regression relationships that resulted were very similar to those produced when optimizing them separately, and therefore produced similarly poor streamflow estimates. *Bastola et al.* [2008], however, report improved prediction of streamflow and uncertainty compared to traditional two step regression methods (optimization of model parameters followed by regression of these against catchment attributes). Similarly, *Hundecha and Bardossy* [2004] and *Hundecha et al.* [2008] optimize, in a single step, rainfall-runoff model parameters and their spatial structure within physiographic-climatic spaces. Parameters for catchments not used to define these spaces are defined via ordinary kriging; however, the authors suggest that the approach may be useful for extrapolating to catchments that lie slightly outside the bounds of the defined physiographic-climatic spaces.

[7] Techniques which apply multiple parameter sets to create an ensemble streamflow estimate provide an opportunity to overcome these problems, as they forego the need to introduce simplifying assumptions about parameter distributions and interdependence. *McIntyre et al.* [2005] describe a method which applies calibrated parameter sets intact to an ungauged catchment. The method is a generalization of the spatial proximity approach; however, it is based on proximity in catchment attribute space (physical similarity) rather than in geographic space. The premise is that there is some continuous surface relating model parameter sets to catchment attributes, and if sampled adequately in the correct region, this should represent viable models for the target ungauged catchment. Although *McIntyre et al.* [2005] demonstrated advantages of this method over the regression approach, their exploration of catchment similarity measures was very limited, and this issue was highlighted as a priority for further research.

[8] *Oudin et al.* [2008] applied a method similar to that of *McIntyre et al.* [2005] to 913 French catchments, again with promising results. Some exploration of catchment similarity measures was undertaken, but in a basic sense, with simple combinations of attributes tested. This study focused on investigating the respective interests of spatial proximity and physical similarity in a model averaging context. *Oudin et al.* [2008] found that an approach based simply on spatial proximity performed slightly better than the model averaging approach based on physical similarity. The benefits of using spatial proximity as a surrogate for more complex information about hydrologic behavior have been discussed in several papers [*Merz and Blöschl*, 2004]. *Oudin et al.* [2008] use a relatively dense gauging network, which no doubt increases the validity of the assumptions behind approaches based on spatial proximity.

[9] This study takes the work of *McIntyre et al.* [2005] and *Oudin et al.* [2008] further, by describing a method for identification of region-specific measures of catchment hydrologic similarity within an optimization framework. Section 2 presents a discussion of catchment hydrologic similarity, and a review of catchment attributes considered important in this context. Section 3 describes the data set and rainfall-runoff model used, the model averaging framework and equations, and the methodology used for identification of similarity measures. Section 4 presents and discusses results of application to 184 gauged catchments in Australia. First, these are discussed in terms of the identification of catchment attributes that will be useful in the optimization of similarity measures; next, interpretation of the hydrologic significance of these similarity measures is discussed, along with their sensitivity to assumptions within the method and their usefulness in estimating ungauged streamflow. The quality of predictions from the model averaging framework is compared with those using traditional regression techniques and with those based on spatial proximity. Finally, conclusions are drawn as to the suitability of the optimization method to identification of catchment hydrologic similarity measures, and of the model averaging framework to estimation of ungauged Australian streamflow.

2. Measuring Catchment Hydrologic Similarity

[10] The question of what makes two catchments hydrologically similar is of fundamental importance to the understanding of catchment hydrology. In addressing it, we are addressing the underlying question of what controls various aspects of hydrological response. The current limited ability to predict the runoff response of an ungauged catchment to a particular forcing regime may point in part to a lack of fundamental understanding. *Wagener et al.* [2007, p. 901] raised the issue of a need for a general classification system in hydrology to provide "an organizing principle, create a common language, guide modeling and measurement efforts, and provide constraints on predictions in ungauged basins, as well as on estimates of environmental change impacts." Optimizing similarity metrics that are useful in estimating ungauged catchment behavior aims to serve the immediate problem of ungauged streamflow estimation, but through the proposition of important attributes and exclusion of others, it also contributes toward the goal of hydrologic classification, and to debate on the question of hydrologic similarity.

[11] A review of the literature yields many attributes that the hydrologic community has thought important in the

Table 1. Catchment Attributes Used in This Study and Their Descriptions

Attribute	Description	Category	Symbol
μCosAspect	mean of cos(aspect)	geomorphic	$\mu_{\cos(As)}$
σCosAspect	standard deviation of cos(aspect)	geomorphic	$\sigma_{\cos(As)}$
MinElevation	minimum elevation (m)	geomorphic	Min_z
MaxElevation	maximum elevation (m)	geomorphic	Max_z
μElevation	mean elevation (m)	geomorphic	μ_z
ERR	elevation-relief ratio (μElevation-MinElevation)/(MaxElevation-MinElevation)	geomorphic	ERR
μSlope	mean slope (deg)	geomorphic	μ_β
σSlope	standard deviation of slope (deg)	geomorphic	σ_β
μX	mean and standard deviation of the X and Y projections of the unit normal vector to the elevation surface	geomorphic	μ_X
σX	mean and standard deviation of the X and Y projections of the unit normal vector to the elevation surface	geomorphic	σ_X
μY	mean and standard deviation of the X and Y projections of the unit normal vector to the elevation surface	geomorphic	μ_Y
σY	mean and standard deviation of the X and Y projections of the unit normal vector to the elevation surface	geomorphic	σ_Y
μX-Y	mean distance in X-Y space	geomorphic	$\mu_{X\text{-}Y}$
Links	number of stream links	geomorphic	Links
Area	catchment area (km^2)	geomorphic	Area
LinkDensity	stream links per area (km^{-2})	geomorphic	ρ_{Link}
μWP	mean winter precipitation (mm)	climatic	μWP
μSP	mean summer precipitation (mm)	climatic	μSP
μAP	mean annual precipitation (mm)	climatic	μAP
Aridity	mean annual areal potential evapotranspiration (MAAPET)/MAP	climatic	Ar
SoilDepth	mean soil depth (m)	soils	Soil
PAWHC	mean plant available water holding capacity (mm)	soils	PAWHC
A_KSAT	A-horizon saturated hydraulic conductivity (m^3/s)	soils	A_KSAT
Transmissivity	lateral transmissivity (m^2/s)	soils	Tr
FractionNative	fraction covered by native woody vegetation	vegetation	F_{NWV}
FractionWoody	fraction covered by woody vegetation	vegetation	F_{TWV}
Proximity	centroid-centroid (DD)	-	Pr

rainfall-runoff process. They are many, varied, and often strongly correlated or slightly different representations of the same attributes, confirming the sentiment expressed by *Wagener et al.* [2007], that there is a need for a general, common classification system.

[12] Reviewing many studies in Australia, the United Kingdom, Europe and the United States [*Beven and Kirkby*, 1979; *Chiew and Siriwardena*, 2005; *Fernandez et al.*, 2000; *Hundecha and Bardossy*, 2004; *Hundecha et al.*, 2008; *Laaha and Blöschl*, 2006; *Lowe and Nathan*, 2006; *McIntyre et al.*, 2005; *Merz and Blöschl*, 2004; *Nathan and McMahon*, 1990; *Parajka et al.*, 2005; *Peel et al.*, 2000; *Post and Jakeman*, 1999; *Sefton and Howarth*, 1998; *Vogel*, 2005; *Wagener and Wheater*, 2006; *Young*, 2006], yields a collection of catchment attributes considered by the hydrological community to be important in this context. Table 1 contains the attributes used in our study, grouped into geomorphologic, climatic, vegetation and soil attribute types. There is an obvious commonality of attributes used in these studies. They all use indicators of geomorphology, vegetation cover, climate and soil properties. It is likely that this commonality is largely the result of the data that is available, but also of the collective understanding of hydrologists of the importance of these catchment attributes.

[13] Considering the many efforts toward this problem of regional streamflow estimation, and the obvious need for more structured and defensible approaches to the question of catchment similarity, this paper aims to contribute to these debates by optimizing dimensionless measures of catchment similarity.

3. Methods and Data

3.1. Model

[14] The method was developed using the lumped conceptual daily rainfall-runoff model SimHyd. SimHyd is driven by daily precipitation and long-term monthly average potential evapotranspiration (PET), and simulates daily streamflow and evapotranspiration (ET). It has been tested and used extensively across Australia [*Chiew et al.*, 2002]. This version of SimHyd has seven model parameters, $\theta_{1:7}$, described in Table 2. *Chiew et al.* [2002] provides the full description of the model algorithm. SimHyd runs at a daily time step, however *Chiew and Siriwardena* [2005] recommend that it be conditioned to reproduce monthly flow,

Table 2. Parameters of the SimHyd Lumped Conceptual Rainfall-Runoff Model

Parameter	Description
θ_1	interception store capacity (mm)
θ_2	maximum infiltration loss (mm)
θ_3	infiltration loss exponent
θ_4	soil moisture store capacity (mm)
θ_5	constant of proportionality in interflow equation
θ_6	constant of proportionality in groundwater recharge equation
θ_7	base flow linear recession parameter

Figure 1. Maps showing the Köppen climate Cfa (medium gray) and Cfb (dark gray) regions as well as the distribution of catchments (in black) across the country for (left) development and (right) test groups. Note that 12 development and 11 test catchments fall outside these climate regions.

particularly when creating regional regression relationships. This reduces the complexity of the model (2 routing parameters are required for simulation of daily flow). The simpler model suffers less from strong parameter interactions which hinder meaningful identification of these relationships. SimHyd is therefore conditioned to minimize the difference between monthly modeled and observed flows using a quasi-Newtonian optimizer (see section 3.3 for objective functions).

