AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2013WR015203

Using multiple donor sites for enhanced flood estimation in ungauged catchments

T. R. Kjeldsen[1], D. A. Jones[2], and D. G. Morris[2]

[1]Department of Architecture and Civil Engineering, University of Bath, Bath, UK, [2]Centre for Ecology and Hydrology, Crowmarsh Gifford, Wallingford, UK

Key Points:
- Statistical framework for optimal transfer of data from multiple donor sites
- New performance measure accounting for at-site sampling variability
- A minimum of six donor sites gives the optimal index flood prediction

Correspondence to:
T. R. Kjeldsen,
t.r.kjeldsen@bath.ac.uk

Citation:
Kjeldsen, T. R., D. A. Jones, and D. G. Morris (2014), Using multiple donor sites for enhanced flood estimation in ungauged catchments, *Water Resour. Res.*, *50*, 6646–6657, doi:10.1002/2013WR015203.

Received 20 DEC 2013
Accepted 22 JUL 2014
Accepted article online 24 JUL 2014
Published online 18 AUG 2014

Abstract A new generalized method is presented enabling the use of multiple donor sites when predicting an index flood variable in an ungauged catchment using a hydrological regression model. The method is developed from the premise of having an index flood prediction with minimum variance, which results in a set of optimal weights assigned to each donor site. In the model framework presented here, the weights are determined by the geographical distance between the centroids of the catchments draining to the subject site and the donor sites. The new method was applied to a case study in the United Kingdom using annual maximum series of peak flow from 602 catchments. Results show that the prediction error of the index flood is reduced by using donor sites until a minimum of six donors have been included, after which no or marginal improvements in prediction accuracy are observed. A comparison of these results is made with a variant of the method where donor sites are selected based on connectivity with the subject site through the river network. The results show that only a marginal improvement is obtained by explicitly considering the network structure over spatial proximity. The evaluation is carried out based on a new performance measure that accounts for the sampling variability of the index flood estimates at each site. Other results compare the benefits obtained by adding relevant catchment descriptors to a simple regression model with those obtained by transferring information from local donor sites.

1. Introduction

Estimating the magnitude of design floods (here defined as the discharge associated with a predefined return period) in ungauged basins is an important practical problem in applied hydrology, and one which has also received considerable attention in the scientific literature; in particular as part of the IAHS Prediction in Ungauged Basin (PUB) decade [*Blöschl et al.*, 2013]. A class of methods which has found favor among practitioners and academics is regional flood frequency analysis, where, using statistical analysis of samples of extreme flood or rainfall data from a geographical region, estimates at ungauged sites are obtained through transfer of data from gauged sites. Numerous techniques for undertaking regional frequency analysis have been reported in the literature [*Cunnane*, 1988; *Blöschl et al.*, 2013]. One method in particular, the index flood method, described by *Dalrymple* [1960] and more recently made popular in the L-moment version described by *Hosking and Wallis* [1997], has been the subject of numerous studies [e.g., *Pearson*, 1991; *Vogel et al.*, 1993; *Parida et al.*, 1998; *Kachroo et al.*, 2000; *Lim and Voeller*, 2009; *Yang et al.*, 2010; *Salinas et al.*, 2013].

Underpinning the index flood method is a set of assumptions of which the most prominent, but simplistic, is that within a homogeneous region the flood series from different sites are independent and identically distributed apart from a scale parameter, the index flood, often defined as the mean or median annual maximum flood [*Hosking and Wallis*, 1997; *Institute of Hydrology*, 1999]. The estimation of the index flood at ungauged sites is typically implemented using a regression-type model relating the index flood at gauged sites to a set of relevant physiographic, geomorphologic, and climatic catchment descriptors such as catchment area, mean annual rainfall, soil type, urban extent, etc. [*Grover et al.*, 2002]. Estimation of the regression model parameters can be undertaken using, for example, techniques such as ordinary, weighted, or general least squares (OLS, WLS, or GLS) depending on the level of complexity adopted in the model building phase [*Stedinger and Tasker*, 1985]. Alternatively, the model parameters can be estimated using maximum-likelihood [*Kjeldsen and Jones*, 2009; *Mediero and Kjeldsen*, 2014] or Bayesian [e.g., *Reis et al.*, 2005; *Haddad et al.*, 2012] techniques.

The standard errors of estimates obtained at ungauged sites using this type of regression model are generally relatively large. For example, the regression model developed by *Kjeldsen and Jones* [2009] for estimating the

median annual flood at ungauged sites in the United Kingdom was based on annual maximum peak flow data from 602 gauged catchments and has a factorial standard error (fse) of 1.431. Another example is the study by *Meigh et al*. [1997] developing regression models linking the mean annual flood (MAF) to catchment descriptors in regions from around the world. They reported values of fse in the range between 1.36 (24 basins in South Korea larger than 1000 km^2) and 2.88 (162 arid and semiarid basins worldwide). In recognition of this high level of uncertainty, the guidance provided in the United Kingdom on flood frequency estimation in ungauged catchments [*Institute of Hydrology*, 1999] suggests that pure regression-based estimates should be adjusted, where possible, through data transfer from hydrologically similar gauged donor catchments. This strategy is in keeping with similar conclusions drawn by other researchers. For example, *Merz and Blöschl* [2008] highlighted the benefit of incorporating local knowledge and data into flood frequency analysis. Conceptually, the use of local data to adjust the regression-based estimate at a particular site can be viewed as an attempt to compensate for exclusion of local flood controlling factors in the explanatory variables used in the regression model [*Kjeldsen and Jones*, 2010]. The rules provided by *Institute of Hydrology* [1999] for selection of suitable donor catchments were heuristic and included sites that were considered hydrologically similar in terms of catchment area, mean annual rainfall, and soil type. Assessing the benefit of data transfer in the United Kingdom, *Kjeldsen and Jones* [2010] found that superior performance was achieved when donor sites were selected based on geographical proximity rather than hydrological similarity as measured by catchment descriptors. A similar conclusion was reached by *Merz and Blöschl* [2005] and *Viglione et al*. [2013]. In a separate study involving spatial generalization of flood statistics, *Morris* [2003] found that where donor sites were upstream or downstream of the subject site, utilizing their location on the river network relative to the subject site (primarily determined by similarity of catchment area) could significantly enhance the performance of the data transfer methodology. Other researchers [*Skøien et al*., 2006; *Guse et al*., 2009; *Ganora et al*., 2013] have also reported benefits when including river network geometry in flood frequency regionalization studies.

