Prediction uncertainty in index flood modelling at ungauged catchments

Thomas R Kjeldsen and David A Jones
Centre for Ecology & Hydrology, Wallingford, UK: trkj@ceh.ac.uk/daj@ceh.ac.uk

Abstract
This paper discusses the prediction variance of an index flood when estimated for an ungauged catchment. Three different methods are investigated: i) using only a newly developed regression model linking median annual flood to a set of four catchment descriptors, ii) an extension using the FEH data transfer method from a nearby gauged catchment to an ungauged catchment, and iii) using a modified version of the data transfer scheme. The results illustrate the link between the structure of the errors of the regression model and the utility of the data-transfer from gauged to ungauged catchments.

Introduction

Flood frequency analysis based on the index-flood method is the most widely applied method for design flood estimation at ungauged catchments in the UK, as described in the Flood Estimation Handbook (Institute of Hydrology, 1999). The method is based on analysis of annual maximum peak flow data. The index-flood method assumes that data from all catchments within a specified homogeneous region have identical frequency distributions, except for a site-specific scale parameter, the index flood (Hosking and Wallis, 1997). The FEH adopted the median annual maximum flood as the index flood, which differs slightly from the more traditional choice of the mean annual maximum flood. Estimates of the index flood can be obtained using both direct and indirect methods, depending on the availability of data at the site of interest. Direct methods include estimation of the index flood directly from available at-site annual maximum data, whereas indirect methods attempt to estimate the index flood at ungauged sites where no observed flow data are available.

This paper discusses and compares the uncertainty of the prediction errors of the index flood when estimated at an ungauged site using three different methods: i) using a newly developed regression model linking median annual flood to a set of four catchment descriptors, ii) using an extension of this with the FEH data transfer method to incorporate data from a nearby gauged catchment, and iii) using a modified version of the data transfer scheme. Both data transfer methods rely on the regression model, and it will be shown that the correlation structure of errors of the regression model are important when evaluating the uncertainty of the prediction errors of estimates obtained using data transfer.

A hydrological regression model

The estimation of a regression model linking the index flood to a set of catchment descriptors in the UK is described in detail by Kjeldsen and Jones (2008) and only a short summary is given here. To relate the index flood variables from *n* different catchments to a set of catchment descriptors, consider a vector of sample (log transformed) median annual maximum floods, **y**, where individual sites are denoted with a subscript *i*. Each sample value is described in terms of a population regression model and two individual error components representing the modelling and sampling errors, η and ε, respectively so that

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta} + \eta_i + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\theta} + \omega_i \qquad (1)$$

where $\boldsymbol{\theta}$ is a vector of regression model parameters and $\mathbf{x}_i$ is a vector of catchment descriptors with a value of one in the first location. Both errors are assumed normally distributed with zero mean values. The covariance matrix of the sampling errors is denoted $\boldsymbol{\Sigma}_\varepsilon$, the corresponding covariance of the modelling errors denoted $\boldsymbol{\Sigma}_\eta$, and the two errors are assumed mutually independent. Further, it is assumed that the elements along the diagonal of the modelling error covariance matrix are identical and equal to σ_η^2. The covariance matrix of the vector of total errors, $\boldsymbol{\omega}$, is defined as

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

where $\mathbf{R}_\eta$ is the correlation matrix of the modelling error. The matrix **G** is introduced for computational convenience and is derived from values of σ_η^2 and $\mathbf{R}_\eta$. In pioneering the use of the Generalised Least Squares (GLS) procedure in hydrology, Tasker and Stedinger (1989) assumed the modelling covariance matrix to be of the form $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \mathbf{I}$, i.e. there is an assumption of no cross correlation between the modelling errors. In contrast, the model formulated here is more general and assumes the cross correlation to be represented by the associated modelling error correlation matrix $\mathbf{R}_\eta$.

The sampling and model error components represent two distinctly different sources of error in the regression model. Start by assuming that a ‘true’ value of the index flood could be estimated for each catchment if an infinite long series of annual maximum peak flow data was available. In practice, the index flood has to be estimated from finite series which introduces a *sampling error* representing the difference between this sample estimate and the notional true value. The *modelling error* represents the inability of a particular regression model to adequately predict the true value of the index flood. For hydrological models such as the regression model studied here, the model error is often much larger than the sampling error if a reasonable number of years have been used to estimate the index flood.

