Statistical modelling of flood risk at ungauged sites

By Thomas R. Kjeldsen, David A. Jones and Adrian C. Bayliss

Centre for Ecology & Hydrology, Wallingford, Oxfordshire

Abstract

The use of multivariable regression models which provide linkage between a particular hydrological variable and a set of physical catchment descriptors is a long established practice in applied hydrology. This paper focuses on the modelling and prediction of the median annual maximum index flood at gauged and ungauged sites through the use of regression modelling and on data transfer from gauged to ungauged catchments as outlined in the Flood Estimation Handbook (FEH). Through an extension of the commonly used regression model to include, in addition to cross correlation of sampling errors, non-zero cross correlation of model errors, it is possible to establish a more formal relationship between the regression model and the use of data transfer from a gauged (donor) catchment to an ungauged catchment. By explicitly considering the correlation between the regression model errors, a revised data transfer scheme has been developed, which was found to perform better in terms of predictive error than the established FEH scheme and the case where only the regression model is used. In fact, the automated version of the original FEH data transfer scheme used in this study was found to give estimates of the index flood with higher prediction variance than estimates obtained using regression only.

1. Introduction

The Flood Estimation Handbook (FEH) published by IH (1999) is used as the standard for flood frequency analysis in the UK. The statistical method for flood risk assessment outlined in the FEH is based on the index flood method, where, for an ungauged catchment, the index flood is estimated through a multivariable linear regression model linking the index flood, defined in the FEH as the logarithm of the median annual maximum flood (QMED), to a set of catchment descriptors. The FEH guidelines then emphasise the importance of data transfer from nearby gauged catchments (donor catchments) to enhance the initial regression estimate. However, little guidance was previously available to practitioners to assist in the selection of donor catchments. This paper presents a revised regression model for estimation of the index flood at ungauged sites combined with a more formalised framework for data transfer, as an alternative to the procedure used in the FEH methodology. The regression model is formulated to represent a covariance structure, including both a sampling error component resulting from the limited sample sizes and a model error component arising from the inability of a simple linear regression model to accurately represent the complex dynamics of real catchments. By introducing a specific model component accounting for the correlation between the model errors, it is possible to derive an optimal data transfer scheme from a gauged donor catchment to an ungauged site of interest. The results presented in this paper are based on analysis of annual maximum series of peak flow and the associated FEH catchment descriptors for 602 catchments included in the web-based dataset produced by the Environment Agency led HiFlows-UK project (www.environment-agency.gov.uk/hiflowsuk).

2. Regression of the index flood on catchment descriptors

Consider a vector of sample (log-transformed) median annual floods, y_i, where individual sites are denoted with a subscript i. Each sample value is considered an estimate of the underlying true population value of the median, i.e.

$$\ln[y_i] = \ln[\xi_i] + \varepsilon_i \qquad (2.1)$$

where ε_i is the sampling error of the log-transformed median annual flood with a mean value of $\mathrm{E}\{\varepsilon_i\} = 0$ and a covariance structure which will be specified later. The notion of a true value, ξ_i, is defined here as a hypothetical median of an infinite sample of annual maximum flood peaks from a catchment in a stationary condition. Next, consider the actual model, where it is assumed that the true log -transformed index flood can be estimated as a linear combination of a set of catchment descriptors and a site specific model error

$$\ln[\xi_i] = \mathbf{x}_i^T \boldsymbol{\beta} + \eta_i \tag{2.2}$$

where $\boldsymbol{\beta}$ is a vector of true regression model parameters, $\mathbf{x}_i$ is a vector of catchment descriptors and η_i is the regression modelling error with the statistical properties

$$\begin{aligned} &\mathrm{E}\{\eta_i\} = 0 && i = j \\ &\mathrm{cov}\{\eta_i, \eta_j\} = \sigma_\eta^2 r_{\eta,ij} && i \neq j \end{aligned} \tag{2.3}$$

where the model error correlation $r_{\eta,ij}$ will be estimated from a maximum-likelihood procedure outlined below. By combining equation (2.1) and equation (2.2), the sample estimate of the index flood can be expressed in terms of the true regression model and two error components representing the sampling and modelling error, respectively

$$\ln[y_i] = \mathbf{x}_i^T \boldsymbol{\beta} + \eta_i + \varepsilon_i \tag{2.4}$$

