VOLUME 7 N°01 CONFLUENCE
Journal of Watershed Science and Management

Fitting Power-Law Relationships in Watershed Science and Analysis, with an Example Using the R Language

R. Dan Moore, *Department of Geography, University of British Columbia*

Abstract

Many quantitative relationships in the environmental sciences, and specifically in watershed science, can effectively be modelled using a power-law function. Such relationships are often estimated using ordinary least squares regression after linearizing the relationship by log-transforming both the x and y variables. Alternative approaches include nonlinear least squares regression and generalized nonlinear least squares regression. However, there are some differences in the underlying characteristics of these models that can result in the generation of different relationships and associated prediction limits. This article provides an overview of the statistical models underlying these approaches, then illustrates their application using the R language for an example based on fitting a regional relationship to predict flood quantiles from catchment area.

Keywords: *law relationships, log-transformation, nonlinear least squares regression, generalized nonlinear least squares regression*

Introduction

Many quantitative relationships in the environmental sciences can effectively be modelled using a power-law relationship. Mathematically, a power-law relationship is expressed as $y = \alpha x^{\beta}$, where x and y are variables and α and β are parameters. Some examples of power-law relationships in the watershed sciences include hydraulic geometry relationships (Gleason, 2015; Eaton & Davidson, 2022), drainage basin morphometry (Gardiner & Park, 1978; Brardinoni & Hassan, 2006), relationships between glacier volume and area (Adhikari & Marshall 2012; Bahr, Pfeffer, & Kaser, 2015), relationships between sapflow from individual trees and basal area for computing stand-level transpiration (Chiu et al., 2016), relationships between suspended sediment concentration and stream discharge (Ferguson, 1986; Bywater-Reyes, Bladon, & Segura, 2018), concentration-discharge relationships for streamwater chemistry (Basu et al., 2010; Knapp et al., 2020), stage-discharge relationships for weirs and flumes (Boiten, 1993), regional analysis of flood quantiles (McCuen, Leahy, & Johnson, 1990; Eaton, Church, & Ham, 2002; Northwest Hydraulic Consultants, 2020), paired-catchment analyses of streamflow response to forestry (Gronsdahl et al., 2019), relationships between rainfall quantiles and duration in intensity-duration-frequency analysis (Moore & Hutchinson, 2017), and fractal analysis of snow cover (Shook & Gray, 1997).

The majority of applications in watershed science and analysis fit power-law relationships using ordinary least squares (OLS) regression after linearizing the relationship by log-transforming both the x and y variables, followed by a back-transformation to the original measurement scale by exponentiation. However, as has been documented for decades in the watershed science literature, the back-transformed predicted values will underpredict the mean response (e.g., Ferguson, 1986; Cohn et al., 1989). Several bias corrections have been proposed (e.g., Bradu & Mundlak, 1970; Baskerville, 1972; Duan, 1983). However, simpler methods as recommended by Baskerville (1972) and Duan (1983) may not

Moore. (2024). Fitting Power-Law Relationships in Watershed Science and Analysis, with an Example Using the R Language. http://confluence-jwsm.ca /index.php/jwsm/article/view/53/7 . doi:10.22230/jwsm.2024v7n1a53

entirely remove bias (Cohn et al., 1989), while the method developed by Bradu and Mundlak (1970) involves an iterative solution. An additional complication of back-transforming from log-transformed values is that prediction limits for new observations will not be symmetric about the predicted values (Baskerville, 1972).

McCuen, Leahy, and Johnson (1990) and Chen et al. (2020) promoted the use of nonlinear least squares (NLS) regression for fitting power-law relations. One issue that McCuen, Leahy, and Johnson (1990) and Chen et al. (2020) did not address in detail is heteroscedasticity of the error variance, which occurs when the error variance exhibits systematic variability in time, in space, or in relation to the predictor variables. A common occurrence in environmental data sets in which values span multiple orders of magnitude is that the variance is not constant (i.e., is heteroscedastic), and typically tends to increase as the magnitude of the *y* variable increases. Indeed, one of the benefits of the log-log transformation is that it tends to stabilize the error variance for power-law relations, albeit at the cost of bias associated with back-transformation.