By studying the error structure of a hydrological regression model, *Kjeldsen and Jones* [2009, 2010] developed an optimal procedure for transferring data from a single gauged site to an ungauged subject site utilizing a functional relationship between the spatial correlation of regression model errors and the geographical distance between catchment centroids. Here, optimality is defined as minimizing the prediction variance of the adjusted estimates of the index flood at the ungauged sites. *Kjeldsen and Jones* [2010] found that the prediction error obtained when using estimates derived from the nonoptimal data transfer procedure presented by the *Institute of Hydrology* [1999] in the Flood Estimation Handbook (FEH) is about twice as large as the optimal data transfer procedure. The FEH method is to choose "similar" catchments as donors, and these are not necessarily geographically close.

This paper presents a novel and generalized method for adjusting regression-based estimates of the index flood at ungauged sites in the United Kingdom using data transfer from multiple gauged donor sites. The analytical results are also extended to develop a new performance metric for the evaluation of regional models that filters out the effect of at-site sampling noise. Based on a case study from the United Kingdom considering an existing model for predicting an index flood in ungauged catchments, the benefit of using data transfer from gauged donor catchments is illustrated and quantified.

2. A Framework for Data Transfer From Multiple Sites

The optimal data transfer procedure is intimately related to the structure of the regional regression model used for predicting the index flood based on catchment descriptors only. Thus, before discussing the data transfer procedure, the details of the regression model are described.

2.1. A Regression Model for Predicting the Index Flood at Ungauged Sites

Consider a region where annual maximum series (AMS) of peak flow events are available from $i=1,\ldots,n$ different gauged catchments. The median of each individual AMS is denoted m_i and the corresponding log-transformed value is y_i. Following *Stedinger and Tasker* [1985], the sample estimate of y_i can be written in terms of a regression model as

$$y_i=\mathbf{x}_i^T\theta+\eta_i+\varepsilon_i=\xi_i+\varepsilon_i \quad (1)$$

where subscript i refers to catchment number, ξ_i is the true (but unknown) value of the log-transformed median, ε_i is the sampling error of the log-transformed index flood (y_i) and is assumed to be normally

distributed with mean zero and a covariance matrix $\boldsymbol{\Sigma}_\varepsilon$. The value of individual elements of $\boldsymbol{\Sigma}_\varepsilon$ depends on: (i) record-length, (ii) the assumed distribution of the AMS series and, (iii) for nondiagonal elements, the distance between catchment centroids and length of overlapping record. An example of the sampling covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ of the log-transformed median annual maximum flood can be found in *Kjeldsen and Jones* [2009]. The regression model parameters are denoted by θ, and $\mathbf{x}_i$ is a $(q+1))$ vector of q catchment descriptors for the ith site with a value of one in the first location. The term $\mathbf{x}_i^T\theta+\eta_i$ is then the true value expressed as a linear regression model based on catchment descriptors plus a model error terms due to the inability of a simple regression model to represent complex basin hydrology. The model error, η, is assumed to have zero mean and the elements of the covariance matrix $\boldsymbol{\Sigma}_\eta$ are defined as

$$\boldsymbol{\Sigma}_{\eta,ij}=\mathrm{cov}\left(\eta_i,\eta_j\right)=\begin{cases}\sigma_\eta^2 & i=j\\ \sigma_\eta^2 r_{\eta,ij} & i\neq j\end{cases} \tag{2}$$

where $r_{\eta,ij}$ is the correlation between model errors which *Kjeldsen and Jones* [2009] related to the geographical distance between catchment centroids, d_{ij}, as

$$r_{\eta,ij}=\varphi_1\exp\left[-\varphi_2 d_{ij}\right]+(1-\varphi_1)\exp\left[-\varphi_3 d_{ij}\right] \tag{3}$$

where φ_1, φ_2, and φ_3 are model parameters that must be estimated along with the regression model parameters θ and the model error variance σ_η^2. The model and sampling errors are assumed to be mutually independent but cross correlated within each set, and the spatial correlation of the model errors is a key part of the data transfer scheme. The introduction of correlated model errors is an extension to the GLS model presented by *Stedinger and Tasker* [1985], who assumed negligible or no correlation between model errors. *Kjeldsen and Jones* [2009] provide a more detailed discussion of the difference between the two error types.

In the next section, a new procedure will be developed which allows for the estimates of the index flood obtained using a regional model, such as equation (1), to be moderated using local data from neighboring sites.

2.2. Optimal Data Transfer From Multiple Sites

Suppose that a set of estimated regression model parameters, $\hat{\theta}$, such as those in Table 1, is available which enables the prediction of the log-transformed index flood, y, at any site in the region of interest. The estimate $\hat{y}_s$ is given as

$$\hat{y}_s=\mathbf{x}_s^T\hat{\theta} \tag{4}$$

where the subscript s indicates a specific site, and the estimate in equation(4) therefore constitutes the regression-only estimate at a specific site. Subsequently, this first estimate should be adjusted by using data transfer from multiple donors. The regression-only estimate at the site of interest, $\hat{y}_s$ from equation (4), is adjusted using residuals from p nearby and gauged donor sites as a weighted average: a weight α (to be determined) is applied to the regression residual $(y_i-\hat{y}_i)$ at each of the p sites. The resulting adjusted estimate is denoted $\tilde{y}_s$ and is derived as

$$\begin{aligned}\tilde{y}_s&=\hat{y}_s+\sum_{i=1}^{p}\alpha_i(y_i-\hat{y}_i)\\ &=\underbrace{\mathbf{x}_s^T\hat{\theta}}_{\hat{y}_s}+\sum_{i=1}^{p}\alpha_i\left(\underbrace{\mathbf{x}_i^T\theta+\eta_i+\varepsilon_i}_{y_i}-\underbrace{\mathbf{x}_i^T\hat{\theta}}_{\hat{y}_i}\right)\\ &=\mathbf{x}_s^T\hat{\theta}+\sum_{i=1}^{p}\alpha_i\left(\mathbf{x}_i^T\left(\theta-\hat{\theta}\right)+\eta_i+\varepsilon_i\right)\end{aligned} \tag{5}$$

The set of weights α_i is specified so as to ensure the prediction error variance of $\tilde{y}_s$ is as small as possible. Note that there is no requirement for the weights α_i to sum to unity. For example, if only one donor site is used ($p=1$), then the weight should decrease toward zero as the distance between the donor and the subject site increases [*Kjeldsen and Jones*, 2010]. Given that a reasonably large number of gauged sites is used in the estimation of the regression model parameters, a reasonable assumption is $\theta\approx\hat{\theta}$, which reduces the complexity of the equation above to

Table 1. Summary Statistics for Regression Model Linking the Log Median Annual Maximum Peak Flow to Catchment Descriptors [from *Kjeldsen and Jones*, 2009]