The covariance matrix of the sampling errors is denoted $\boldsymbol{\Sigma}_\varepsilon$, the corresponding covariance matrix of the modelling errors is denoted $\boldsymbol{\Sigma}_\eta$ and the two errors are assumed mutually independent. It is assumed that the elements along the diagonal of the modelling error covariance $\boldsymbol{\Sigma}_\eta$ are identical (σ_η^2) and that the associated correlation matrix $\mathbf{R}_\eta$ has unit diagonal elements. The two error terms are generated from two very different processes. The sampling covariance is caused by similarity of the flood generating rainfall events striking two catchments located close to one another, whereas correlation of the model errors is caused by the inability of a simple regression type model to adequately represent the relationship between catchment descriptors and the index flood at different sites. The covariance matrix of the total errors is the sum of the sampling and model error components

$$\boldsymbol{\Sigma}_\eta + \boldsymbol{\Sigma}_\varepsilon = \sigma_\eta^2\left(\mathbf{R}_\eta + \boldsymbol{\Sigma}_\varepsilon / \sigma_\eta^2\right) = \sigma_\eta^2 \mathbf{G} \tag{2.5}$$

where **G** is a composite matrix which plays a particular role in the computations. Each element in the sampling error covariance $\boldsymbol{\Sigma}_\varepsilon$ is estimated based on considerations of the asymptotic variance of the sampling median and defined as

$$g_{\varepsilon,ij} = \begin{cases} 4\beta_i^2 / n_i & i = j \\ 4\beta_i \beta_j \dfrac{n_{ij}}{n_i n_j} r_{\varepsilon,ij} & i \neq j \end{cases} \tag{2.6}$$

where β_i is the scale parameter of the Generalised Logistic (GLO) distribution standardised to unit median and estimated using the L-moment ratios as shown by Kjeldsen and Jones (2006). Here n_{ij} denotes the number of years where both series have data, while n_i and n_j denote the total number of years for the two series separately. Note that the conventional notation for the GLO distribution and the regression model both use “beta” but with two distinct meanings. In addition, estimation of the off-diagonal elements requires estimates of the correlation coefficient between the log-transformed median annual maximum flood for each site, $r_{\varepsilon,ij}$, which can be estimated directly from the dataset through bootstrapping. Based on 1000 bootstrap replications, Kjeldsen and Jones (2007a) used all pairs of records with more than 39 overlapping years to estimate the correlation between the log-transformed median annual maximum peak flow values and related it to geographical distance between catchment centroids using a weighted sum of two exponential distributions, as shown in Figure 1

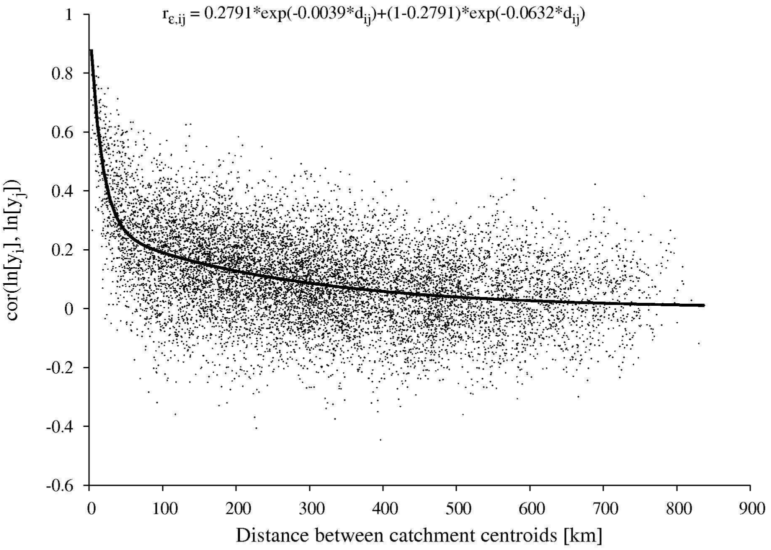

FIGURE 1. Correlation between sampling errors of log-transformed median annual maximum flood as a function of distance between catchment centroids.

Similarly to the correlation between sample values of the log-transformed median illustrated in Figure 1, the non-diagonal elements in the model error correlation matrix, $\mathbf{R}_{\eta}$, are described as a weighted sum of two exponential distributions as

$$r_{\eta,ij} = \psi \exp(-\varphi_1 d_{ij}) + (1-\psi)\exp(-\varphi_2 d_{ij}) \tag{2.7}$$

where ψ, φ_1 and φ_2 are model parameters and d_{ij} is the geographical distance [km] between catchment centroids. It is important to note that the model parameters are not necessarily equal to those used for the correlation between the sample values shown in Figure 1. Justification for equation (2.7) can be found in Kjeldsen and Jones (2007a) but the interpretation is that the regression residuals from nearby catchments have a tendency to be more positively correlated for nearby pairs of gauges, i.e. a simple regression model fails to encompass some local factors controlling flood response. It seems reasonable to assume that a simple regression model cannot fully represent the complexity of real catchments. However, it is important to include this known behaviour of the model error explicitly into the modelling framework to ensure statistically correct estimates. The relationship in equation (2.7) plays an important role when transferring information from donor sites to ungauged sites as we shall demonstrate later.