Similarly, the correlations between catchments of the individual error terms have very different interpretations for the two types of error. Correlation between sampling errors is a result of rainfall events causing increased flow in neighbouring catchments at the same time. The existence of correlation in model errors on the other hand, signifies an inability of a particular regression model to adequately represent the true values of the index flood in neighbouring catchments, i.e. the existence of regional clusters of under and over prediction. It can be argued that the existence of model error correlation is a result of an inadequate regression model and should be removed by improving the regression model. However, the approach taken here argues that a simple regression model is unlikely to capture the complex behaviour of real catchments and acknowledges this inability by explicitly allowing model error correlation into the modelling framework.

While the sampling errors are related to the data set used for estimation of the index flood at each individual site, the model errors are specific to a particular regression model, i.e. each choice of a set of catchment descriptors will result in its own specific model error structure. This means that the statistical properties of the sampling error can be estimated once and used in all regression models whereas those of the model error need to be estimated for each regression model tested. Details of the estimation of the sampling error covariance, $\boldsymbol{\Sigma}_\varepsilon$, are not shown here but can be found in Kjeldsen and Jones (2008). Based on detailed investigations, Kjeldsen and Jones (2008) found that model error correlation across sites could reasonably be described for most regression models as

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

Here, d_{ij} is the distance between catchment centroids [km] and φ_1, φ_2 and φ_3 are model parameters that must be estimated for each individual regression model.

Having specified the error structure, the regression model parameters can be estimated using a maximum-likelihood procedure which incorporates what are essentially the steps involved in calculating the GLS estimates of the regression parameters. If it is assumed that the regression residuals are normally distributed with mean zero and a total covariance matrix, $\sigma_\eta^2\mathbf{G}$, as described in equation (2), the objective of the overall estimation procedure is to minimise the negative log-likelihood function,

$$-2\ln(L) = \ln\left[\det\left(\sigma_\eta^2\mathbf{G}\right)\right] + (\mathbf{y}-\mathbf{X\theta})^T\left(\sigma_\eta^2\mathbf{G}\right)^{-1}(\mathbf{y}-\mathbf{X\theta}), \qquad (4)$$

with respect to the three model error correlation parameters (φ_1, φ_2 and φ_3), the model error variance, σ_η^2, and the regression parameters, $\mathbf{\theta}$. The problem is simplified by noting that, for given values of σ_η^2 , φ_1, φ_2 and φ_3 (which between them determine **G**) , the value of $\mathbf{\theta}$ which minimises (4) is given the GLS estimator

$$\hat{\mathbf{\theta}} = \left(\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{G}^{-1}\mathbf{y}\,. \qquad (5)$$

Thus, estimation by Maximum Likelihood is implemented as a search over the four parameters σ_η^2, φ_1, φ_2 and φ_3. The results are shown in Table 1.

Table 1: Summary statistics for regression model describing the (log transformed) median annual maximum flood. (Maximum-Likelihood estimates)

Coefficient	Parameter θ_p	Standard error	t-value	p-value
Intercept (θ_0)	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	-3.0800	0.1158	-26.60	0.000
$\sigma_\eta^2 = 0.1286$, df = 597, r^2 = 0.945 (log scale) $\varphi_1 = 0.4598$ $\varphi_2 = 0.0200$ $\varphi_3 = 0.4785$				

Using annual maximum data from 602 rural catchments located throughout the UK, a five parameter regression model was developed, linking the log-transformed median annual maximum flood to a set of four different catchment descriptors. The estimated model parameters are shown in Table 1 where AREA, SAAR, FARL and BFIHOST are catchment descriptors describing catchment area [km^2], standard average annual rainfall 1960-90 [mm], upstream reservoir attenuation and a measure of the relative baseflow contribution as derived from HOST soil data. These catchment descriptors are available for all gauged and ungauged catchments in the UK larger than 0.5 km^2 (Institute of Hydrology, 1999).

The particular choice of catchment descriptors in Table 1 and their transformation used here has been based on other analyses which are not described here but which included the examination of the model residuals by plotting them against catchment descriptors.