Coefficient	Parameter	Standard Error	*t*-Value	p-Value
Intercept	2.1170	0.1172	18.06	0.000
Ln[AREA]	0.8510	0.0114	74.35	0.000
(SAAR/1000)$^{-1}$	−1.8734	0.0968	−19.35	0.000
Ln[FARL]	3.4451	0.2654	12.98	0.000
BFIHOST^2	−3.080	0.1158	−26.60	0.000

$$\tilde{y}_s \approx \mathbf{x}_s^T\hat{\theta}+\sum_{i=1}^{p}\alpha_i(\eta_i+\varepsilon_i). \tag{6}$$

The prediction error of $\tilde{y}_s$ is denoted e_s and will be derived by utilizing the fact that the true value at the subject site, ξ_s, is defined as $\xi_s=\mathbf{x}_s^T\theta+\eta_s$ as discussed previously (equation (1)). Thus, e_s is defined as

$$\begin{aligned} e_s &= \tilde{y}_s-\xi_s \\ &= \underbrace{\mathbf{x}_s^T\hat{\theta}+\sum_{i=1}^{p}\alpha_i(\eta_i+\varepsilon_i)}_{\tilde{y}_s}-\xi_s \\ &= \mathbf{x}_s^T\hat{\theta}+\sum_{i=1}^{p}\alpha_i(\eta_i+\varepsilon_i)-\mathbf{x}_s^T\theta-\eta_s \\ &\approx -\eta_s+\sum_{i=1}^{p}\alpha_i(\eta_i+\varepsilon_i)=-\eta_s+\alpha^T(\eta+\varepsilon) \end{aligned} \tag{7}$$

which, as above, assumes that $\theta\approx\hat{\theta}$. The vector $\boldsymbol{\alpha}$ contains the weights assigned to each of the p donor sites, and similarly η and ε denote vectors of the model and sampling errors. Next, the variance of this prediction error e_s is defined as

$$\begin{aligned} \mathrm{var}\,(e_s) &= \mathrm{var}\left\{-\eta_s+\alpha^T(\eta+\varepsilon)\right\} \\ &= \sigma_\eta^2+\boldsymbol{\alpha}^T\mathrm{var}\,(\eta+\varepsilon)\alpha-2\alpha^T\mathrm{b} \\ &= \sigma_\eta^2+\alpha^T\boldsymbol{\Omega}\alpha-2\mathbf{b}^T\alpha \end{aligned} \tag{8}$$

The $p\times p$ covariance matrix of the combined error terms $(\eta_i+\varepsilon_i)$ is denoted by $\boldsymbol{\Omega}$. The vector containing the covariance between the model error at the subject and each of the donor sites $\mathrm{cov}(\eta_s,\eta_i)$ is denoted by $\mathbf{b}$ where the p elements are obtained directly from equation (3) considering the geometric distance, d_{ij}, between the subject site s and each of the $i=1,\ldots,p$ donor sites. In deriving equation (8), independence between sampling and model errors is assumed. Next, the minimum prediction variance is found through straight-forward differentiation of equation (8) with respect to the weights α as

$$\frac{\partial\mathrm{var}\,(e_s)}{\partial\boldsymbol{\alpha}}=0 \;\Rightarrow\; \alpha^T\left(\boldsymbol{\Omega}^T+\boldsymbol{\Omega}\right)-2\mathbf{b}^T=0 \tag{9}$$

As $\boldsymbol{\Omega}$ is symmetric, the optimal set of weights $\boldsymbol{\alpha}$ can finally be derived by isolating α in equation (9) as

$$\boldsymbol{\alpha}=\boldsymbol{\Omega}^{-1}\mathbf{b} \tag{10}$$

If only one donor site is selected, the solution to equation (10) reduces to the corresponding analytical solution for a single donor presented by *Kjeldsen and Jones* [2010]. Thus, the result obtained in equation (10) constitutes a new and more general framework for including local information into the estimation of the index flood variable at ungauged sites.

Given that the variance of the model errors are typically much larger than the variance of the sampling errors, it is likely that the weights will primarily be determined by the model errors (i.e., how well the regression model describes the data) rather than the sampling errors.

3. A Revised Performance Measure

A common approach for assessing the performance of a regional method is the mean squared error, or root mean squared error (RMSE), based on the sum of squared errors, S. The version, S_C that compares the regional estimate with the sample estimate for each of the n gauged site in the region in turn is

$$S_C=\sum_{s=1}^{n}(\tilde{y}_s-y_s)^2=(\tilde{\boldsymbol{y}}\text{-}\boldsymbol{y})^T(\tilde{\boldsymbol{y}}\text{-}\boldsymbol{y}) \tag{11}$$

where $(\tilde{\boldsymbol{y}}\text{-}\boldsymbol{y})$ is a vector containing the n residuals. However, the at-site sample estimates are themselves only best estimates of the true values, ξ, and thus what would be more interesting to know is the value of

$$S_T = \sum_{s=1}^{n} (\tilde{y}_s - \xi_s)^2 = (\tilde{\mathbf{y}}\text{-}\xi)^T (\tilde{\mathbf{y}}\text{-}\xi) \tag{12}$$

In the following, the relationship between S_C and S_T will be derived, allowing an estimate of S_T rather than S_C to be used for assessing the performance of the data transfer method. As before, the difference between θ and $\hat{\theta}$ is ignored (i.e., $\theta \approx \hat{\theta}$) which is the same as ignoring the sampling error of θ, a reasonable assumption when a large number of sites are included in the regression analysis. This analysis starts from a slightly revised version of the earlier expression in equation (6), where the adjustment to the index flood estimate is represented here as a weighted sum of the residuals from all n sites in the national data set. This leads to equation (6) being replaced by

$$\tilde{y}_s = \mathbf{x}_s^T \hat{\theta} + \sum_{i=1}^{n} \alpha_{si} (y_i - \hat{y}_i) \tag{13}$$

but where the weights α_{si} are mostly zero except for the p sites chosen for the data transfer for the site s. The difference between the donor adjusted estimate and the at-site estimate and true value, respectively, can be expressed by combining equations (1) and (4) with the assumptions listed above. First, consider the difference between the adjusted estimate $(\tilde{y}_s, \text{equation } (13))$ and the at-site sampling value, $(y_s, \text{equation } (1))$:

