Predictions in Ungauged Basins: PUB Kick-off (Proceedings of the PUB Kick-off meeting held in Brasilia, 20–22 November 2002). IAHS Publ. 309, 2007.

The effects of errors in measuring drainage basin area on regionalized estimates of mean annual flood: a simulation study

R. T. CLARKE & L. C. BRUSA

Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS Brasil

clarke@iph.ufrgs.br

Abstract Regionalized estimates of mean annual flood in ungauged drainage basins are typically found by using multiple regression equations fitted to data from neighbouring gauging stations. Amongst the explanatory variables used in fitting such regressions, basin area is one of the most important, commonly showing close correlation with mean annual flood, the variable to be predicted. In regions of very low relief, however, determining the boundaries of drainage basins, whether from maps or on the ground, is not straightforward. Using data from tributaries of the River Uruguay in southern Brazil, this paper reports results of a simulation study which quantifies the magnitude of errors in estimates of mean annual flood, that result from errors in drainage basin area.

Key words catchment area errors; mean annual flood; regionalization

INTRODUCTION

There is a frequent need to estimate the flow characteristics of rivers at sites where no flow records exist. Techniques of hydrological "regionalization" seek to obtain estimates of flow characteristics by using data from neighbouring drainage basins to establish relationships between flow characteristics and variables describing the geomorphology and climate. Multiple regression is commonly used for this purpose (Thomas & Benson, 1970; NERC, 1975). If Q denotes a flow characteristic derived from n_i years of record at site i (I = 1, ..., N) for N sites on rivers within a region, a model is postulated of the form $Q = \gamma A^{\alpha}B^{\beta}C^{\delta}$.... where α, β, γ, δ... are parameters to be determined from flow records available at the N sites, and A, B, C... are the geomorphological and climate characteristics measured or estimated for each of the N catchment areas. The model is then put into a form in which the parameters appear linearly by transforming to logarithms, giving a multiple regression model of the form:

$$\ln Q = \beta_0 + \beta_1 \ln A + \beta_2 \ln B + \beta_0 \ln C + ... + \varepsilon \tag{1}$$

where the residuals are taken to be uncorrelated random variables with a probability distribution having zero mean and constant variance independent of i. Extensions to this simplified model have been developed by Stedinger & Tasker (1985, 1986) and Tasker & Stedinger (1989). Where the flow characteristic to be estimated at the ungauged site is the mean annual runoff volume V, or the annual peak discharge Q_T with return period T, catchment area, here denoted by A, is an extremely important predictor. If the corresponding relationships are denoted by:

$$Q_T = c_1 A^{\theta} \tag{2a}$$

$$V_m = c_2 A^{\phi} \tag{2b}$$

then the empirical values of the scaling exponents θ, ϕ have been found to lie in the range $0.5 < \theta < 1$ and $0.6 < \phi < 1$ (Leopold *et al.*, 1964; Benson, 1962, 1964; Alexander, 1972). Benson (1962, 1964) found that the exponent θ tended to decrease with increasing aridity (for example, $\theta = 0.85$ for humid New England, whilst $\theta = 0.59$ for semi-arid Texas and New Mexico). Benson also found, and was later supported by Alexander (1972) using world-wide data, that θ actually decreased with increasing catchment area. Goodrich *et al.* (1997) recently examined peak runoff Q_T and annual water yield V_m over a range of catchment sizes on the Walnut Gulch watershed in Arizona, concluding that θ changed from 0.90 for small subcatchments to 0.55 for larger subcatchments, whilst ϕ varied from 0.97 to 0.82 over the same area ranges. Similarly, in a detailed regionalized analysis of mean annual peak discharge in Great Britain and Ireland, NERC (1975) found values of the scaling exponent θ extending up to 0.970.