3.2. Data

[15] The precipitation, potential evapotranspiration and streamflow data used for this study are a subset of the National Land and Water Resources Audit (NLWRA) data set [*Peel et al.*, 2000]. All data are observed rather than modeled data. Since the full set of 329 NLWRA catchments have data of varied lengths and periods a common period was sought in order to eliminate potentially confounding effects of using disparate time periods, such as the model parameter sensitivity to hydroclimatic variability reported by *Le Lay et al.* [2007]. Following the report of *Jakeman and Hornberger* [1993] on the data requirements of models of varying degrees of parameterization, it is considered that 10–15 years of monthly data are adequate for determination of SimHyd's parameters, allowing for significant instances of missing data. A requirement is that the data period contains, for the majority of catchments, good quality data for both wet and dry periods, in order to adequately exercise the model parameters [*Gupta and Sorooshian*, 1983]. The period 1972–1985 fulfils the requirement, while maintaining 184 catchments for the study (Figure 1). These catchments and their physical characteristics are included in the auxiliary material.[1] The set of catchments is randomly split into two subsets, one to be used for the optimization of similarity measures (the "development" set of 95 catchments) and one for independent validation (the "test" group of 89 catchments). Several catchments were discounted due to suspect data identified after this random split. These catchments had negative Nash-Sutcliffe efficiency results in validation. This resulted in the small difference in the numbers of catchments in the two groups.

[16] The catchments are further divided into Köppen climatic groups, shown in Figure 1, using a recently updated Köppen-Geiger climate map [*Peel et al.*, 2007]. The Cfa and Cfb climate types predominate and are defined as being temperate and without a dry season, and are distinguished by having hot and warm summers, respectively. Coastal NSW and southern Queensland are predominantly of Cfa type, and Tasmania and Victoria are predominantly of Cfb type. Twenty-seven of the 95 development catchments are of Cfa climate type and 56 of Cfb; 23 of the 89 test catchments are of Cfa climate type and 55 of Cfb. Note that when conducting experiments on the two Köppen climate regions, catchments from the whole country are available to donate model parameter sets (these catchments are termed "donors" or "donor catchments" herein). This generally results in better simulations than if donors are available only from within the region. It allows development of region-specific similarity measures, but with a larger pool of available donor catchments.

[17] The soil data are available as a 1 km grid developed by A. Western and N. J. McKenzie (Soil hydrologic properties of Australia, 2004, Cooperative Research Centre for Catchment Hydrology, Canberra, available at www.toolkit.net.au/shpa). The vegetation data are available as a 1 km grid developed by A. Western (Land cover for the intensive use zone of Australia, 2005, Cooperative Research Centre for Catchment Hydrology, Canberra, available at www.toolkit.net.au/liza), and are based on the vegetation cover as of 1995. The geomorphic data are taken from a 250 m DEM (Spatial Information Council of Australia and New Zealand (ANZLIC), GEODATA 9 second digital elevation model (DEM-9S) ANZCW0703005624, 2002, available at www.ga.gov.au). Table 1 provides a list of the 27 catchment attributes used in this study, grouped into climatic, geomorphic, vegetation and soil properties. The catchment-average values of these were extracted from the above data sets.

3.3. Ensemble Techniques and Model Averaging

[18] For ease of comparison, terminology and symbology have, where possible, been kept consistent with that of *McIntyre et al.* [2005]. It is important to note, especially in light of recent debate on the subject [e.g., *Beven*, 2006; *Montanari*, 2005; *Stedinger et al.*, 2008] that while this paper refers to "likelihoods" (described in this section and sections 3.3.1 and 3.3.2), we do not derive formal likelihood measures. The definitions of likelihood that we use are akin to those of *Beven and Freer* [2001], and are considered defensible in that they provide us with an objective assessment of how likely the candidate model is to suit the target catchment.

[1] Auxiliary materials are available at ftp://ftp.agu.org/apend/wr/2008wr007248.

[19] The concept of model averaging lends itself ideally to the problem of ungauged streamflow prediction. As discussed by *McIntyre et al.* [2005], the relationship between model parameters and catchment attributes should be thought of not as a linear or deterministic one, based on a single realization, but as a continuous likelihood surface. Accepting that each catchment is unique, defined by a unique set of attributes and having a unique response, in theory this surface has infinite dimensions, and is therefore impossible to define deterministically. All we can hope to do is identify, by sampling sets of model parameters and attributes derived from gauged catchments, the features of the surface which best allow us to constrain uncertainty in predictions; also to identify, again by sampling, the apparently stochastic variability of responses over catchments. An ensemble prediction for an ungauged ''target'' catchment may then be obtained by sampling from the available models within the part of the parameter attribute space to which the ungauged catchment belongs. The available candidate models are assigned a likelihood that they will represent the target catchment well. This consists of an ''informative'' likelihood prior to considering the nature of the catchment, based upon evidence of its suitability to the gauged catchment on which it was identified, i.e., its local calibration or validation performance; and an updating likelihood which considers the physical similarity of the target and potential donor catchments, according to a prespecified similarity measure. *McIntyre et al.* [2005] noted that the scheme has the potential to integrate uncertainty in the model structures, parameter sets, input data and catchment attributes. Here, however, we focus the investigation on identifying the best similarity measures for estimating the likelihood that a model will adequately represent the target catchment, given a model structure, sets of input data and catchment attributes.

[20] Streamflow at time step t is represented as $Q(t)$. In the case of an ungauged estimation, it is derived thus:

$$Q(t) = \sum_{k=1}^{K} \sum_{i=1}^{M} W_{k,i} h(\theta_{k,i}, X(t)) \tag{1}$$

where $h(\theta_{k,i}, X(t))$ is the output of candidate model k of K from gauged catchment i of M, given the set of model parameters $\theta_{k,i}$ and forcing data for the target catchment, $X(t)$. In this study, forcing data are precipitation and PET data. $W_{k,i}$ is the posterior likelihood assigned to model k from gauged catchment i. All W are scaled such that they sum to one. W (equation (2)) is defined as the product of the prior likelihood, C (equation (4)), and the updating likelihood, B (equation (8)). *McIntyre et al.* [2005] varied the number of candidate models retained from each donor catchment, and showed that it had little effect on predictive performance. For this reason, and for the sake of simplicity and clarity, in this study we choose only one candidate model from each donor catchment, and focus instead on the definition of the similarity metric used to update the likelihood. The number of donor catchments, M, is later optimized (see section 4.2):

$$W_i = \frac{B_i C_i}{\sum_{i=1}^{M} B_i C_i} \tag{2}$$

The streamflow prediction then becomes

$$Q(t) = \sum_{i=1}^{M} W_i h(\theta_i, X(t)) \tag{3}$$

We present three different scenarios: First, as a benchmark, we consider neither prior nor posterior information, instead selecting donor catchments at random; next, we consider posterior information (about catchment similarity) but not prior; and finally we consider both prior and posterior information. This allows for clear discussion of the value of adding both of these extra sources of information.

3.3.1. Suitability of the Calibrated Model

[21] The prior likelihood, C, of a candidate model reflects its suitability prior to considering the nature of the donor catchment to which it will be applied. We can objectively assess this suitability via a measure of the goodness of fit of streamflow estimate produced by that model to the observed streamflow data of its local catchment:

$$C_i = \frac{E_i^{\alpha}}{\sum_{i=1}^{M} E_i^{\alpha}} \tag{4}$$

where E_i is the Nash-Sutcliffe efficiency (equation (5)) of the candidate model following calibration to data from gauged catchment i. Prior to defining equation (4), all E for the donor catchments $i = 1{:}M$ are scaled so that the set of E values covers the range 0 to 1. The variable α accentuates the differences between the better and worse candidate models. The value of α is later optimized (see section 4.2).

[22] The Nash-Sutcliffe efficiency [*Nash and Sutcliffe*, 1970], E, is defined in equation (5):

$$E = 1 - \frac{\sum_{t=1}^{T} [Q(t) - Qobs(t)]^2}{\sum_{t=1}^{T} [Qobs(t) - \mu(Qobs)]^2} \tag{5}$$

$Q(t)$ is the streamflow estimate and $Qobs(t)$ the observed streamflow at time step t of T for the catchment in question. $\mu(Qobs)$ is the mean of the observed streamflow over the T time steps considered. In addition to the Nash-Sutcliffe efficiency, a goodness-of-fit measure VE was used which places more importance on the smaller flows:

$$VE = 1 - \frac{\sum_{t=1}^{T} |Q(t) - Qobs(t)|}{\sum_{t=1}^{T} Qobs(t)} \tag{6}$$

Clearly the choice of goodness-of-fit statistic is up to the modeler, and should reflect the aims of the particular exercise. Remember that we may also choose an uninformative prior likelihood in which case we still use all M calibrated models, but assign all C to be equal.