$$\begin{aligned} \tilde{y}_s - y_s &= \mathbf{x}_s^T \hat{\theta} + \sum_{i=1}^{n} \alpha_i (y_i - \hat{y}_i) - y_s \\ &= \mathbf{x}_s^T \hat{\theta} + \sum_{i=1}^{n} \alpha_i \left(\underbrace{\mathbf{x}_i^T \theta + \eta_i + \varepsilon_i}_{y_i} - \underbrace{\mathbf{x}_i^T \hat{\theta}}_{\hat{y}_i} \right) \\ &\quad \underbrace{- \left(\mathbf{x}_s^T \hat{\theta} + \eta_s + \varepsilon_s \right)}_{y_s} \\ &= \sum_{i=1}^{n} \alpha_i (\eta_i + \varepsilon_i) - (\eta_s + \varepsilon_s) \end{aligned} \tag{14}$$

which in a vector format considering all n sites is given as

$$\hat{y}\text{-}y = (\mathbf{A}\text{-}\mathbf{I})\eta + (\mathbf{A}\text{-}\mathbf{I})\varepsilon \tag{15}$$

where $\mathbf{A}$ is a matrix of the weights α_{si}, $\mathbf{I}$ is an identity matrix, and η and ε are $(n \times 1)$ vectors of model and sampling errors, respectively. Analogous to equation (7), the difference between the adjusted estimate and the true value is derived as

$$\begin{aligned} \tilde{y}_s - \xi_s &= \mathbf{x}_s^T \hat{\theta} + \sum_{i=1}^{n} \alpha_{si} (y_i - \hat{y}_i) - \xi_s \\ &= \mathbf{x}_s^T \hat{\theta} + \sum_{i=1}^{n} \alpha_{si} \left(\underbrace{\mathbf{x}_i^T \theta + \eta_i + \varepsilon_i}_{y_i} - \underbrace{\mathbf{x}_i^T \hat{\theta}}_{\hat{y}_i} \right) \\ &\quad \underbrace{- \left(\mathbf{x}_s^T \hat{\theta} + \eta_s \right)}_{\xi_s} \\ &= \sum_{i=1}^{n} \alpha_i (\eta_i + \varepsilon_i) - (\eta_s) \end{aligned} \tag{16}$$

Again, the corresponding vector notation considering all n sites simultaneously is

$$\hat{y}\text{-}\xi = (\boldsymbol{A}\text{-}\boldsymbol{I})\eta + \boldsymbol{A}\varepsilon \tag{17}$$

Considering the two different definitions of residuals in equations (15) and (16), it can be seen that comparing the predictions of the index flood with the observations directly (equation (14)) will contaminate the residuals with a sampling error, ε_s of the observed index flood. The result is that performance measures such as RMSE will be inflated by this sampling error. As the sampling error depends on the record-length of the at-site record, the RMSE will also become a function of record-length and it is not possible to ascertain how much of the difference between predictions and observations is caused by model deficiency and how

much is down to sampling noise in the at-site observations. However, the framework developed here allows for a revised performance measure to be developed where the influence of the at-site sampling noise is removed.

From the matrix representation of the differences in equations (15) and (17), the sum of squares defined in equations (11) and (12) can now equally be written in matrix form as

$$S_C=\eta^T(\mathbf{A}-\mathbf{I})^T(\mathbf{A}-\mathbf{I})\eta+\varepsilon^T(\mathbf{A}-\mathbf{I})^T(\mathbf{A}-\mathbf{I})\varepsilon -2\eta^T(\mathbf{A}-\mathbf{I})^T(\mathbf{A}-\mathbf{I})\varepsilon \tag{18}$$

$$S_T=\eta^T(\mathbf{A}-\mathbf{I})^T(\mathbf{A}-\mathbf{I})\eta+\varepsilon^T\mathbf{A}^T\mathbf{A}\varepsilon -2\eta^T(\mathbf{A}-\mathbf{I})^T\mathbf{A}\varepsilon \tag{19}$$

By subtracting the two expressions in equations (18) and (19), the following relationship between S_C and S_T is obtained

$$S_C-S_T=2\eta^T(\mathbf{A}-\mathbf{I})^T\varepsilon+\varepsilon^T\left(\mathbf{I}-\mathbf{A}^T-\mathbf{A}\right)\varepsilon \tag{20}$$

Next, the mean value of the difference between the two sums of squares is

$$\begin{aligned}\mathrm{E}\{S_C-S_T\}&=\mathrm{E}\left\{2\eta^T(\mathbf{A}-\mathbf{I})^T\varepsilon+\varepsilon^T\left(\mathbf{I}-\mathbf{A}^T-\mathbf{A}\right)\varepsilon\right\}\\&=0+E\left\{\varepsilon^T\left(\mathbf{I}-\mathbf{A}^T-\mathbf{A}\right)\varepsilon\right\}\\&=\mathrm{tr}\left\{\left(\mathbf{I}-\mathbf{A}^T-\mathbf{A}\right)\boldsymbol{\Sigma}_\varepsilon\right\}\\&=\mathrm{tr}\{\boldsymbol{\Sigma}_\varepsilon\}-2\mathrm{tr}\{\mathbf{A}\boldsymbol{\Sigma}_\varepsilon\}\end{aligned} \tag{21}$$

where tr is a trace function and, as previously, the model and sampling errors are assumed to be independent. For computational convenience, equation (21) is written as sums, i.e.,

$$\mathrm{E}\{S_C-S_T\}=\sum_{s=1}^{n}\Sigma_{\varepsilon,ss}-2\sum_{s=1}^{n}\sum_{i=1}^{n}a_{si}\Sigma_{\varepsilon,si} \tag{22}$$

The term in equation (22) can be considered a bias-correction term to be subtracted from the calculated sum of squares, S_C to obtain an improved estimate of S_T denoted $\hat{S}_T$.

4. A Case Study Using UK Data and Methods

The method developed above for using multiple donor sites to adjust a regression estimate for an ungauged site was tested using annual maximum series of instantaneous peak flow available at 602 nonurban catchments located throughout the United Kingdom as shown in Figure 1. The data set is part of the Hiflows-UK data set, consisting of peak flow series where the gauging authorities have sufficient confidence in the rating curves that these stations can be recommended for use in flood studies.

Using the same data set, *Kjeldsen and Jones* [2009] adopted a maximum-likelihood method to estimate the regression model parameters as well as the parameters in equation (3) controlling the relationship between model error correlation and distance between catchment centroids, and the results are replicated in Table 1.

The four catchment descriptors used in the regression model are catchment area (AREA) in km^2, standard annual average rainfall as measured from 1961 to 1990 (SAAR) in mm, an index of flood attenuation due to online reservoirs and lakes (FARL) which takes values between zero and one, (where a value of one indicates no attenuation), and BFIHOST which is related to the hydrological properties of the catchment soils and can take values between zero (impermeable) and one (completely permeable). Each of these descriptors is transformed as shown in Table 1. A detailed description of the model development and performance is reported in *Kjeldsen and Jones* [2009] and not repeated here.