It is of interest to consider why the scaling exponents are always less than unity in all the cases reported here, and what factors cause increased departures from this value. Intuition would suggest that V_m would be directly proportional to drainage basin area *A*, but this is not the case. If the linear regression model in (1) is correct (that is, if all explanatory variables have been included on the right-hand side, and all terms are linear), then estimates of their coefficients are unbiased estimates of their true values; however, if—as is invariably the case—the model is incorrect, estimates of the regression coefficients will be biased (Draper & Smith, 1981). However, there is no reason why the omission of explanatory variables, or incorrect model form on the right-hand side of (1), should result in values consistently less than unity, as is the case with the coefficient of ln*A*, and hence in values of the scaling exponents in (2a, 2b) which are also less than one. A partial explanation for the phenomenon may be as follows. To be explicit, denote the linearized form of (2a) by $\ln Q = \beta_0 + \beta_1 \ln A + \varepsilon$. Then ordinary regression theory assumes that that the independent or explanatory variable ln*A* is free from measurement error; but if this variable is subject to error, then the usual linear regression of β_1 will be biased downwards. A simple example illustrates the point. Suppose that the flow variable *Q* is directly proportional to basin area *A* with constant of proportionality one, so that $Q = A$. Let *A* take the values 1, 2...10, so that *Q* also has these values. Since the true values of *Q* will not be observed exactly but will be subject to errors that are assumed to be random and independent, introduce Normal errors N(0, 0.1) to the true *Q* values. So far, there are errors in the *Qs*, but not in the areas *A*. Now introduce errors of *X*% in each of the areas $A = 1, ..., 10$, by: (a) tossing an unbiased coin; (b) if the result is heads, impose an error of + *X*% on the current value of *A*, otherwise impose an error of –*X*%. Now repeat the calculation 200 times, with the result shown in Table 1.

It is clear that the larger the measurement errors in *A*, the smaller is the mean value $\bar{b}_1$ of the 200 estimates of β_1 obtained from the 200 simulated pairs of values (*Q*,*A*), and the larger the mean intercept $\bar{b}_0$ of the 200 regressions. Thus the effect of errors in the measurement of drainage area is to "flatten" the regression by making it more horizontal, and this confirms intuition: namely that large errors in *A* will tend to

Table 1 Means $\bar{b}_0$, $\bar{b}_1$ of 200 estimates of regression coefficients β_0, β_1 (true values: $\beta_0 = 0$, $\beta_1 = 1$) with values *A*, *Q* simulated from $A = 1...10$, $Q = A + \varepsilon$, $\varepsilon \sim N(0,0.1)$, and with each *A*-value disturbed by $\pm X\%$ with equal probability.

X	0%	5%	10%	15%	20%
$\bar{b}_0$	0.0006	0.0429	0.1834	0.3654	0.6921
$\bar{b}_1$	1.0014	0.9904	0.9688	0.9341	0.8910

depress the true relation between *Q* and *A*, reducing the proportion of variance in ln*Q* explained by the relationship with ln*A*. This may explain, in part, the results of Benson (1962, 1964), Alexander (1972) and Goodrich *et al.* (1997) cited above. The fact that the United Kingdom is a small and well-mapped country may also explain why the NERC Report (1975) found scaling exponents θ extending up to 0.970, not much less than one. However, in areas of very low relief covered by tall vegetation, which is the case over large parts of Brazil, identification of catchment boundaries is difficult whether from the use of maps, aerial photographs or satellite images, so the errors in estimating *A* may be considerable.

The question remains whether measurement errors implicit in the estimation of catchment areas are sufficient to explain in full the departures from unity for values of the scaling coefficients θ, ϕ given in the literature. This is the hypothesis explored in this paper, using data from a drainage basin in the south of Brazil.

THE DATA

The data used in the study were from 21 sub-catchments within the drainage basin of the Rio Ibicuí, which is in the southwestern part of the Brazilian State of Rio Grande do Sul, near the frontiers with Uruguay and Argentina, as shown in Fig. 1. The Ibicuí is the main left-bank tributary of the River Uruguay, and for much of its length flows in the east–west direction along the 386 km from its source to its junction with the River Uruguay. Its principal tributaries are the Rivers Santa Maria, Jaguarí, Ibirapuitã, Toropi, Itú, Ibirocá and Ibicui-Mirim, giving a total drainage area of roughly 47 740 km^2.