3.3.2. Physical Similarity to the Target Catchment

[23] We can now update the likelihood of the donor catchment model representing the target catchment well, by considering the physical similarity of the two catch-

ments. We use a variation of the Euclidean distance metric, known as the Minkowski metric (equation (7)) [*Afifi et al.*, 2004; *Jobson*, 1992; *Krzanowski*, 1988]:

$$D_{i,j} = \left[\sum_{a=1}^{N} w_a \left|A_{a,i} - A_{a,j}\right|^{\lambda}\right]^{1/\lambda} \quad (7)$$

where $D_{i,j}$ is the dissimilarity of gauged catchment i to target catchment j, w_a is an importance coefficient of catchment attribute A_a of N, and λ accentuates differences between more and less similar catchments. All attributes, A, used in equation (7) are standardized to have mean zero and standard deviation one. This removes artificial importance due to the units used, and allows the importance of each attribute to be more easily compared via w. The importance coefficients, w and λ are later optimized (see section 4.2). This dissimilarity metric is converted into the likelihood, B, that a model from a particular gauged catchment will represent the target catchment well:

$$B_{i,j} = \frac{1}{\sum_{i=1}^{M}\left[Z - \frac{D_{i,j}}{D\max_j}\right]}\left[Z - \frac{D_{i,j}}{D\max_j}\right] \quad (8)$$

where $Dmax_j$ is the maximum dissimilarity of the M donor catchments to target catchment j. Z can take any value greater than one (to ensure that the least similar of the M donor catchments receives a nonzero likelihood). The results are insensitive to choice of Z, so it was fixed at $Z = 1.1$. A threshold of similarity may be applied, beyond which we consider the gauged catchment to be so dissimilar to the target catchment that it would not be useful as a donor. This can take the form of a threshold value of the dissimilarity, D, or of a limit on M, the number of donor catchments for each target catchment. Gauged catchments falling outside this threshold receive $B = 0$. In this study we will only use limits on M. This has the disadvantage that potentially some unsuitable donors may be included; however it has the advantages of achieving a consistently large ensemble to represent uncertainty and to smooth out the effects of the poorer donors in the averaged result [*McIntyre et al.*, 2005; *Oudin et al.*, 2008].

[24] *McIntyre et al.* [2005] defined D via a measure of catchment dissimilarity from the Institute of Hydrology Flood Estimation Handbook, and highlighted, as a priority for future research, identification of improved similarity measures. This is the focus of the current study, and is described in sections 3.4–4.2.

3.4. Selection of Attributes and Optimization of Similarity Measures

[25] The steps that were undertaken to optimize multi-attribute similarity measures are as follows.

[26] 1. As a first pass to identifying which of the many possible catchment attributes should be included in the optimized similarity measure, model averaging experiments are run using each attribute individually to assess similarity. The usefulness of each is determined by comparing the results to those achieved by selecting donor catchments at random. Selecting donors at random is equivalent to assigning all W (model likelihood) as equal and summing to one for donors $i = 1{:}M$, and $W = 0$ for all donors $i > M$. Here M, the number of donor catchments, is set at 5. Initial proof-of-concept studies show this to be a reasonable choice and to have no effect on the conclusions from this step (note that M is optimized in later steps).

[27] 2. Once we have some (hopefully objective) insight into which attributes should prove useful, the dimensionality of the following optimization step must be reduced, in order to make it viable (i.e., to match the number of degrees of freedom with the amount of information available). Attributes may be chosen based upon evidence of general significance in the previous step and/or on prior knowledge of key attributes. Strongly correlated attributes are not included.

[28] 3. The model averaging variables are optimized to maximize predictive performance for the development group of catchments. The optimization algorithm is the shuffled complex evolution of *Duan et al.* [1992], using the algorithm settings recommended by *Duan et al.* [1994]. This optimization is done in two parts. First, an uninformative prior likelihood for all candidate models is used, such that W (the posterior likelihood) is equal to B (the updating likelihood, based on catchment similarity). Next, informative prior likelihood is also introduced. This allows conclusions as to the value of including prior information in the scheme.

[29] 4. The optimized similarity measures are then used to estimate streamflow for 89 independent catchments. Estimates are compared with those from a traditional regression approach and one based on spatial proximity. These methods are described in section 4.3.

[30] The variables that require optimization in step 3 are the importance coefficients, w_a for $a = 1{:}N$; the number of donor catchments M; the similarity accentuation exponent, λ; and the prior likelihood accentuation exponent, α (for informative prior likelihood only).

[31] It should be noted that this is a very large optimization problem, which in some ways restricts the analysis. Each function call generates a set of model averaging variables, similarity is assessed and the ensemble of donor models is run for each of the 95 development catchments. Due to the interaction of the variables, the response of the objective function is very nonsmooth, necessitating the use of thorough, but slow, global optimization algorithms. Of course, it would be ideal to design the experiment somewhat differently, by optimizing the variables using each combination of 183 catchments, and testing on the one independent catchment. This would give us 184 optimized similarity measures rather than 1. However, it is prohibitively computationally expensive to conduct the experiment in this way.

4. Results and Discussion

4.1. Similarity Using Individual Attributes

[32] As a first pass at identifying which of the many possible catchment attributes should be included in the optimized similarity measure, model averaging experiments are run using each attribute individually to assess similarity. For each attribute, therefore, $N = 1$ and $w_1 = 1$ in equation (7).

[33] Figure 2 shows the model averaging results for each region (Australia, Cfa and Cfb) using individual attributes

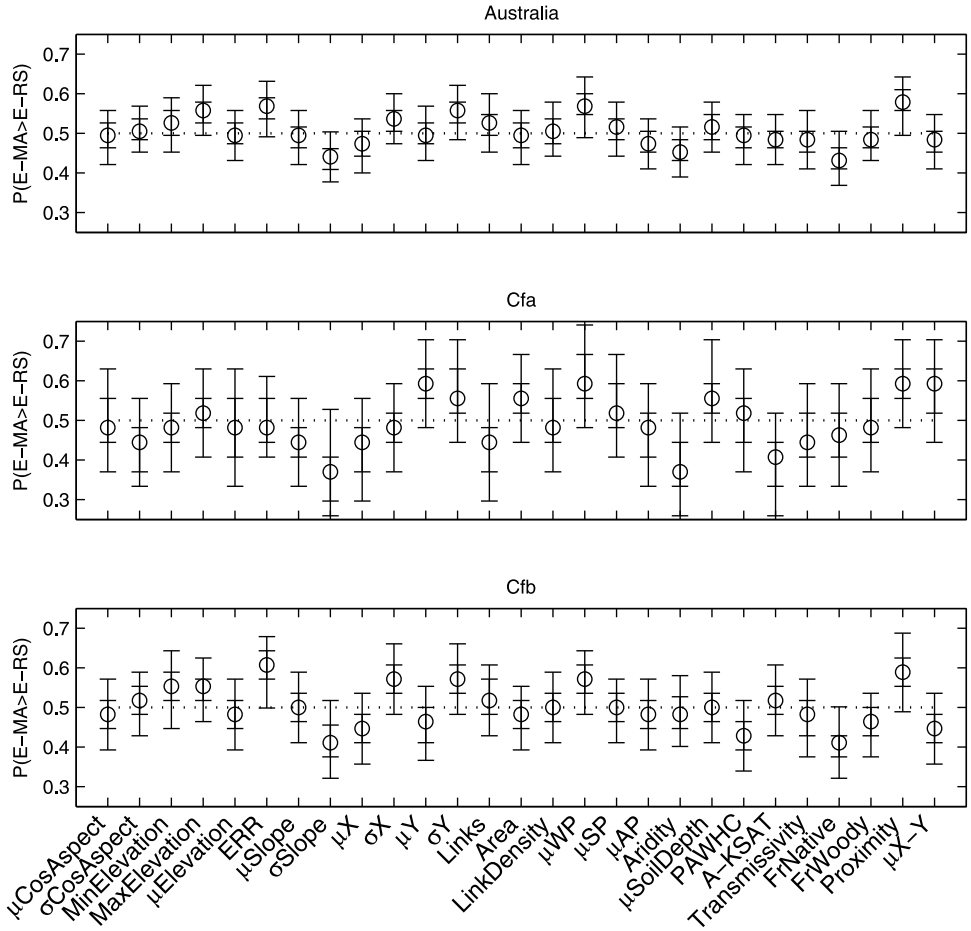

Figure 2. Performance of the model averaging procedure when using individual attributes as measures of similarity, presented in terms of the probability that the model-averaging (MA) E value will exceed that when donor catchments are selected randomly (RS). The circle represents the 50th percentile, and the bars represent the 5th, 25th, 75th, and 95th percentiles. See section 4.1 for an explanation of how these attributes are derived.

as similarity measures. It is presented in comparison with random selection of donors, using the following procedure: the probability that, for any given target catchment, the model averaging (MA) using individual attributes as similarity measures will be better than when selecting donors at random (RS) is denoted P(E − MA > E − RS). This is estimated for the group of 95 catchments. The 90% confidence limits around this probability are estimated using 1000 repetitions of random donor selection. We can see from Figure 2 that the 90% confidence limits of our estimates of P(E − MA > E − RS) encompass 0.5 for each individual attribute. Figure 2 is presented in terms of the Nash-Sutcliffe efficiency, E, however the conclusions for VE are the same.