The relative performance of the proposed donor transfer method was initially investigated by comparing the performance for predicting in ungauged catchments based on three different cases:

1. Using the regression model only

2. Identify the geographically nearest gauged catchments

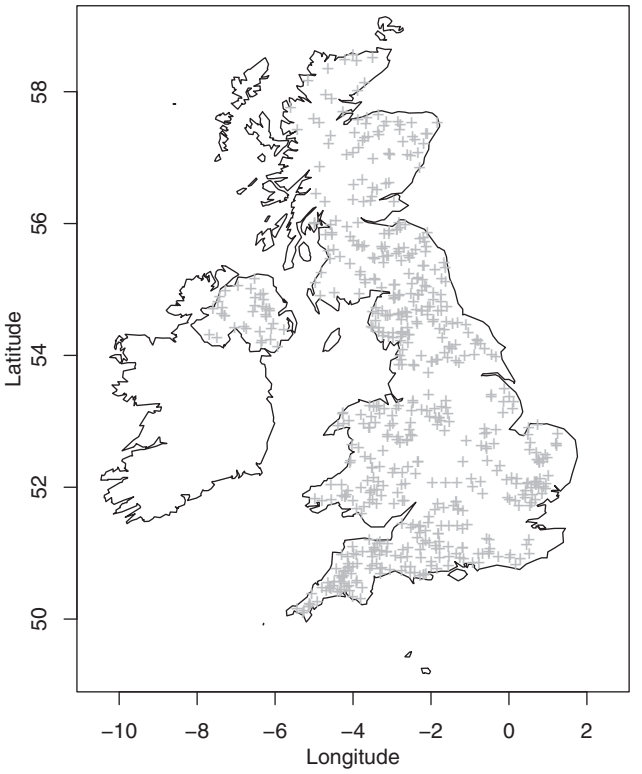

Figure 1. Location of the 602 gauging stations used in this study.

3. Identify donors connected to the subject site by being located on the same river network

Finally, an investigation was conducted into the relative value of the donor transfer method using local data by contrasting the performance to that of a number of existing regional models of varying complexity.

For all investigations, the performance of the models will be reported in terms of the factorial standard error (fse), which is defined as the exponential of the standard error as derived from the sum of squared errors as

$$\text{fse}=\exp\left(\sqrt{\frac{S}{m-(q+1)}}\right) \quad (23)$$

where S is the sum of square errors defined either as S_C from equation (18) or the bias corrected version S_T, m is the total number of catchments, and q is the number of catchment descriptors in the regional regression model. The RMSE is defined from equation (23) as RMSE$=log(\text{fse})$, but the advantage of the fse measure is that it can more easily be translated into confidence intervals than the RMSE measure, e.g., *Kjeldsen* [2014].

4.1. Nearest Geographical Neighbors

For the data set at hand, the potential number of donors for each subject site is 601 when selecting based on geographical distance only. Evaluating the performance of the adjustment procedure, the performance criteria $\text{fse}_C=\exp\left(\sqrt{S_C/(602-5)}\right)$ and $\text{fse}_T=\exp\left(\sqrt{\hat{S}_T/(602-5)}\right)$ were both evaluated as a function of the number of donor sites used to adjust the regression-only estimates of the index flood at the subject sites. The results are plotted as a function of the number of donor-sites in Figure 2 for a range of different cases: (i) the regression-only estimates (no donor adjustment), (ii) donor-adjusted estimates with and without neglecting sampling errors when calculating weights using the matrix $\mathbf{\Omega}$ in equation (10), and (iii) donor-adjusted estimates using the revised performance measure in combination with the full solution.

For the data set used in this study, a minimum level of fse (RMSE) values was reached when using six or more donors, and the corresponding fse (RMSE) values are reported in Table 2 below. It is also evident from Figure 2 that, in the case presented here, the omission of the sampling errors when calculating the weight assigned to each donor site has a relatively minor effect on the performance.

It is clear that the inclusion of multiple donors is beneficial in terms of reducing the prediction variance. The results suggest that the prediction accuracy is relatively insensitive to the actual number of donors, as long as this number exceeds about five.

4.2. Drainage Network Structure

Results presented by *Morris* [2003], *Skøien et al*. [2006], *Guse et al*. [2009], and *Ganora et al*. [2013] suggest that the connectivity of the site of interest and different gauging stations as determined by the river

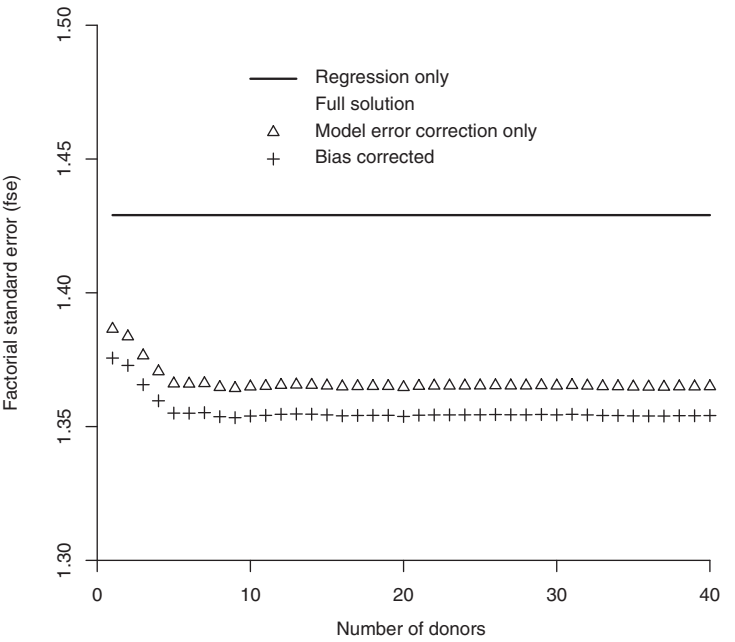

Figure 2. Factorial standard error (fse) plotted as a function of the number of donor sites for (i) the regression-only estimates (no donor adjustment), (ii) donor-adjusted estimates with and without neglecting sampling errors when calculating weights, and (iii) donor-adjusted estimates with and without using the revised performance measure.

network structure can be an important source of information when estimating the index flood at ungauged sites. Thus, an assessment of the benefit from selecting donor sites connected to the target site via the river network was undertaken. Based on the available data set of 602 AMS, river network-based donors were selected by first searching downstream from the subject site until a gauging station was encountered or the sea was reached, and then searching in an upstream direction up every tributary in turn, in each case stopping when a gauging station was encountered or the head of the tributary was reached. A maximum of seven donors was allowed as this was considered to be fully adequate for the needs of the original FEH statistical procedures. There are only 10 UK river reaches where the number of eligible gauges exceeds seven, and here the preferred upstream donors are those whose catchment area is closest to that of the subject site. At sites with fewer than seven network-based donors, additional donors are selected on the basis of proximity, provided that their catchments are sufficiently similar to that of the subject site. Because of the need for catchment similarity, not all of the subject sites have seven donors. Of the 602 locations in this study, 29 had no donors and 53 had between one and five donors.