To estimate the variance of the prediction errors, consider first an estimate of the (log transformed) index flood obtained at an ungauged subject site

$$\hat{y}_s = \mathbf{x}_s^T \hat{\boldsymbol{\theta}} \qquad (6)$$

which is considered an estimate of the true (log transformed) index flood, ξ_s, defined as

$$\xi_s = \mathbf{x}_s^T \boldsymbol{\theta} + \eta_s \qquad (7)$$

where subscript *s* indicates the ungauged subject site. The prediction error is then defined as

$$\hat{y}_s - \xi_s = \mathbf{x}_s^T \hat{\boldsymbol{\theta}} - \mathbf{x}_s^T \boldsymbol{\theta} - \eta_s = \mathbf{x}_s^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) - \eta_s \qquad (8)$$

The full variance of the prediction error in equation (8) is given by Kjeldsen and Jones (2007). However, under the assumption that the model error variance is significantly larger than the sampling error in such estimates, the prediction error variance can reasonably by simplified to the model error variance only, i.e.

$$\mathrm{var}\{\hat{y}_s - \xi_s\} \approx \sigma_\eta^2 \qquad (9)$$

Table 1 also contains estimates of the three parameters describing the correlation between model errors across sites through equation (3). While this correlation does not have a significant effect on the variance of the prediction error when using the regression model at a particular site, it is important in determining the prediction error variance when combining these estimates with data transferred from neighbouring gauged sites, as will be discussed in the next section.

Using data transfer

As the UK has a relatively dense gauging network, the FEH generally recommends using data transfer from ‘hydrologically similar’ sites for which annual maximum series are available. The data transfer from the gauged to the ungauged catchment is conducted using a scaling factor applied to the non-transformed index flood estimate:

$$m_{s,adj} = m_{s,cds} \frac{m_{g,obs}}{m_{g,cds}}, \quad m = \exp(y) \qquad (10)$$

where the subscripts are as follows. *s* and *g*: the ungauged subject site and the gauged sites, respectively, *cds:* catchment descriptor estimates at the gauged and ungauged sites; *obs:* the observed value at the gauged site; *adj:* the adjusted value at the subject site. Kjeldsen and Jones (2007) found the variance of the prediction error of the (log transformed) adjusted index flood, $\hat{y}_{s,adj}$, to be given as

$$\mathrm{var}\{\hat{y}_{s,adj} - \xi_s\} \approx 2\sigma_\eta^2 \left(1 - r_{\eta,sg}\right) + h_{gg} . \qquad (11)$$

Here $r_{\eta,sg}$ is the correlation between the model errors at the subject site and the gauged site and h_{gg} is the sampling error of (the logarithm of) the median at the gauged site (see Kjeldsen and Jones (2007) for an analytical expression of h_{gg} not shown here). The record length at the gauged site is often sufficiently long that the expression above is dominated by the first term only. Note that if the model error correlation, $r_{\eta,sg}$, was assumed zero, as done in the GLS model proposed by Tasker and Stedinger (1989), then the prediction error variance becomes almost twice as large as the variance of the error from the regression model alone.

Kjeldsen and Jones (2007) suggested an alternative data transfer scheme

$$m_{s,adj} = m_{s,cds}\left(\frac{m_{g,obs}}{m_{g,cds}}\right)^{\alpha}, \quad m = \exp(y) \tag{12}$$

where the new parameter α is estimated by minimizing the variance of the prediction error for the (log transformed) adjusted index flood $\hat{y}_{s,adj}$ and is given by

$$\alpha = r_{\eta,sg}\,\frac{\sigma_{\eta}^{2}}{\sigma_{\eta}^{2} + h_{gg}}\,. \tag{13}$$

Consequently, the variance of the prediction error for the (log transformed) adjusted index flood is given by

$$\operatorname{var}\{\hat{y}_{s,adj} - \xi_{s}\} \approx \sigma_{\eta}^{2}\left(1 - r_{\eta,sg}^{2}\right) + r_{\eta,sg}^{2} h_{gg}\,. \tag{14}$$

If a sufficiently long record is available at the gauged site the adjustment factor reduces to $\alpha = r_{\eta,sg}$ and the prediction error variance in equation (14) will be dominated by the first term.