The regression model error, σ_{η}^2, and the additional three parameters in equation (2.7) describing the model error correlation are estimated using the maximum-likelihood method. Assuming an initial set of regression model parameters, $\boldsymbol{\beta}$, the four model error parameters are estimated by minimising the negative of the log-likelihood function

$$-\ln[L] = \frac{1}{2}\ln\left[\det\left(\sigma_{\eta}^2\mathbf{G}\right)\right] + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\left(\sigma_{\eta}^2\mathbf{G}\right)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{2.8}$$

For any trial value of the four parameters, an estimate of the composite matrix **G** is used for obtaining an updated estimate of the regression model parameters using the well-known generalised least square (GLS) estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{y} \tag{2.9}$$

where **y** is a vector of sample values of the log-transformed median. The resulting regression model parameters as well as the parameters of the model error correlation model are shown in Table 1.

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^{-1}	-1.8734	0.0968	-19.35	0.000
Ln[FARL]	3.4450	0.2654	12.98	0.000
BFIHOST^2	-3.0800	0.1158	-26.60	0.000
$\sigma_\eta^2 = 0.1286$ df = 598 r^2 = 0.945				
$\psi = 0.4598$ $\varphi_1 = 0.0200$ $\varphi_2 = 0.4785$				

TABLE 1. Results of regression analysis

The resulting regression model used for predicting the index flood in ungauged catchments is given as

$$y_i = 8.3062 AREA^{0.8510}\, 0.1536^{\left(\frac{1000}{SAAR}\right)} FARL^{3.4451}\, 0.0460^{BFIHOST^2} \tag{2.10}$$

where the catchment descriptors *AREA*, *SAAR*, *FARL* and *BFIHOST* represent catchment area (km^2) and catchment average values of annual average rainfall 1961-1990 (mm), flood attenuation due to upstream reservoirs and lakes, and the hydrological properties of catchment soils, respectively. The descriptors are available for any UK catchment larger than 0.5 km^2 and are further described by Bayliss (1999). The model error variance reported in Table 1 is equal to $fse = 1.431$ which is 7.5 % less than the corresponding value of *fse* = 1.549 reported for the corresponding QMED model in the FEH (IH, 1999), where the factorial standard error (*fse*) is defined as $fse = \exp(\sigma_\eta)$. The choice of catchment descriptors in Table 1 and the particular transformations used here has been based on other analysis, which are not described here but included examining the model residuals by plotting them against catchment descriptors. Note that the model presented in Table 1 and equation (2.10) is provided as an example of a model estimated from the dataset rather than a direct substitute for the QMED equation presented in the FEH.

3. Using donor adjustment

When conducting a flood frequency analysis at an ungauged site, the FEH strongly recommends transferring data from catchments judged to be hydrologically similar to the subject site and for which annual maximum flood data are available. However, in a comprehensive assessment of the FEH statistical method, Morris (2003) found inappropriate adjustment of the regression model estimate using donor catchments to be a major source of potential error. In a separate study, Kjeldsen and Jones (2007b) analysed the benefits of using data transfer from donor sites from the perspective of reducing prediction variance at the site of interest. The results obtained by Kjeldsen and Jones (2007b) enabled a more analytical approach than that of Morris (2003) to be carried out and the resulting improved data transfer scheme is presented below.

3.1 The FEH donor adjustment

Once a suitable donor site has been identified, the index flood at the site of interest is estimated as

$$y_{s,adj} = y_{s,cds} \frac{y_{g,obs}}{y_{g,cds}} \tag{3.1}$$

where the subscript *s* refers to the ungauged subject site and *g* the gauged donor site, the subscript *cds* refers to catchment descriptor estimates at the gauged and subject sites, *obs* the observed value at the gauged site and *adj* the adjusted value at the subject site. While this adjustment assumes the residuals from the regression equation at both the subject and the donor site exhibit the same behaviour, the recommended procedure makes no use of the distance-based model for the model error correlation that is included in the FEH model (IH, 1999). The linkage between the model error correlation and the prediction variance of ln[$y_{s,adj}$] was derived by Kjeldsen and Jones (2007b) to be approximately