[34] Figure 2 shows some interesting results. Various attributes are clearly useful similarity measures. Most notably, terrain properties (MaxElevation, ERR, σX and σY) and winter precipitation consistently provide better predictions than the random models, although not always better. However, there are also some attributes which give consistently poorer predictions that the random models, notably aridity, fraction of native forest and variability of slope. In fact, approximately half of the attributes do worse in general than the random model. As there are no physically based reasons why more similar catchments should consistently produce less similar responses, this is presumed to be due to the sparse sampling of the 27-dimensional attribute space provided by the 95 development catchments. For example, the catchments which have similar fractions of native forest in our sample happen to be dissimilar in more important respects. A much larger coverage of the attribute space would be needed to begin to address this issue. Therefore, the evidence in Figure 2 does not allow us conclude that the apparently important attributes (MaxElevation, ERR, σX and σY and winter precipitation) would be important outside of our development sample; rather they are likely to be an artifact of sampling error. There is also the possibility that some of the apparently less important attributes are important for specific catchment types, and therefore their value can only be seen when combined with other attributes. The next stage of analysis looks at combining attributes in search of an optimally powerful similarity measure.

[35] The recognition that step 1 has not lead to a safe list of attributes means that attributes to be used in step 3 are instead chosen based on experience and intuition. They should represent if possible geomorphic, vegetation, climatic and soil attribute types, and should have low correlation

Table 3. Correlation of Catchment Attributes[a]

	μ_z	μ_β	Area	μ_{AP}	Ar	Soil	F_{TWV}	Seasonality	Latitude	Longitude
μ_z	1	0.31	−0.00	−0.04	0.02	0.03	0.15	−0.09	−0.04	0.27
μ_β		1	0.04	**0.52**	**0.57**	0.48	**0.70**	−0.04	−0.11	0.35
Area			1	−0.13	−0.14	−0.00	0.01	0.10	−0.00	−0.12
μ_{AP}				1	**0.97**	**0.57**	**0.52**	−0.32	0.37	0.32
AR					1	**0.55**	**0.54**	−0.23	0.21	0.34
Soil						1	0.46	−0.06	0.05	0.26
F_{TWV}							1	−0.10	0.06	0.13
Seasonality								1	−0.49	−0.42
Latitude									1	0.09
Longitude										1

[a]"Strong" correlation coefficients (>0.5) are in bold. Seasonality is MWP/MAP (the fraction of annual precipitation falling in the lower-PET half of the year).

with one another. It is judged by the authors that the attributes in Table 3 should be useful in this context. Table 3 shows the correlation between attributes; strong correlation coefficients (greater than 0.5) are in shown in bold. After removing strong correlations, the list is reduced to seasonality, mean elevation, mean slope, area, soil depth, latitude and longitude. Each attribute is standardized to have a mean of zero and standard deviation of one. Note that seasonality was not used in the original analysis: it is used from this point on to represent climate, as the other climatic attributes were correlated to several other important attributes, whereas seasonality is not. Seasonality is defined as mean winter precipitation divided by mean annual precipitation, or the fraction of precipitation falling in the lower-PET half of the year.

4.2. Optimization of Multiattribute Similarity Metrics

[36] Using the attributes from section 4.1, the multiattribute similarity measure D (equation (7)) is defined. The variables within the similarity measure and weighting method are now optimized: the attribute importance weights, $w_{1:7}$, and the dissimilarity accentuation exponent, λ (equation (7)), the number of donors, M (equations (1) and (3)) and the local calibration accentuation exponent α (equation (4)). The results are insensitive to Z (equation (8)), so this is fixed at $Z = 1.1$ (in this form of the equation Z can take any value greater than 1). These optimizations show good repeatability, giving us confidence that they terminate in the region of global optimality. The optimization results are discussed in the sections 4.2.1 and 4.2.2.

4.2.1. Attributes' Importance and Sensitivity

[37] Table 4 presents the values of the optimized variables for each region and for the two objective functions. There are two versions of each, the first using an uninformative prior, and the second using an informative prior.

[38] Interpreting the relative importance of each attribute from Table 4 is problematic. The attributes are standardized to have mean zero and standard deviation one. While this removes artificial importance due to units it accounts incompletely for kurtosis, and not at all for skewness. A highly kurtic distribution (leptokurtic) is narrow and tall about the mean, whereas its opposite, a platykurtic distribution is flat and wide about the mean. If, in calculating a distance metric within a group of catchments, we wish for the same influence to be imparted from an attribute with a leptokurtic distribution as from one with a platykurtic distribution, we require a higher weight to be given to the leptokurtic attribute. This potentially clouds our interpretation of the importance of individual attributes. Figure 3 shows frequency distributions of the seven standardized attributes for the development catchments, and their kurtoses. The distributions for the test catchments are very similar. When the kurtoses are plotted against the average values of the importance coefficients for the Australia region (these are the average of four values: two for each objective function), there is a positive correlation, which

Table 4. Optimized Variables for Australia, Cfa, nd Cfb[a]

	w_1 Seasonality	w_2 μ_z	w_3 μ_β	w_4 Area	w_5 Soil	w_6 Latitude	w_7 Longitude	λ	M	α
Aus *E*, no prior	0.21	0.37	0.25	0.87	0.32	0.02	0.82	8.61	9	-
Aus *E*, prior	0.17	0.52	0.09	0.77	0.30	0.37	0.42	2.51	13	3.22
Aus *VE*, no prior	0.34	0.27	0.84	0.44	0.27	0.28	0.85	1.07	10	-
Aus *VE*, prior	0.88	0.42	0.50	0.24	0.50	0.90	0.46	1.01	5	0.3
Cfa *E*, no prior	0.60	0.01	0.97	0.44	0.27	0.14	0.93	2.01	5	-
Cfa *E*, prior	0.62	0.06	0.92	0.23	0.65	0.78	0.47	3.2	3	2.2
Cfa *VE*, no prior	0.96	0.00	0.10	0.05	0.24	0.24	0.93	4.78	5	-
Cfa *VE*, prior	0.47	0.17	0.74	0.32	0.96	0.45	0.52	1.37	6	0.1
Cfb *E*, no prior	0.87	0.71	0.04	0.07	0.79	0.19	0.04	5.52	15	-
Cfb *E*, prior	0.64	0.46	0.08	0.45	0.69	0.02	0.29	3.2	9	4.8
Cfb *VE*, no prior	0.95	0.47	0.59	0.06	0.87	0.78	0.30	2.15	4	-
Cfb *VE*, prior	0.67	0.88	0.42	0.08	0.36	0.89	0.69	1.24	4	0.2

[a]Variables use two objective functions: median Nash-Sutcliffe efficiency (*E*, equation (5)) and median monthly volume error (*VE*, equation (6)).

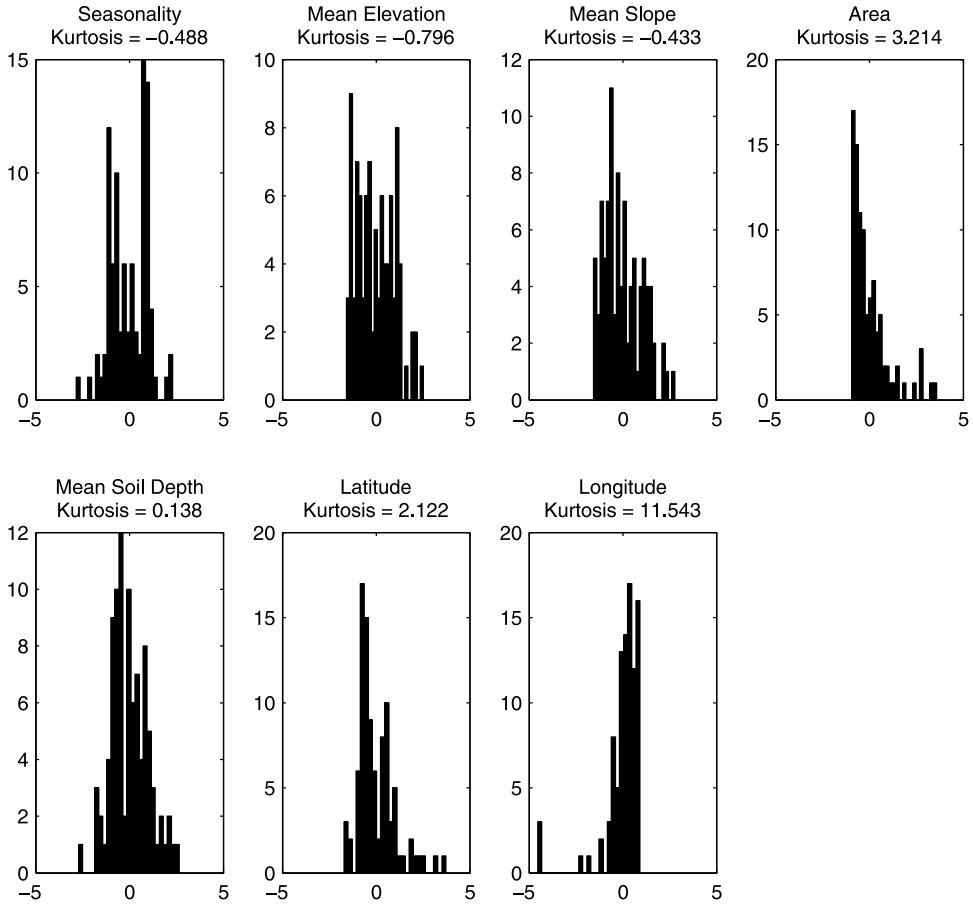

Figure 3. Frequency distributions and kurtoses for the development catchment group of the seven standardized catchment attributes used in the optimization of multiattribute similarity measures.

indicates that this is a confounding factor. We can instead look at interpreting importance a different way, by setting $w = 0$ for each attribute in turn, and judging importance by the deterioration of the performances.