Example

The effect of data transfer on the prediction error variance at an ungauged site as compared to the prediction error variance of an estimate obtained from the regression model only is illustrated in Figure 1 by comparing the standard deviation of the prediction error from each of the three methods. Assuming that a long enough record would be available at a gauged site, the only parameter controlling the value of the prediction error variance obtained using data transfer is the distance between catchment centroids, via equation (3) with parameter values listed in Table 1.

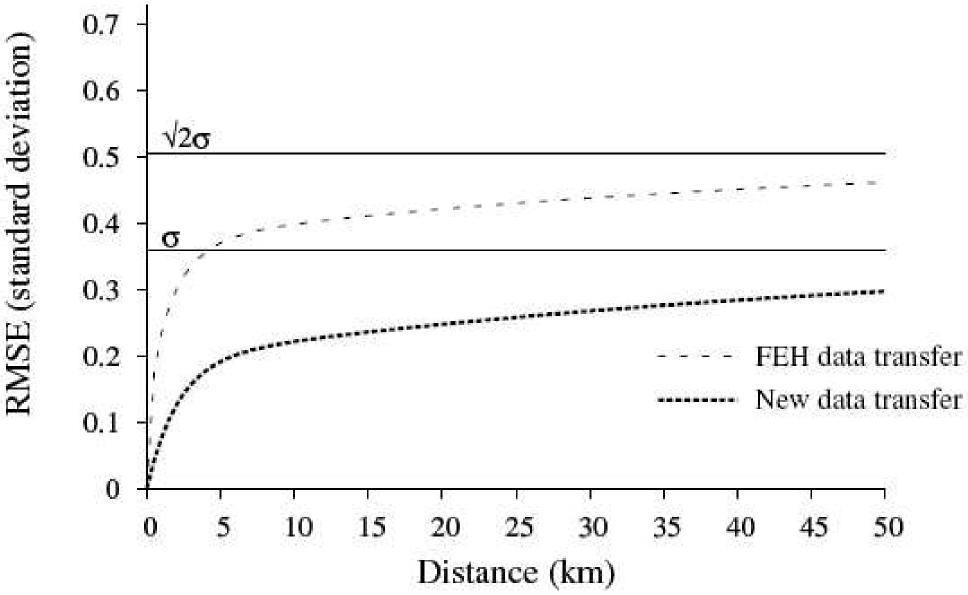

Figure 1: Comparison of standard error of prediction errors at ungauged sites using the regression model only, FEH data transfer and the new data transfer scheme.

From Figure 1 it is clear that unless catchments are located very close together, estimates of the index flood obtained using the original FEH transfer scheme will have prediction error variance twice the size of the corresponding estimate from the regression model only. The new transfer scheme, however, ensures that the prediction error variance never exceeds the corresponding variance obtained from the regression model only as the effect of the gauged site is reduced as the distance increases. In Figure 1, the standard deviation has been plotted without considering the sampling uncertainty and, hence, the two curves representing the data transfer methods show zero variance at distance zero, where in fact they should show the sampling variance of the gauged site, which would depend on the record-length for that particular record.

Conclusions

By explicitly identifying and estimating the two error sources in a regression model it is possible to derive analytical expressions of the prediction variance of estimates obtained at the ungauged site using data transfer from a gauged site. The results clearly show any improvement to be gained from data transfer arises from correlation in the modeling error. If the model error correlation is not fully taken into account (as in the FEH method) the resulting prediction error of the index flood will have a variance twice as large as the variance of the error from using the regression model only.

References

Hosking, J. R. M. and Wallis, J. R. (1997) *Regional frequency analysis: An approach based on L-moments*. Cambridge University Press, USA.

Institute of Hydrology (1999) *Flood Estimation Handbook*, 5 Volumes, Institute of Hydrology, Wallingford, UK.

Kjeldsen, T. R. and Jones, D. A. (2007) Estimation of an index flood using data transfer in the UK. *Hydrological Sciences Journal*, 52(1), 86-98.

Kjeldsen, T. R. and Jones, D. A. (2008) An exploratory analysis of error components in hydrological regression modelling. Submitted to *Water Resources Research*

Tasker, G. D. and Stedinger, J. R. (1989) An operational GLS model for hydrological regression. *Journal of Hydrology*, 111, 361-375.