$$\mathrm{var}\{\ln[y_{s,adj}] - \ln[\xi_s]\} \approx 2\sigma_\eta^2 (1 - r_{\eta,sg}) + g_{\varepsilon,gg} \tag{3.2}$$

where $r_{\eta,sg}$ is the correlation of the model errors of the subject and donor catchment derived from equation (2.7) and $g_{\varepsilon,gg}$ is the sampling variance of the log-transformed median at the donor site. In most cases $\sigma_\eta^2 >> g_{\varepsilon,gg}$ and therefore, unless the donor and subject catchments are located very closely together, the prediction variance arising from the donor transfer quickly increases to twice that obtained using the regression model only. In fact, from equation (3.2), it is clear that unless $r_{\eta,sg} > 0.5$ the donor transfer is not preferable with the FEH data transfer method. Based on equation (2.7) and the parameters in Table 1 this corresponds to a maximum distance between catchment centroids of about 4 km.

3.2 A new data transfer scheme

A major advance of the FEH statistical method developed as part of this project is the ability to identify and estimate a separate model for the model error correlations (Kjeldsen and Jones, 2007a). Kjeldsen and Jones (2007b) showed that knowledge of the model error correlation can be use to define an alternative data transfer scheme of the form

$$y_{s,adj} = y_{s,cds}\left(\frac{y_{g,obs}}{y_{g,cds}}\right)^{\alpha} \qquad (3.3)$$

where the new parameter α is estimated by minimising the prediction variance of ln[$y_{s,adj}$] given as

$$\mathrm{var}\{\ln[y_{s,adj}] - \ln[\xi_s]\} \approx \sigma_\eta^2 + \alpha^2(\sigma_\eta^2 + g_{\varepsilon,gg}) - 2\alpha\sigma_\eta^2 r_{\eta,sg} \qquad (3.4)$$

where σ_η^2 is the model error variance, $g_{\varepsilon,gg}$ is the sampling variance of ln[y_g] and $r_{\eta,sg}$ is the model error correlation between the subject *s* and the donor *g* sites calculated using the model specified in equation (2.7), i.e. based on the geographical distance between the subject and the donor site. The resulting estimator of α is

$$\alpha = r_{sg}\,\frac{\sigma_\eta^2}{\sigma_\eta^2 + g_{\varepsilon,gg}}\,. \qquad (3.5)$$

As mentioned before, the sampling error of ln[y_g] ($g_{\varepsilon,gg}$) is generally much smaller than the model error variance and, thus, for most practical purposes, the α parameter in equation (3.5) reduces to $\alpha = r_{\eta,sg}$ which is given by equation (2.7) with the model parameters shown in Table 1.

4. Application

The effect of data transfer when predicting the index flood for ungauged catchments has been investigated based on estimates obtained for 602 catchments from the HiFlows-UK dataset and using three different approaches:

i) using only the regression model and predicting the index flood based on catchment descriptors only,
ii) identifying the geographically closest catchment, using catchment centroids, out of the 601 other gauged catchments and using the FEH data transfer procedure equation (3.1); and
iii) identifying the donor as in ii) but using the new data transfer procedure in equation (3.3).

To assess the performance of each of the three methods, the root mean square error (RMSE) was derived for each method as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\left(\ln[y_{s,adj,i}] - \ln[y_{g,i}]\right)^2}{M-5}} \qquad (4.1)$$

where the subscripts *s*, *g*, *adj* and *obs* are described in a previous section. The degrees of freedom are M-5 = 602 – 5 = 597 corresponding to the five parameters in the regression model. Note that the use of RMSE as defined above is somewhat flawed since it has to make use of the sample median as the “target” for the estimation rather than the true median. Thus this empirical measure of performance is affected by the sampling error and by the correlation of these. The RMSE values obtained for each of the three options are shown in Table 2, where it can be observed that, while the new data transfer method improves the RMSE when compared to using regression only, the FEH data transfer scheme has, in fact, a higher RMSE than regression only. The latter finding indicates that, on average, the FEH data transfer scheme does not improve the prediction compared to using the regression model only.

Method	RMSE
Regression only	0.357
FEH data transfer	0.377
New data transfer	0.327

TABLE 2. RMSE for each of the three methods predicting the index flood in ungauged catchments

To further investigate the structure of the RMSE values, the 602 catchments were divided into 20 groups according to the distance from a particular catchment and its closest donor catchment. Each of the 20 groups span a distance of 1 km and within each group the RMSE was estimated as

$$RMSE_i = \sqrt{\frac{1}{M_i}\sum_{j=1}^{M_i}\left(\ln\left[y_{s,adj,j}\right]-\ln\left[y_{j,obs}\right]\right)^2} \tag{4.2}$$

where M_i is the number of catchment pairs in the i’th group. For each of the three methods, the RMSE was estimated for each of the 20 groups and the results plotted on Figure 2.