Comparing the results in Table 3 with the corresponding results in Table 2 obtained without consideration of the river network structure, it can be observed that the resulting RMSE and fse values are generally reduced slightly when incorporating information on river network structure, especially when not considering the influence of the sampling errors of the at-site estimates. Removing the error contribution from the at-site samples, the resulting RMSE values are almost similar, suggesting that the actual network structure does not add additional information over and above that already contained in the distance between catchment centroids. This conclusion is derived from analysis of AMAX data, and it is of course possible that analysis of other types of hydrological data and indices would provide more insight into the role of the drainage network geometry and its influence on runoff processes.

4.3. A Regional Model Versus Local Data

A comprehensive set of catchment descriptors are available for each of the catchments included in the UK case study. However, other regions might have only a smaller subset of descriptors available, and in this section the value of local data will be evaluated for two cases of catchment descriptor availability. The performance of the data transfer scheme developed in this study will be compared when combined with a full regional model, i.e., the model in Table 1 based on four different catchment descriptors (AREA, SAAR, FARL, and BFIHOST) with a fse value equal to 1.43, and a second more simple model using catchment area (AREA) only as a covariate and a fse value of 2.76. The catchment area (AREA) only model was developed previously by *Kjeldsen and Jones* [2009]. The two models were contrasted in *Kjeldsen and Jones* [2010] who showed that the omission of catchment descriptors in the simple model resulted in a much higher degree of model error correlation. Based on these results, it was argued that the existence of high model error correlation increases the value of data transfer as a compensation of the lack of explanatory power of the simple regression model.

Table 2. RMSE (fse) When Using Six Donor Sites for Estimating the Index Flood at Each Target Site

Incl. Sampling Errors in Ω	Apply Bias Correction?	RMSE	fse
Yes	No	0.311	1.365
Yes	Yes	0.303	1.354
No	No	0.312	1.366
No	Yes	0.304	1.355

Table 3. RMSE and fse When Selecting Donors Site Based on Shared Network Structure

Incl. Sampling Errors in Ω	Apply Bias Correction?	RMSE	fse
Yes	No	0.308	1.360
Yes	Yes	0.304	1.356
No	No	0.308	1.360
No	Yes	0.304	1.355

Figure 3 shows a comparison between the fse values obtained using the full and the simple regression model, respectively, in combination with the data transfer scheme developed in this study. A third model based on AREA and SAAR is also shown, and this model has an fse value of 1.92 [*Kjeldsen and Jones*, 2009].

The horizontal lines represent the fse value for the three regression models, i.e., using only catchment descriptors but without the use of data transfer. Clearly the fse of the simple AREA only model is much higher than for the two more complex models, showing the importance of additional catchment descriptors beyond catchment area for describing the between-catchment variation in the index flood.

When incorporating local data through use of the data transfer scheme, the drop in fse observed in connection with the simple AREA only model is much larger than the corresponding reduction in fse when using the full model. In fact, the fse values obtained using the AREA only model in combination with data transfer are lower than the fse values obtained when using a more complex regional model based on both AREA and SAAR without data transfer. However, even with data transfer, the fse values obtained for the simple AREA-only model do not reach the low level associated with the four-descriptor model without data transfer. Nonetheless, these results suggest that there is potentially a large gain to be had in predictive power when using data transfer in combination with a simple regional model developed using catchment area only.

A more detailed assessment of the link between the complexity of the regional model and the benefit of data transfer can be made by studying more closely the behavior of the differences between the log-transformed at-site values, and corresponding log-transformed values predicted using the regionalized models with and without data transfer (residuals). Figure 4 (top row) shows the residuals (gray points) of the full regression model plotted against the four catchment descriptor values used in the model (AREA, SAAR, FARL, and BFIHOST), and the corresponding residuals (bottom row) from the area-only model. Note that the scale of the *y* axis is the same in both rows to better enable the comparison. In each of the eight plots, the spread of the residuals is indicated by a convex hull spanning 95% of all the points (solid line). A second convex hull (dashed line) spans 95% of the second set of residuals obtained when comparing at-

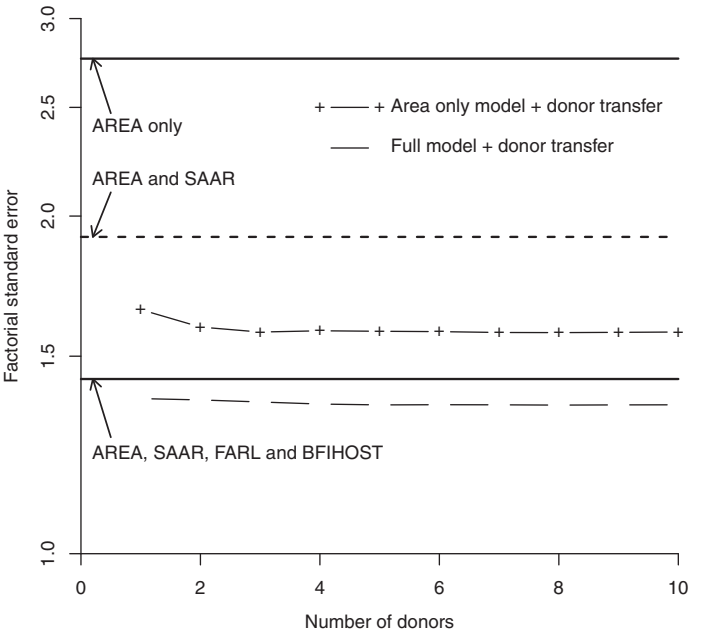

Figure 3. Factorial standard error (fse) plotted as a function of the number of donor sites for the regression-only estimates (no donor adjustment) and donor-adjusted estimates for (i) the full model and (ii) the simple catchment area only model. The fse for an intermediate regression-only model based on catchment area and mean annual rainfall is shown as a dashed line.