[39] Figure 4 shows the percentage deterioration in the median of each set of objective functions, for both development and test groups, and uninformative and informative priors, when w is set to zero for each attribute in turn. Since these importance coefficients are optimized, it is expected that when they are set to zero the median result for the development catchments should decrease, and this is seen in Figure 4. However, it is immediately obvious from Figure 4 that for the test catchment group, this often results in an improvement in the median result (shown as a negative deterioration in Figure 4). There is no discernable difference in this improvement when using uninformative or informative prior likelihoods. One explanation for this is that the optimized variables are not truly "optimal" but are dependent on the development catchments. This is not surprising given that what is effectively a seven-dimensional attribute space is represented by only 95 and 89 samples of the development and test catchments, respectively. However, it is appropriate at this point to highlight another possible explanation for this result. Given that our system is extremely complex, that our representation is coarse and simplistic and that it relies heavily upon often erroneous and averaged indicators, our ability to predict behavior is necessarily limited [*Koutsoyiannis*, 2009]. Whichever the correct explanation, though, this result does not mean that the approach is not useful, as will be demonstrated in section 4.3.

4.2.2. Sensitivity to Averaging Variables

[40] To understand the interactions of the averaging variables some sensitivity analyses are presented. Figures 5a and 5b show the sensitivity of the median objective function to changing α and λ when using informative priors. Figures 5c and 5d show λ under the special case of uninformative priors ($\alpha = 0$). The *E*-optimized similarity measure with informative prior likelihood, Figure 5a, is sensitive to λ up until around $\lambda = 10$, after which it is less insensitive. E is sensitive to α over the whole range tested, $0 > \alpha > 20$, with a distinct region of optimality around $\lambda = 2.5$ and $\alpha = 3$. The *VE*-optimized similarity measure, Figure 5b shows the same behavior; sensitive to λ up until around $\lambda = 8$, but sensitive to α in the whole of the range tested. This sensitivity to λ is repeated in Figures 5c and 5d. Note that the small ranges of values on the y axes serve to amplify the noise in the trends. The sensitivity to λ and α shows that there is useful information in both the prior and updating likelihoods.

[41] There is an obvious difference, when looking at Table 4 and Figures 5a and 5b, between the prior likelihood accentuation exponent, α, for the different objective functions. The *E*-optimized α is consistently greater than one

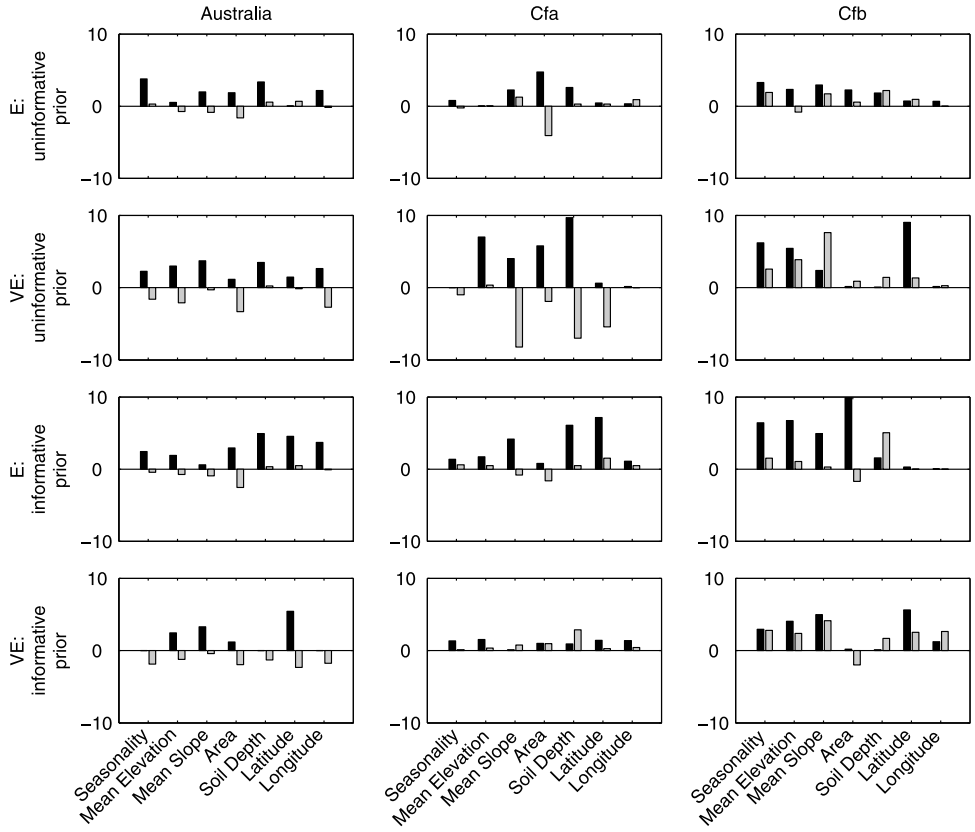

Figure 4. Deterioration in median prediction result when setting $w = 0$ (w is the importance coefficient for each catchment attribute) for each attribute, one at a time. Development catchments are shown in black, and test catchments are shown in gray.

(3.2 for the Australia region); the *VE*-optimized α is consistently smaller than one (0.3 for the Australia region). The reason for this is again due to the distributions of calibration *VE* and *E* values. The *E* and *VE* values have a coefficient of variation (CV) of 0.14 and 0.4, respectively. Therefore, to achieve the same influence α must also be significantly larger for *E* than for *VE*.

[42] Figure 6 shows the change in median objective function result when changing the number of donors, *M*. For *E* when using the informative prior likelihoods, the median result increases from a minimum at $M = 1$ to a maximum at $M = 13$; for *VE* the same applies with a maximum at $M = 5$. The general result from Figure 6, that it is beneficial to use more than one donor model, is consistent with *McIntyre et al.* [2005], who found that generally the best number of donors was 10 when using a simple PDM model; and with *Oudin et al.* [2008], who preferred 3–4 when using the parsimonious GR4J model, and 7–8 when using a variation of TOPMODEL. *Oudin et al.* [2008, paragraph 42] point out that "choosing only one donor catchment was detrimental for regionalization purposes: a larger number of donor catchments allows avoiding strong errors in streamflow simulation by smoothing the responses with other sources."

4.3. Ungauged Streamflow Prediction

[43] The results of the ungauged streamflow predictions are shown in Figure 7. They are presented in terms of the probability, for any given catchment, of returning a better result than a random selection of donors. The results are presented for model averaging with an uninformative prior likelihood (MA, no prior); and model averaging with an informative prior likelihood (MA, prior). The model averaging results are compared with those using local calibration, those using a traditional regional regression technique, those based on spatial proximity using the single nearest donor (SP1) and spatial proximity using the *M* nearest donors (SP, *M*). The model averaging result (MA, simple) when using equal importance coefficients (w) and unoptimized averaging variables ($\alpha = 0$, $\lambda = 2$ and $M = 5$) is also presented for comparison.