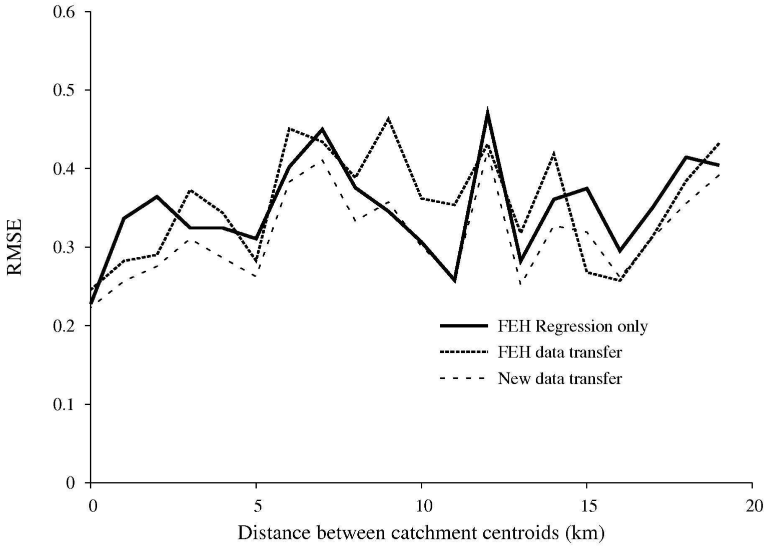

FIGURE 2. RMSE for 1 km intervals in distance between subject and donor catchments.

As observed on Figure 2, both the FEH and the new transfer scheme have improved the predictions compared to using regression only for very short distances less than 3 km. In general, the new transfer scheme is consistently performing better than both the regression-only option and the FEH data transfer scheme, whereas the FEH method often gives higher RMSE values than the regression model on its own. This is confirmed by the average RMSE values reported in Table 2.

5. Conclusion

The analytical framework presented in this paper represents a significant improvement in the ability to estimate the index flood (or any other hydrological variable) at both gauged and ungauged sites. By estimating the correlation between the regression model errors and successfully linking it to geographical distance between catchments, it is possible to make an objective assessment of the weight attached to data transferred from a neighbouring donor catchment.

The results obtained in the comparison of the predictive ability of the different methods showed that the performance of the traditional FEH data transfer scheme (3.1) performs rather more poorly than expected. In fact, on average, a smaller prediction error was obtained using the regression model only than when using the regression model with the FEH donor transfer method. The simple automated donor selection method implemented in this study might have been improved somewhat if the selection for each catchment had been carried out manually, but the conclusion is not likely to have changed much, i.e. the FEH donor transfer scheme should be used with care and does not necessarily guarantee an improved estimate. In comparison, the improved transfer scheme developed in this study is consistently outperforming the regression model and the FEH transfer scheme. It is therefore recommended that this scheme should be adopted for practical use.

Acknowledgements

The work presented in this paper was partly funded by the Joint Defra/EA Flood and Coastal Erosion Risk Management R&D Programme Project SC050050.

REFERENCES

BAYLISS, A C. 1999

Catchment Descriptors, Flood Estimation Handbook, Volume 5, Institute of Hydrology, Wallingford, UK.

INSTITUTE OF HYDROLOGY (IH) 1999

The Flood Estimation Handbook, Institute of Hydrology, Wallingford, UK.

KJELDSEN, T.R. & JONES, D.A. 2006

Prediction uncertainty in a median based index flood method using L-moments. *Water Resources Research*, **42**, W07414, doi :10.1029 / 2005WR004069.

KJELDSEN, T.R. & JONES, D.A. 2007a

Recursive estimation of a hydrological regression model. *In: Proceedings of World Water and Environmental Resources Congress 2007 ASCE*, Tampa, Florida, USA.

KJELDSEN, T.R. & JONES, D.A. 2007b

Estimation of an index flood using data transfer in the UK. *Hydrol. Sci. J.* **52**(1), 86-98.

MORRIS, D.G. 2003

Automation and appraisal of the FEH statistical procedures for flood frequency estimation. *Final Report to Defra*, Centre for Ecology & Hydrology, Wallingford, UK.