METHOD

Table 2 shows the number of years of record *n*, mean annual mean daily flow *Q*, and the drainage area for each of 21 subcatchments within the drainage basin of the Rio Ibicuí. For each subcatchment, the value of *Q* was obtained as follows. In each day, water-level was recorded at 07:00 and 17:00 hours, and a rating curve was used to estimate the corresponding discharges. The two discharges so estimated were averaged to give a mean daily flow, and the maximum of the mean daily flows was selected from each of the *n* years of record. The value *Q* shown in Table 2 is the mean of these *n* values.

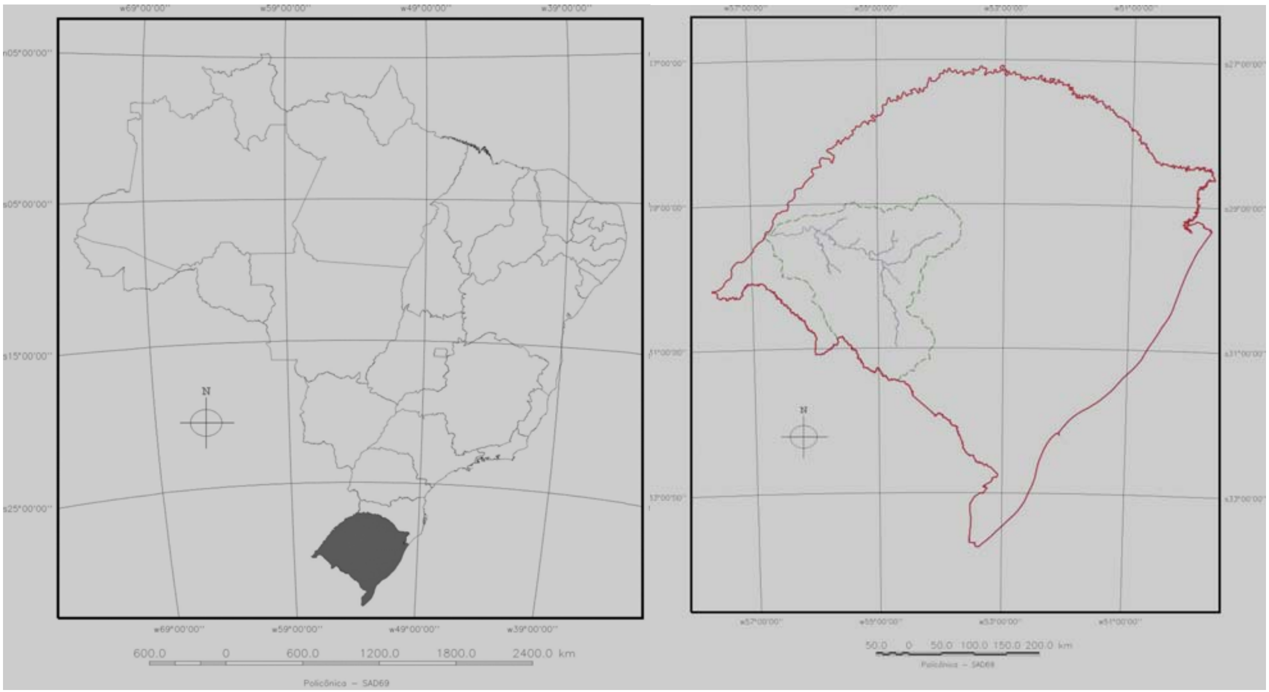

Fig. 1 Location of the State of Rio Grande do Sul (RS) within Brazil (left figure): position of R. Ibicuí and its tributaries within RS (right figure).