Heteroscedasticity can be addressed when fitting a relationship using NLS regression by specifying weights for each point used to fit the relationship (Neter et al., 1996). Alternatively, generalized nonlinear least squares (GNLS) can be applied, in which the parameters in a variance function can be estimated as part of the overall model fitting procedure.

The dominant practice of fitting power-law relationships using OLS regression with log-transformed variables likely reflects the fact that most watershed analysts have learned how to apply OLS in an introductory statistics course, but not NLS or GNLS. In addition, OLS regression is supported in all statistical software packages, and can be readily performed in spreadsheet applications. While it is possible to implement the NLS algorithm in a spreadsheet application (e.g., using the *Solver* add-in in Excel), it can be challenging to incorporate spreadsheet-based analyses into integrated, reproducible workflows that can be applied to data sets that are large and/or contain multiple subsets (Moore & Hutchinson, 2017).

The R programming language includes functions that facilitate fitting power-law relationships using the OLS, NLS, and GNLS approaches and, in conjunction with associated software such as RStudio and Quarto, provides an efficient framework for integrated, reproducible workflows (Marwick, Boettiger, & Mullen, 2018). A further advantage for watershed science and analysis is that the R language is popular for hydrological applications, and a range of contributed packages have been developed specifically for hydrologic analyses (Moore & Hutchinson, 2017; Slater et al., 2019). For example, the tidyhydat package provides easy, programmatic access to the Water Survey of Canada HYDAT data base (Albers, 2017), while the CSHShydRology package provides functions for a broad range of hydrologic applications, including catchment delineation and visualization of hydrologic time series data (Shook et al., 2022).

The objective of this article is to illustrate and compare the use of OLS regression on log-transformed variables, NLS regression with and without weights, and GNLS regression for fitting power-law relations, with specific reference to functions in the R programming language. The application of these approaches is illustrated through a regional analysis of flood quantiles, which is a common approach for estimating design floods for ungauged locations. All scripts, data sets, and files used in this analysis and the production of the manuscript are publicly available in a repository; see data availability statement for details.

Overview of approaches

Ordinary least squares regression with log-transformed variables

The use of OLS regression with log-transformed variables is based on a model with multiplicative errors, which can be expressed as follows:

$$y_i = \alpha x_i^{\beta} \epsilon_i \tag{1}$$

where ϵ_i is a random error term. Taking the logarithms of both sides of equation (1) yields the following expression as follows:

$$log(y_i) = log(\alpha) + \beta log(x_i) + log(\epsilon_i) \tag{2}$$

which can be expressed in the form of a linear regression model as follows:

$$\widehat{log(y_i)} = b_0 + b_1 log(x_i) \tag{3}$$

where $log(y_i)$ is the predicted value of $log(y_i)$, and b_0 and b_1 are the estimated intercept and slope, respectively. It is usually assumed that $log(\epsilon_i)$ are drawn from a normal distribution and have con-

stant variance. Under these conditions, ϵ_i would follow a log-normal distribution with a variance that varies with x_i.

Predictions of y_i can be made by back-transforming the result of equation (3) by exponentiation. Alternatively, the coefficients of the back-transformed relationship can be estimated as

$$\hat{\alpha} = exp(b_0) \tag{4}$$

and

$$\hat{\beta} = b_1 \tag{5}$$

where the hat indicates an estimated value. However, if the error terms ϵ_i follow a log-normal distribution, the back-transformed relationship would predict the median response for a given value of x, which would be less than the mean response. As mentioned in the introduction, several bias corrections have been derived. For illustrative purposes, back-transformation will be applied here using the non-parametric smearing estimator proposed by Duan (1983):

$$c_s = \frac{1}{n}\sum_{i=1}^{n} e^{v_i - \hat{v}_i} \tag{6}$$

where n is the number of data points used to fit the relation, v_i=$log(y_i)$, and $\hat{v}_i$ is the predicted value of $log(y_i)$ based on equation (3). The bias correction involves multiplying predictions based on equation (3) by c_s. The smearing estimate approach can be used for a number of transformations applied to the y variable, not just logarithmic; see Duan (1983) for details.