[44] The regression relationships were derived using forward stepwise regression between catchment attributes and model parameters. Attributes were only added to the existing relationship if they reduced the standard error significantly at the 5% level. Separate regressions were developed for the two objective functions, and for Australia, Cfa and Cfb. No detailed cluster analysis or similar analysis was conducted to identify homogeneous regions with the aim of strengthening the regressions; however, since they were developed separately for the three regions, just as the model averaging variables were, it is considered a basis for fair comparison. One arguably unfair point of comparison is that the model averaging variables were optimized to maximize predictive potential for a group of catchments, whereas the regressions were optimized to maximize the fit

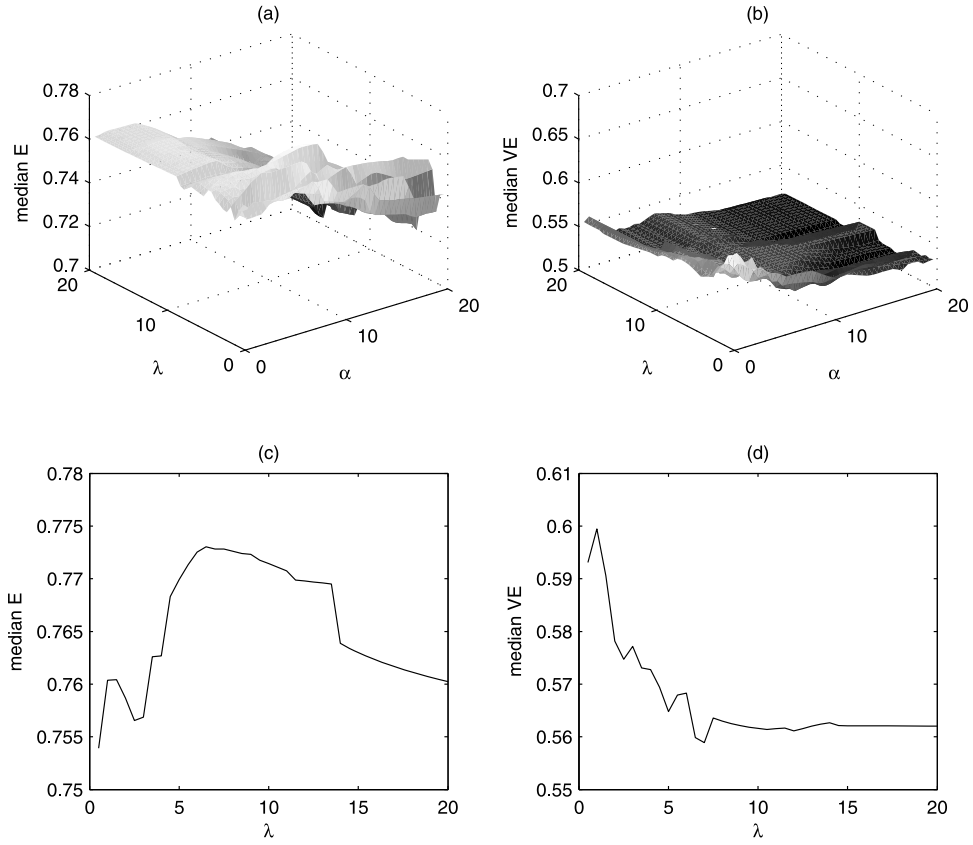

Figure 5. Sensitivity of the median result for the Australian development catchments when varying the prior likelihood accentuation factor, α, and the similarity accentuation factor, λ, (a and b) when using an informative prior likelihood and (c and d) when using an uninformative prior likelihood. (c and d) Variation with λ only. Note that the ranges of the two variables are $0 < \lambda \leq 20$ and $0 \leq \alpha \leq 20$.

of the relationships, as is standard practice. Optimizing the regressions against target catchment performance would provide a better comparison, although would be an additional large optimization effort.

[45] Two spatial proximity variants are presented. The benchmark is the use of the calibrated model from the single nearest gauged catchment. In addition to this, a weighted ensemble of predictions from the *M* nearest gauged catchments is considered. Here *M* takes the optimized value for the appropriate region and objective function (optimized using informative prior likelihoods, see Table 4). This will give a clear indication of the benefits of using the extra information contained in the catchment attributes. The methodology and equations for this variant are exactly the same as for model averaging (equations (2), (6), and (7)), but with only two attributes (latitude and longitude), $\lambda = 2$, and with an uninformative prior likelihood.

[46] All regionalization methods presented deteriorate significantly in quality from local calibration. This is an important point to note, and was introduced in section 1: We cannot hope to perfectly represent complex catchment processes using even locally calibrated models when utilizing temporally and spatially lumped, often seriously erroneous data. Accepting this, the task of inferring catchment behavior based on surrogate data, which is therefore one step less adequate, becomes daunting. However, as the goal is an important one, and positive attempts toward it continue to appear in the literature, it is certainly worthy of further effort. The model averaging approaches represent an improvement upon past methods.

[47] The model averaging methods generally outperform the spatial proximity approaches. This shows the benefit of using the extra information included in the multiattribute catchment similarity measures. Of the two model averaging methods, that with an informative prior likelihood is always better than with an uninformative prior likelihood for the test catchments, with the exception of *VE* for Australia where there is minimal difference. This illustrates that the informative prior likelihood is indeed useful. The unoptimized model averaging (MA, simple) performs consistently worse than the optimized model averaging approaches, showing the value of our new approach. As expected, all model averaging approaches outperform the regression approach overall, although there are exceptions. Of the two spatial proximity methods shown, that using the *M* nearest donors (SP, *M*) is clearly superior to that using the single nearest donor (SP1). Indeed, the latter method consistently does worse than choosing a random donor catchment. In general, less variable results are achieved for Australia than for the Cfa and Cfb subsets. This is thought to be due to the reduced data available for the smaller regions for optimization.

[48] The ungauged prediction results are also presented in Figure 8. These show *E* and *VE* for each regionalization method against local calibration results for the test catchment group. All methods are clearly inferior to local

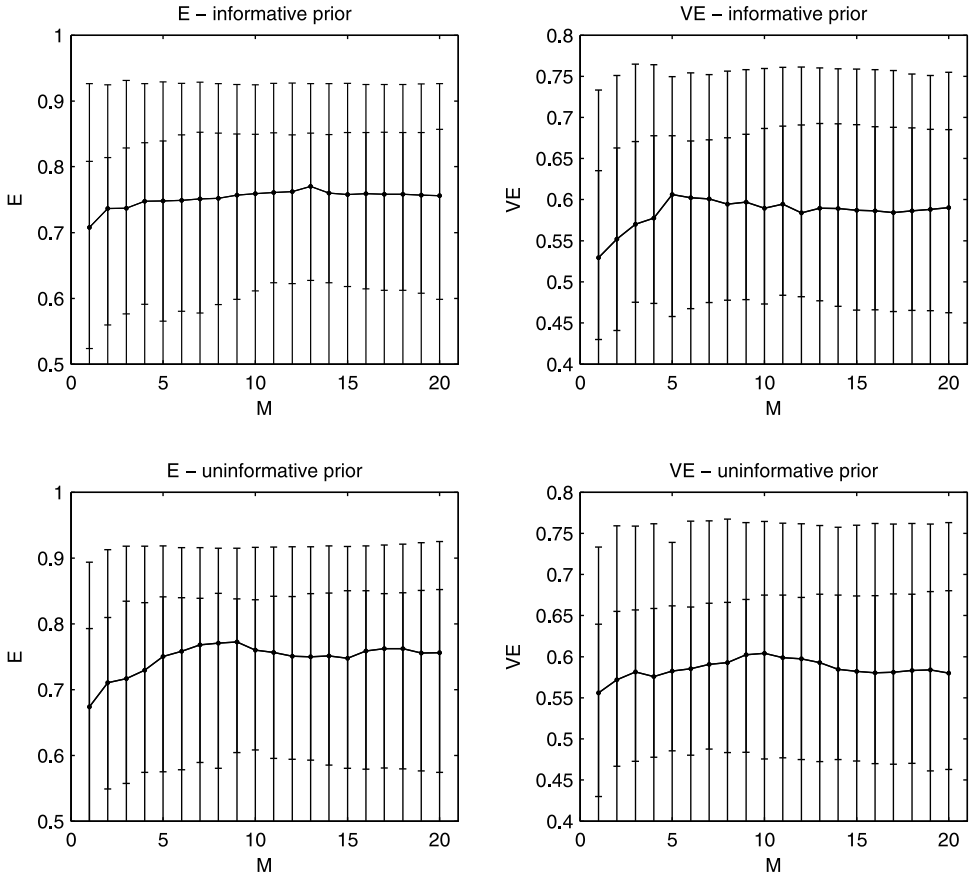

Figure 6. Sensitivity of the results for the Australian development catchments to change in the number of donor catchments. The lines show the median result, and the bars show the 5th, 25th, 75th, and 95th percentiles (note that the axes are scaled such that the 5th percentile does not appear).

calibration, indicating room for significant improvement in ungauged streamflow estimation. Figure 8 shows, generally, better streamflow estimates with fewer very poor results for the optimized model averaging methods than for the other methods presented.

[49] The poor performance of spatial proximity contrasts with previous reports. *Merz and Blöschl* [2004] found spatial proximity using a single donor to perform better than methods based on catchment attributes, although used a far more spatially concentrated group of catchments; *Oudin et al.* [2008] found spatial proximity (with multiple donors) to be clearly better than model averaging (without an optimized similarity measure), which was in turn better than regression, again using a far denser set of catchments. However, another key difference is the optimization, and it may be that, if both studies were placed within an optimization framework, the extra information contained in the similarity measures would improve them above those of spatial proximity and regression.