Table 2 Numbers of years of record (n), mean annual peak discharge (Q), and drainage areas for 21 subcatchments within the Rio Ibicui drainage basin.

n	21	53	28	37	25	29	30
Q	451.9	862.3	646.1	1219.5	155.6	798.4	1681.4
Area	1635	2783	3310	5679	2101	6005	12077
n	39	55	34	10	31	32	19
Q	371.9	822.7	490.6	304.4	1019.3	2520.2	70.6
Area	1826	2296	933	1345	4578	27771	376
n	32	16	49	15	21	56	40
Q	3094.7	2539.4	415.5	474.4	207.2	991.1	4063.7
Area	29321	31008	2562	3194	1163	5942	42498

A weighted regression analysis of the data given in Table 2 showed a highly significant relationship between ln$Q$ and ln *Area*; the regression on ln *Area* accounted for 82.6% of the variation in ln$Q$, the fitted regression being ln$Q$ = 0.701 + 0.713 ln *Area*. The standard errors of $\hat{\beta}_0, \hat{\beta}_1$ were ±0.617 and ±0.0729 respectively. Thus, although the slope of the weighted regression was significantly less than one, there was no evidence that the regression intercept was significantly greater than zero, such as would occur if the regression had been flattened by errors in determining drainage areas. As a next step, values of the variable A (catchment area) were perturbed, either positively or negatively with equal probability, by adding or subtracting 5% of their values, and after each such perturbation of the 21 values A, the weighted regression of ln$Q$ on the perturbed ln A was recalculated. This calculation was repeated 500 times. The same procedure was followed for perturbations of 10%, 15% and 20%, with the results shown in Table 3 (although errors as large as 15% or 20% in the estimation of catchment area are improbable).

Table 3 shows that the same features of Table 1 are broadly reproduced when a similar procedure is used with the real data of Table 2. However, even when errors of ±10% are introduced in the catchment areas A, the effect on the regression slope $\bar{b}_1$ is rather small: namely a reduction from 0.713 to 0.709. In the case of the regression intercept, the effect of introducing errors of ±10% in the areas A is larger, with $\bar{b}_0$ increased from 0.701 to 0.745. However the real point at issue is not how errors in A affect the regression slope and intercept, but how such errors would affect the estimates of Q, the annual maximum mean daily discharge estimated from the regionalized equation $\ln Q = \hat{\beta}_0 + \hat{\beta}_1 \ln A$. To explore this point, a set of hypothetical areas A = 500, 1000, 5000, 10000, 20000 and 40000 km^2 was taken, covering approximately the range of A values given in Table 2.

Table 3 Means $\bar{b}_0, \bar{b}_1$ of 500 estimates of regression coefficients β_0, β_1 after perturbation of values of *Area* shown in Table 2, by ±X% with equal probability and $X\%$ =5, 10, 15 and 20%.

X:	0%	5%	10%	15%	20%
$\bar{b}_0$	0.701	0.710	0.745	0.786	0.866
$\bar{b}_1$	0.713	0.712	0.709	0.704	0.697

Table 4 shows the estimated values of Q obtained: (a) from the regression of ln$Q$ on ln$A$, before errors in A were introduced, and (b) from the regression of ln$Q$ on ln$A$, obtained as the mean of 500 simulated regressions within each of which areas A has been perturbed by 10%. Averaged over the 500 simulations, the changes to Q resulting from the 10% perturbations are small. However the smallness of this average difference obscures the variation between the predicted Qs given by individual simulations; the extent of this variation is given in the last line of Table 4 which shows an approximate interval that included 95% of the predicted Qs. Even the most extreme of the predicted Qs do not differ greatly from the value of Q obtained without any perturbation of the areas A.

Table 4 Estimates of Q obtained (a) from the regression of ln$Q$ on ln$A$, before errors in A were introduced, and (b) from the regression of ln$Q$ on ln$A$, obtained as the mean of 500 simulated regressions within each of which areas A has been perturbed by 10%. The last line shows an approximate range that included 95% of 500 Qs, obtained by perturbing A in the way described.