Nonlinear least squares regression

In contrast to OLS on log-transformed variables, the model to be estimated using NLS regression is based on an additive error model, which can be expressed as

$$y_i = \alpha x_i^{\beta} + \epsilon_i \tag{7}$$

where y_i and x_i are the response and predictor variables, respectively, for the i^{th} observation, α and β are the "true" values of the power-law parameters, and ϵ_i is a random error for observation i, assumed to have a mean of 0. The form of equation (7) implies that if one could fix x at a specific value, x_j, and generate an unlimited number of independent values of y, the result would be a distribution of values that would have a mean or expected value, given by

$$E(y|x = x_j) = \alpha x_j^{\beta} \tag{8}$$

where $E(y|x=x_j)$ is the expected value of y conditional upon x=x_j.

The coefficients α and β are generally not known *a priori* and must be estimated from the data. The fitted model can be expressed as:

$$y_i = a x_i^{b} + e_i \tag{9}$$

where a and b are estimates of α and β, respectively, and e_i is the residual for data point i, computed as

$$e_i = y_i - \hat{y}_i = y_i - a x_i^{b} \tag{10}$$

where $\hat{y}_i$ is the predicted value for observation i. Nonlinear least squares regression involves the use of an iterative algorithm to find the values of a and b that minimize the sum of squared errors, computed as:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a x_i^{b}\right)^2 \tag{11}$$

If the error terms are independent and normally distributed with constant variance, prediction limits can be computed for predictions made for new observations (Neter et al., 1996). In addition, the estimated parameters would have the property of being approximately minimum variance estimators as well as being unbiased.

Nonlinear least squares regression with specified weights

One concern with minimizing *SSE* as computed from equation (11) is that it gives all observations equal weight. If the residual error variance is heteroscedastic—which is a natural consequence of the multiplicative error model seen in equation (1), when working on the original measurement scale—the resulting parameter estimates would no longer have the statistical properties that hold in the

case of constant variance, although they would remain unbiased. This concern can be addressed by minimizing a weighted sum of square errors, SSE_w, computed as

$$SSE_w = \sum_{i=1}^{n} w_i \left(y_i - a x_i^b\right)^2 \tag{12}$$

where w_i is the weight for point i, which should be equal to, or at least proportional to, $1/\sigma_i^2$, where σ_i^2, is the error variance for point i.

The challenge in applying weighted NLS regression lies in specifying the error variances. While there are some situations in which the error variance associated with a given observation might be estimated *a priori*, for example, when the y value is the mean of a set of random samples, the typical situation in watershed science and analysis is that error variances will need to be estimated or modelled. Neter et al. (1996) discussed some approaches to estimating variance functions.

Weighted least squares regression falls within the broader scope of generalized least squares (GLS) fitting. When weights are estimated from the data as part of the fitting process, which is a typical situation, the method is called estimated generalized least squares (EGLS) fitting (Gumpertz & Rawlings, 1992; Letcher et al., 2001).

A common pattern of heteroscedasticity found in environmental data is a "megaphone" shape that occurs when residuals are plotted against the fitted value, such that the magnitude of the residuals tends to increase proportionally with the fitted value; this pattern is consistent with the multiplicative error model. In this case, it is reasonable to assume that the variance of the error terms for a given value of x_i is proportional to the square of the expected value (i.e., σ_i^2 will be proportional to $[E(y_i)]^2$, which can be estimated as $\hat{y}_i^2$). This reasoning leads to an iterative approach involving the following steps (Neter et al., 1996):

1. Fit the model by applying OLS regression to log-transformed variables; extract the coefficients and store estimates of a and b as a_{old}, b_{old}.
2. Use the predicted values from this initial model to compute $\hat{y}_i$, which are predictions of the expected values.
3. Refit the NLS model using weights computed as $\hat{y}_i^{-2}$ and extract the coefficients as (a_{new}, b_{new}).
4. Compare the new and the old coefficients. If the maximum fractional difference is less than some specified tolerance (e.g., 10^{-6}), consider that the solution has converged; otherwise, compute $\hat{y}$ using the new coefficients, store the new coefficients as (a_{old}, b_{old}) and go to step 3.