5. Conclusions

[50] This paper describes improvements upon a recently developed approach to the problem of estimating ungauged streamflow. The model averaging method selects a number of candidate models from available gauged catchments, and weights them based on the likelihood that they will represent the target ungauged catchment well. A candidate model's prior likelihood is based on local calibration performance or, alternatively, an uninformative prior can be used. This is then updated by considering the physical similarity of the potential donor catchment and the target catchment. A streamflow estimate for the target ungauged catchment is achieved by combining the weighted outputs of all models with nonzero posterior likelihood, forced with the target catchment's data. Previous applications of this method [*McIntyre et al.*, 2005; *Oudin et al.*, 2008] have highlighted the need to develop a context-specific catchment hydrologic similarity measure for calculation of posterior likelihoods. The key contribution of this paper is describing the objective optimization of catchment hydrologic similarity measures within the model averaging framework. As a case study, monthly flows were simulated using the SimHyd model for 184 gauged catchments in Australia. 95 catchments were used for optimization and 89 for testing. Performance was assessed using Nash Sutcliffe efficiency and relative volume balance error.

[51] First, experiments using individual catchment attributes as similarity measures were run, with the aim of informing selection of attributes for combination in multi-attribute similarity measures. This experiment showed that, while some attributes were successful as similarity measures in terms of the achieved flow prediction performance, an equal number had a counterintuitive negative influence on performance. As there is no physical reason why more similar catchments should produce less similar responses,

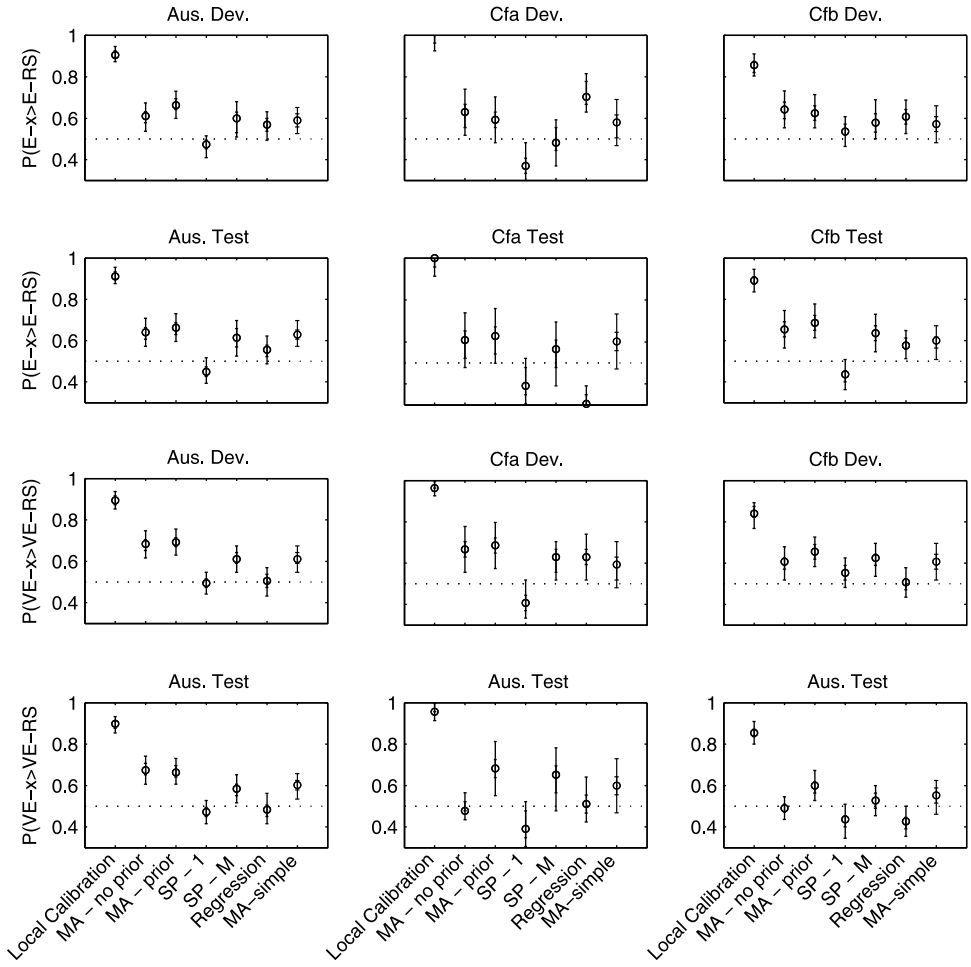

Figure 7. Ungauged prediction results in terms of the probability that the method will do better than when donor catchments are selected at random (represented as P($E - x > E -$ RS) or P($VE - x > VE -$ RS), where x represents the method used to estimate streamflow). The markers represent the median, and the bars represent the 5th, 25th, 75th, and 95th percentiles. See section 4.1 for an explanation of how these attributes are derived.

this indicates that the sample of 95 development catchments does not allow us to confidently select the generally important attributes; rather, any apparently important attributes may appear to be important within this small sample by chance. While even an infinitely large sample will still not truly represent hydrological response due to the various other sources of uncertainty [*Kavetski et al.*, 2002, 2006a, 2006b], it is likely that a larger sample than that used here will give greater insight into important individual attributes. More gauged catchments would also provide a better sample of the apparently stochastic component to the variability. While more gauged catchments are available to add to the set, this would be at the expense of losing the consistent simulation periods used, hence risking more variability due to climate. It is speculated that a larger number of smaller, more homogeneous catchments in the data set would lead to better regionalization performance: the averaging of some attributes over the more heterogeneous catchments leads to a loss of information about the effects of these attributes. Another factor which may have contributed to the unproven importance of individual attributes was that the importance was measured on the basis of monthly flow performance. Using daily flows would allow more information to be retrieved about differences in response between catchments and, potentially, would increase the number of attributes shown to contribute positively to performance. It is also possible that more information would be retrieved using additional performance criteria.

[52] Subsequent to these results, seasonality, mean elevation, mean slope, area, mean soil depth, latitude and longitude were combined in a dissimilarity metric. Several variables were optimized to maximize predictive potential of the method for the 95 development catchments, and also separately for two Köppen climate regions. The variables optimized were the seven attribute importance coefficients, the number of donor catchments, the local calibration (prior likelihood) accentuation exponent (when using an informative prior likelihood only) and the dissimilarity (updating likelihood) accentuation exponent. The flow predictions from the resultant similarity measures were compared with those achieved using other regionalization methods: the same similarity method but without optimizing the attribute weights or other averaging variables; two spatial proximity variants (one with a single donor and one with multiple donors), and a regression approach. Of these, the optimized model averaging approaches were clearly superior. The

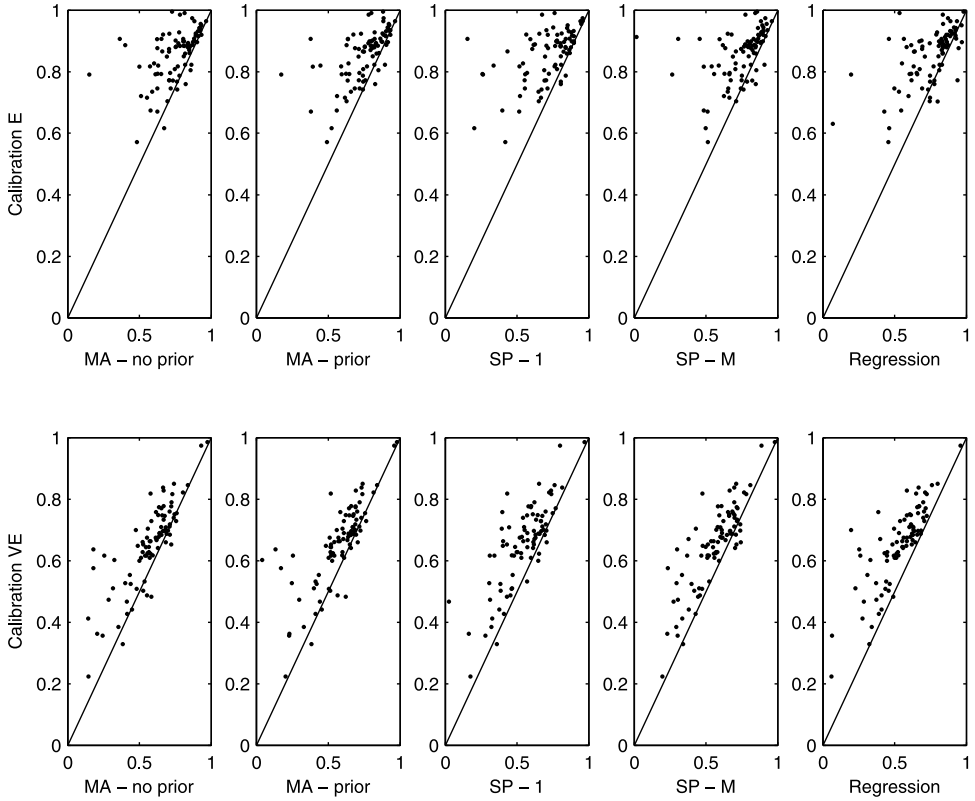

Figure 8. Comparisons between local calibration efficiencies and the regionalization methods in the test catchments: (top) Nash-Sutcliffe efficiencies (*E*) and (bottom) relative volume error (*VE*). Negative values are not shown.

spatial proximity approach with multiple donors was superior to regression, while using the single most spatially proximate donor was worse than random donor selection.