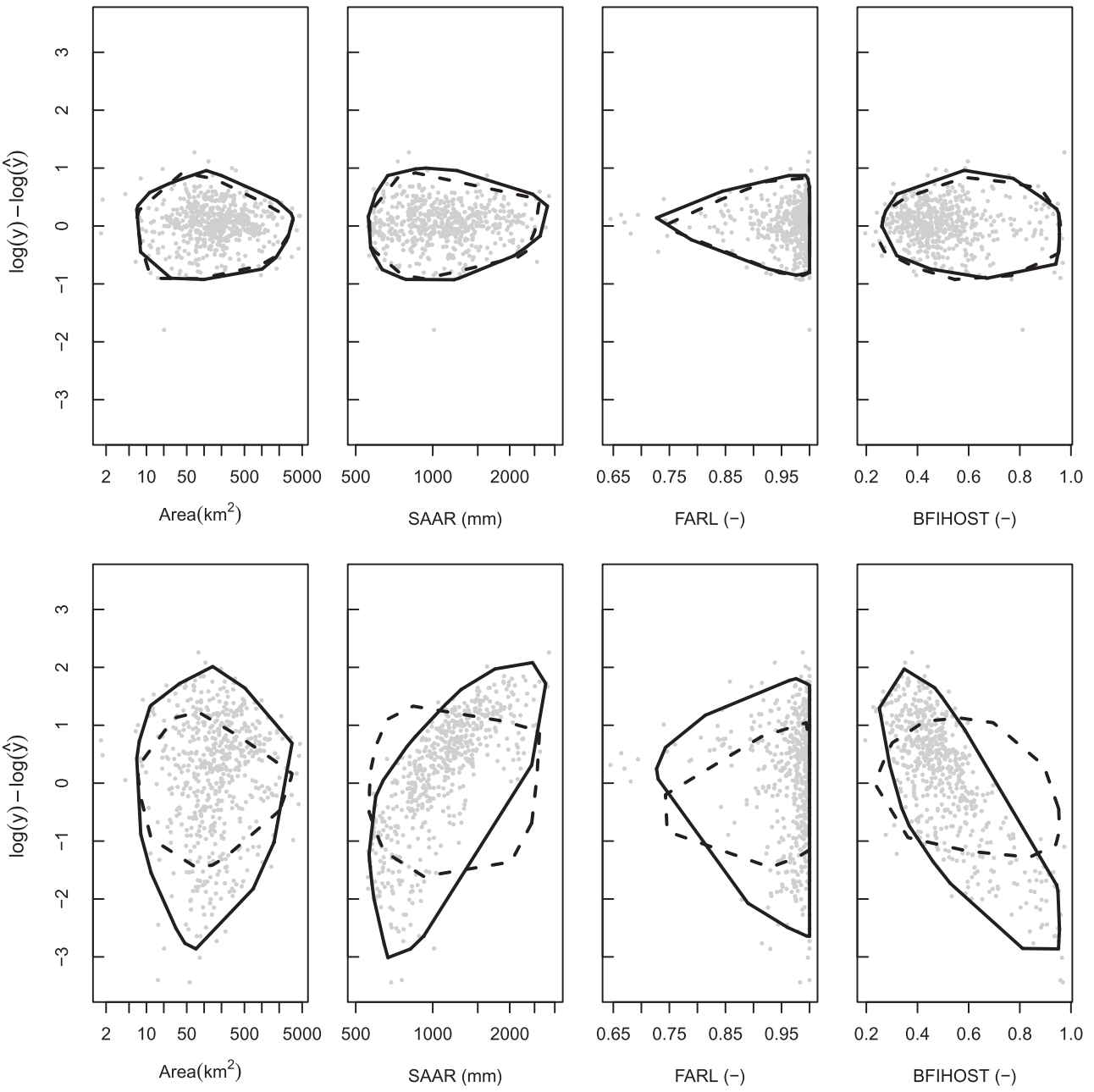

Figure 4. Residuals plotted against key catchment descriptors. 95% convex hull for regression model residuals (solid line—points shown in gray), and residuals obtained using regression model with data transfer from six donor sites (hatched line—points not shown). (top row) Results for full model and (bottom row) results for area-only model.

site values with estimates obtained by using data transfer with five donor sites. For this second set of residuals, only the convex hull is shown but not the actual points. For both sets of residuals, the 95% convex hull was chosen to limit the influence of outliers on a visual assessment of the general behavior of the residuals.

Considering first, the residuals of the full model in the top row, the second convex hull (dashed line) generally has the same shape and occupies an area that is marginally smaller than the convex hull spanning the regression model only residuals (solid line). This is in line with the results shown in Figure 3, and suggests that the utility of data transfer is more limited when the regression model accounts for the differences in the index flood values observed in different types of catchments. In contrast, the difference between the

two convex hulls is much more pronounced when considering the simpler area-only model (bottom row). First, the area spanned by the residuals obtained using data transfer is much smaller than the residuals from the regression model only and more similar to the shape observed for the full model in the top row. Second, there are clear structures in the regression-only residuals (grey points), especially when plotted against SAAR and BFIHOST, with underestimation in wet and impermeable catchments and, conversely, overestimation in dry and more permeable catchments. However, using local flood records largely removes this stratification from the residuals. For FARL (reservoirs and lakes), the results are less clear, but this is caused by the relatively limited number of catchments with low FARL values (stronger influence of upstream reservoirs and lakes) in the data set. These results show that using data transfer can (i) improve the performance of a simple regression model, and (ii) effectively remove most of the structure in the performance of regional models caused by not incorporating key catchment descriptors into the model.

5. Discussion and Conclusions

The procedure outlined in this study is a formal framework that will enable hydrologists to incorporate local data when estimating an index flood variable at an ungauged site using a generalized regional procedure. The adjusted estimates are shown to have a lower fse than those obtained from regression models only. This is particularly important in regions where the available lumped catchment descriptors cannot be considered to adequately capture local flood controlling processes and mechanisms. The methodology was developed for use with index flood estimates in the United Kingdom, where the optimal number of donors was found to be five or more.

The study also presented a generalized procedure for removing the effects of sampling error from RMSE (fse) and hence the effects of sample length available at each site. This is potentially an important result as it will allow comparison of performance of methods between data sets from different regions.

Interestingly, the derivations presented in this study and in *Kjeldsen and Jones* [2010] show that if the correlation between model errors is neglected (i.e., $r_{\eta,ij}=0$ in equation (3)) and not considered as part of the initial construction of the hydrological regression model, then there is seemingly no benefit associated with transfer of data to an ungauged catchment from nearby gauged catchments (as the regression model is then assumed to explain all between-catchment variation). In most practical settings, this would be an untenable position, and thus model error correlation should be considered an integral part of models attempting to predict hydrological variables in ungauged catchments.