A (1000 km ^{2}):	0.5	1	5	10	20	40
(a)	170	278	876	1436	2355	3860
(b)	173	282	883	1444	2360	3858
±2.SD	±12	±16	±28	±58	±130	±280

The above analysis assumed that the estimated drainage areas of the 21 sub-basins are subject to errors that may be positive or negative with value $\pm X$%, with equal probability. However, if all areas are subject to X% errors which are either all positive or all negative, the effects of such errors on estimates Q are greater, as shown in Table 5. In this Table, Column (1) is Q estimated from regionalization, assuming no errors in drainage area A; Column (2) gives adjusted estimates Q when true drainage areas are all consistently overestimated by 10% (obtained by setting drainage areas A equal to 90% of their given values); Column (3) gives adjusted estimates Q when true drainage areas are all consistently underestimated by 10% (obtained by setting drainage areas A equal to 110% of their given values). Even in this case, however, the magnitude of the adjustments to estimated Q are of the order of 8% or less.

Table 5 Effects on estimation of mean annual flood Q ($m^3 s^{-1}$) of errors in drainage basin area A (km^2). For explanation of columns (1), (2), (3), see text.

A	(1)	(2)	(3)
500	169	183	158
1000	278	300	260
5000	876	944	818
10000	1436	1548	1341
20000	2354	2537	2199
40000	3859	4160	3605

CONCLUSION

Based on the evidence of the simulations described, there is no evidence from the Ibicui data that errors in determining catchment area A greatly affect estimates of Q obtained by regionalization of data from 21 sub-basins. Whilst errors in catchment area A influence the slope and intercept of the regression equation $\ln Q = \beta_0 + \beta_1 \ln A + \varepsilon$ by depressing the slope and increasing the intercept, the effect is relatively minor over the range of areas for which the regionalized equation is likely to be used.

REFERENCES

Alexander, G. N. (1972) Effect of catchment area on flood magnitude. *J. Hydrol.* **16**, 225–240.

Benson, M. A. (1962) Factors influencing the occurrence of floods in a humid region of diverse terrain. *US Geol. Survey Water Supply Paper 1580-B*.

Benson, M. A. (1964) Factors affecting the occurrence of floods in the south-west. *US Geol. Survey Water Supply Paper, 1580-D*.

Draper, N. R. & Smith, H. (1981) *Applied Regression Analysis*, 2nd edn. John Wiley & Sons Inc., New York, USA.

Goodrich, D. C., Lane, L. J., Shillito, R. M., Miller, S. N., Syed, K H. & Woolhiser, D. H. (1997) Linearity of basin response as a function of scale in a semi-arid watershed. *Water Resour. Res.* **33**(12), 2951–2965.

Leopold, L. B., Wolman, M. G. & Miller, J. P. (1964) *Fluvial Processes in Geomorphology.* W.H. Freeman, New York, USA.

NERC (1975) *UK Flood Studies Report*, vol. 1: *Hydrological Studies*. Natural Environment Research Council, London, UK.

Stedinger, J. R. & Tasker, G. D. (1985) Regional hydrologic regression. 1: Ordinary, weighted and generalised least squares compared. *Water Resour. Res.* **21**(9), 1421–1432.

Stedinger, J. R. & Tasker, G. D. (1986) Regional hydrologic regression. 2: Model error estimates, estimation of sigma, and log-Pearson type 3 distributions. *Water Resour.Res.* **22**(10) 1487–1499.

Tasker, G. D. & Stedinger, J. R. (1989) An operational GLS model for hydrologic regression. *J. Hydrol.* **111**, 361–375.

Thomas, D. M. & Benson, M. A. (1970) Generalization of streamflow characteristics from drainage-basin characteristics. *USGS Water Supply Paper 1975.*