[53] These results are a clear improvement upon those of *McIntyre et al.* [2005], who found that the averaging method was not as clearly superior to the regression approach as demonstrated here, and confirm the view of *McIntyre et al.* [2005] and *Oudin et al.* [2008] that multiple donors should be considered rather than just one. Furthermore, results showed that using prior likelihoods, based on the relative model performances on the donor catchments, was generally superior to assuming equal prior weights, and that there is significant benefit in objectively optimizing the similarity measures. However, the optimized coefficients were not robust estimates, in that they showed some dependence on the development catchment data.

[54] Despite clear improvements upon previous methods and upon the results of *McIntyre et al.* [2005] and *Oudin et al.* [2008], the resultant streamflow estimates are significantly less efficient than those using locally calibrated models, even when these are in validation mode. This is consistent with previous studies. As discussed, the representation of complex, highly spatially and temporally variable processes with often erroneous, spatially and temporally lumped data will necessarily be significantly less than perfect. This makes the task of modeling streamflow with locally calibrated models a challenge, and that of modeling streamflow using surrogates for local data an even greater challenge. However, there has been continuous improvement in our skill in this area over the past several decades, and given the importance of the task it remains a worthwhile pursuit. It is likely that improved representation of the important physical information that we use as surrogates for streamflow will provide the greatest gains in this area.

Notation

D	catchment dissimilarity.
w	importance weights for catchment attributes.
A	catchment attribute.
N	number of attributes.
M	number of donor catchments.
λ	similarity accentuation exponent.
α	prior likelihood accentuation exponent.
W	posterior likelihood.
C	prior likelihood.
B	updating likelihood.
$Q(t)$	streamflow estimate at time step t.
h	model output.
θ	set of model parameters.
$X(t)$	forcing data at time step t.
$Qobs(t)$	recorded response at time step t.
E	Nash-Sutcliffe efficiency.
VE	monthly relative volume error.

[55] **Acknowledgments.** This study was assisted by the eWater CRC (www.ewatercrc.com.au), which provided project and travel funding, and by a University of Melbourne Research Scholarship. The authors appreciate the significant efforts of three anonymous reviewers, Associate Editor

Demetris Koutsoyiannis, and Editor Scott Tyler, whose comments were very helpful in improving the paper.

References

Afifi, A., V. A. Clark, and S. May (2004), *Computer-Aided Multivariate Analysis*, 4th ed., Chapman and Hall, Boca Raton, Fla.

Bastola, S., H. Ishidaira, and K. Takeuchi (2008), Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving TOPMODEL and basins across the globe, *J. Hydrol.*, *357*, 188–206, doi:10.1016/j.jhydrol.2008.05.007.

Beven, K. J. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36, doi:10.1016/j.jhydrol.2005.07.007.

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, 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.

Chiew, F. H. S. and L. Siriwardena (2005), Estimation of SIMHYD parameter values for application in ungauged catchments, in *MODSIM 2005 International Congress on Modeling and Simulation*, edited by A. Zerger and R. M. Argent, pp. 2883–2889, Modell. and Simul. Soc. of Aust. and N. Z., Canberra.

Chiew, F. H. S., M. C. Peel, and A. W. Western (2002), Application and testing of the simple rainfall-runoff model SIMHYD, in *Mathematical Models of Small Watershed Hydrology and Applications*, edited by V. P. Singh and D. K. Frevert, pp. 335–367, Water Resour. Publ., Highlands Ranch, Colo.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031, doi:10.1029/91WR02985.

Duan, Q., S. Sorooshian, and V. Gupta (1994), Optimal use of the SCE-UA global optimisation method for calibrating watershed models, *J. Hydrol.*, *158*, 265–284, doi:10.1016/0022-1694(94)90057-4.

Fernandez, W., R. M. Vogel, and A. Sankarasubramanian (2000), Regional calibration of a watershed model, *Hydrol. Sci. J.*, *45*(5), 689–707.

Gupta, H. V., and S. Sorooshian (1983), Uniqueness and observability of conceptual rainfall-runoff model parameters: The percolation process examined, *Water Resour. Res.*, *19*(1), 269–276, doi:10.1029/WR019i001p00269.

Hundecha, Y., and A. Bardossy (2004), Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalisation of a watershed model, *J. Hydrol.*, *292*, 281–295, doi:10.1016/j.jhydrol.2004.01.002.

Hundecha, Y., T. B. M. J. Ouarda, and A. Bardossy (2008), Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the "spatial" structures of the parameters within a canonical physiographic-climatic space, *Water Resour. Res.*, *44*, W01427, doi:10.1029/2006WR005439.

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*(8), 2637–2649, doi:10.1029/93WR00877.

Jobson, J. D. (1992), *Applied Multivariate Data Analysis*, vol. 2, *Categorical and Multivariate Methods*, Springer, New York.

Kavetski, D., S. W. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D., G. Kuczera, and S. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.

Koutsoyiannis, D.(2009), A random walk on water, *Geophys. Res. Abstr.*, *11*, 14033.

Krzanowski, W. J. (1988), *Principles of Multivariate Analysis—A User's Perspective*, Oxford Univ. Press, New York.

Laaha, G., and G. Blöschl (2006), A comparison of low flow regionalisation methods—Catchment grouping, *J. Hydrol.*, *323*, 193–214, doi:10.1016/j.jhydrol.2005.09.001.

Le Lay, M., S. Galle, G. M. Saulnier, and I. Braud (2007), Exploring the relationship between hydroclimatic stability and rainfall-runoff model parameter stability: A case study in West Africa, *Water Resour. Res.*, *43*, W07420, doi:10.1029/2006WR005257.

Lowe, L., and R. Nathan (2006), Use of similarity criteria for transposing gauged streamflows to ungauged locations, *Aust. J. Water Resour.*, *10*(2), 161–170.

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

Merz, R., and G. Blöschl (2004), Regionalisation of catchment model parameters, *J. Hydrol.*, *287*, 95–123, doi:10.1016/j.jhydrol.2003.09.028.

Montanari, A. (2005), Large sample behavior of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *41*, W08406, doi:10.1029/2004WR003826.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.

Nathan, R., and T. A. McMahon (1990), Identification of homogeneous regions for the purpose of regionalisation, *J. Hydrol.*, *121*, 217–238, doi:10.1016/0022-1694(90)90233-N.

Oudin, L., V. Andreassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, *Water Resour. Res.*, *44*, W03413, doi:10.1029/2007WR006240.

Parajka, J., R. Merz, and G. Bloschl (2005), A comparison of regionalisation methods for catchment model parameters, *Hydrol. Earth Syst. Sci. Discuss.*, *2*, 509–542.

Peel, M. C., F. H. S. Chiew, A. Western, and T. A. McMahon (2000), Extension of monthly unimpaired streamflow data and regionalisation of parameter values to estimate streamflow in ungauged catchments, Dep. of Environ., Water, Heritage and the Arts, Canberra.

Peel, M. C., B. L. Finlayson, and T. A. McMahon (2007), Updated world map of the Köppen-Geiger climate classification, *Hydrol. Earth Syst. Sci. Discuss.*, *4*, 439–473.

Post, D. A., and A. J. Jakeman (1999), Predicting the streamflow of ungauged catchments in S.E. Australia by regionalising the parameters of a lumped conceptual rainfall-runoff model, *Ecol. Modell.*, *123*, 91–104, doi:10.1016/S0304-3800(99)00125-8.

Sefton, C. E. M., and S. M. Howarth (1998), Relationships between dynamic response characteristics and physical descriptors in England and Wales, *J. Hydrol.*, *211*, 1–16, doi:10.1016/S0022-1694(98)00163-2.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resour. Res.*, *44*, W00B06, doi:10.1029/2008WR006822.

Vogel, R. M. (2005), Regional calibration of watershed models, in *Watershed Models*, edited by V. P. Singh and D. F. Frevert, pp. 549–567, CRC Press, Boca Raton, Fla.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalisation for continuous rainfall-runoff models including uncertainty, *J. Hydrol.*, *320*, 132–154, doi:10.1016/j.jhydrol.2005.07.015.

Wagener, T., M. Sivapalan, P. Troch, and R. Woods (2007), Catchment classification and hydrologic similarity, *Geogr. Compass*, *1*(4), 901–931, doi:10.1111/j.1749-8198.2007.00039.x.

Young, A. R. (2006), Stream flow simulation within UK ungauged catchments using a daily rainfall-runoff model, *J. Hydrol.*, *320*, 155–172, doi:10.1016/j.jhydrol.2005.07.017.

F. H. S. Chiew, CSIRO Land and Water, GPO Box 1666, Canberra, ACT 2601, Australia.

N. R. McIntyre, Department of Civil and Environmental Engineering, Imperial College, London SW7 2AZ, UK.

J. P. C. Reichl and A. W. Western, Department of Civil and Environmental Engineering, University of Melbourne, Melbourne, Vic 3010, Australia. (jpreichl@unimelb.edu.au)