Comparing the performance of regional regression models (in terms of predictive ability) when combined with data transfer from donor catchments showed that the benefit of data transfer in ungauged catchments, when used in combination with a simple regional model using catchment area only outperforms the more advanced regional model using both catchment area and mean annual rainfall as covariates if data transfer is not used. Thus, careful consideration of model error correlation in the model building phase can help to address poor model performance originating from access to only a limited subset of catchment descriptors. As such, the proposed method is considered a valuable addition to the toolbox available for regional hydrological models.

The results obtained in this study are based on flood data from a relatively dense gauging network in the United Kingdom. A case study by *Mediero and Kjeldsen* [2014] using the GLS framework to develop a model linking at-site estimates of a 100 year design flood to catchment descriptors in a north-east Spain identified model error correlation, suggesting that the method might also be useful in other geographical regions. However, further research is needed to verify the extent to which these conclusions are valid. In particular, the influence of reservoirs and lakes on flood characteristics should be further examined. Finally, this study considered only the index flood variable, but similar analysis could have been undertaken in connection with any hydrological variable where a regionalization model is the basis for prediction in ungauged catchments, including statistical moments of high and low flow series as well as parameters in rainfall-runoff models.

Acknowledgments
The authors would like to thank the UK measuring authorities for making the HiFlows-UK peak flow database available. Daniele Ganora and two anonymous reviewers are acknowledged for insightful comments on an earlier version of the manuscript.

References

Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione, and H. Savenije (2013), *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*, Cambridge Univ. Press, Cambridge, U.K.

Cunnane, C. (1988), Methods and merits of regional flood frequency analysis, *J. Hydrol.*, *100*, 269–290.

Dalrymple, T. (1960), Flood frequency analysis, *U.S. Geol. Surv. Water Supply Pap.*, *1543-A*, 80 p.

Ganora, D., F. Laio, and P. Claps (2013), An approach to propagate streamflow statistics along the river network, *Hydrol. Sci. J.*, *58*(1), 41–53.

Grover, P. L., D. H. Burn, and J. M. Cunderlik (2002), A comparison of index flood estimation procedures for ungauged catchments, *Can. J. Civ. Eng.*, *29*(5), 734–741.
Guse, B., A. Castellarin, A. H. Thieken, and B. Merz (2009), Effect of intersite dependence of nested catchment structures on probabilistic regional envelope curves, *Hydrol. Earth Syst. Sci.*, *13*, 1699–1712.
Haddad, K., A. Rahman, and J. R. Stedinger (2012), Regional flood frequency analysis using Bayesian generalized least squares: A comparison between quantile and parameter regression techniques, *Hydrol. Processes*, *26*, 1008–1021.
Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-moments*, Cambridge Univ. Press, Cambridge, U. K.
Institute of Hydrology (1999), *Flood Estimation Handbook*, Wallingford, U. K.
Kachroo, R. K., S. H. Mkhandi, and B. P. Parida (2000), Flood frequency analysis of southern Africa: I. Delineation of homogeneous regions, *Hydrol. Sci. J.*, *45*, 437–447.
Kjeldsen, T. R. (2014), How reliable are design flood estimates in the UK?, *J. Flood Risk Manage.*, doi:10.1111/jfr3.12090, in press.
Kjeldsen, T. R., and D. A. Jones (2009), An exploratory analysis of error components in hydrological modelling, *Water Resour. Res.*, *45*, W02407, doi:10.1029/2007WR006283.
Kjeldsen, T. R., and D. A. Jones (2010), Predicting the index flood in ungauged UK catchments: On the link between data-transfer and spatial model error, *J. Hydrol.*, *387*, 1–9.
Lim, Y. H., and D. L. Voeller (2009), Regional flood estimation in Red River using L moment based index flood and Bulletin 17B procedures, *J. Hydrol. Eng.*, *14*(9), 1002–1016.
Mediero, L., and T. R. Kjeldsen (2014), Regional flood hydrology in a semi-arid catchment using a GLS regression model, *J. Hydrol.*, *514*, 158–171, doi:10.1016/j.jhydrol.2014.04.007.
Merz, R., and G. Blöschl (2005), Flood frequency regionalisation: Spatial proximity vs. catchment attributes, *J. Hydrol.*, *302*, 283–306.
Merz, R., and G. Blöschl (2008), Flood frequency hydrology: 1. Temporal, spatial, and causal expansion of information, *Water Resour. Res.*, *44*, W08432, doi:10.1029/2007WR006744.
Meigh, J. R., F. A. K. Farquharson, and J. V. Sutcliffe (1997), A worldwide comparison of regional flood estimation methods and climate, *Hydrol. Sci. J.*, *42*(2), 225–244.
Morris, D. G. (2003), Automation and appraisal of the FEH statistical procedures for flood frequency estimation, Science report FD1603 to Defra, Centre for Ecology and Hydrology, Wallingford, U. K.
Parida, B. P., R. K. Kachroo, and D. B. Shrestha (1998), Regional flood frequency analysis of Mahi-Sabarmati basin (Subzone 3-a) using index flood procedure with L moments, *Water Res. Manage.*, *23*(11), 1573–1650.
Pearson, C. P. (1991), New Zealand regional frequency analysis using L moments, *J. Hydrol.* (NZ), *30*, 53–64.
Reis, D. S., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regionalised skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.
Salinas, J. L., A. Castellarin, S. Kohnová, and T. R. Kjeldsen (2013), On the quest for a pan-European flood frequency distribution: Effect of scale and climate, *Hydrol. Earth Syst. Sci. Discuss.*, *10*(5), 6321–6358.
Skøien, J. O., R. Merz, and G. Blöschl (2006), Top-kriging: Geostatistics on stream networks, *Hydrol. Earth Syst. Sci.*, *10*, 277–297.
Stedinger, J. R., and G. D. Tasker (1985), Regional hydrological analysis: 1. Ordinary, weighted and generalized least squares compared, *Water Resour. Res.*, *21*, 1421–1432.
Vogel, R. M., T. A. McMahon, and F. H. S. Chiew (1993), Floodflow frequency model selection in Australia, *J. Hydrol.*, *146*, 421–449.
Viglione, A., R. Merz, Salinas, J. L., and G. Blöschl (2013), Flood frequency hydrology: 3. A Bayesian analysis, *Water Resour. Res.*, *49*, 675–692, doi:10.1029/2011WR010782.
Yang, T., C. Y. Xu, Q. X. Shao, and X. Shen (2010), Regional flood frequency and spatial patterns analysis in the Pearly River delta region using L moments approach, *Stoch. Environ. Res. Risk A.*, *24*(2), 165–182.