The *nls* function in base R can be used to apply NLS fitting with or without specified weights. While the calculation of prediction limits is not currently implemented in the *predict* method for *nls* objects, prediction limits can be computed using the *predFit* function in the investr package (Greenwell & Kabban, 2014).

Use of generalized nonlinear least squares

An alternative to applying NLS regression with explicit iterative estimation of the weights is to apply generalized NLS (GNLS) regression, which is provided via the *gnls* function in the nlme package. In this approach, the residual errors are assumed to be independent and normally distributed with a mean equal to zero and a variance that is some function of an independent variable(s) x or the predicted response variable $\hat{y}$. Different variance models can be specified through a "weights" argument, and all model parameters are estimated via maximum likelihood.

The weights are generated by a specified variance function such as *varFixed* or *varPower*. For example, using *weights = varFixed(~ area)* would generate weights based on error variances that are proportional to the variable *area*. The function *weights = varPower( )* with no argument in *varPower( )* would generate weights based on variances that are proportional to $1/\hat{y}^{2t}$, where t is a parameter to be estimated as part of the model fitting. For the multiplicative error model, one would use *weights = varPower(fixed = list(t = 1))*, which computes the weights as $1/\hat{y}^2$.

In cases in which the errors follow a log-normal rather than a normal distribution, an alternative approach to fitting a power-law relationship would be to fit a log-normal distribution to the data using maximum likelihood, with the mean expressed as a power-law function of the x variable. However, to the author's knowledge, this approach has not been applied in the watershed science and management context and is not discussed further.

Prediction limits

In many applications, it is important to quantify the uncertainty in predicted values. If the assumptions underlying the fitted model are valid, then prediction limits can be computed for any new observation for a specified confidence level and used as a measure of precision. For example, if a confidence level of 80 percent is used to compute prediction limits when applying a model to make predictions for a catchment not used in the model fitting, then there is a probability of 0.8 that the interval would contain the new observation.

OLS on log-transformed variables

To simplify the notation, the log-transformed regression model can be expressed as $\hat{v}_i=b_0+b_1u_i$, where $v=log(y)$ and $u=log(x)$. For OLS linear regression, the lower and upper prediction limits for a new observation, u_j, can be calculated as

$$LPL_j = \hat{v}_j - t_c s_e \sqrt{1+\frac{1}{n}+\frac{(u_j-\bar{u})^2}{\sum(u_i-\bar{u})^2}} \tag{13}$$

and

$$UPL_j = \hat{v}_j + t_c s_e \sqrt{1+\frac{1}{n}+\frac{(u_j-\bar{u})^2}{\sum(u_i-\bar{u})^2}} \tag{14}$$

where LPL_j and UPL_j are the lower and upper limits, respectively, $\hat{v}_j$ is the predicted value, t_c is the two-tailed critical value of Student's t for the specified confidence level and $n–2$ degrees of freedom, s_e is the square root of the mean square error for the regression, n is the number of data points used to fit the model, and $\bar{u}$ is the mean of the values of u used to fit the regression (Neter et al., 1996). These equations are based on the assumptions that the error terms are independent, normally distributed and homoscedastic.

Prediction limits for OLS regression can be computed using the *predict* function in base R. The computed prediction limits can then be back-transformed by exponentiation. Note that the back-transformed limits will not be symmetric about the back-transformed prediction of y.

Fitting by NLS and GNLS

Equations of the form of (13) and (14) cannot be used for NLS or GNLS fits; approximate methods must be used. For example, the *predict_gnls* function in the nlraa package uses bootstrapping to generate prediction limits. In contrast, the *predFit* function in the investr package uses Taylor-series approximations to estimate the standard errors used to compute confidence limits on the predicted mean response and prediction limits for new observations based on NLS fits; see, for example, Chapter 13 in Neter et al. (1996) for more information. For fits that do not involve weights, the *predFit* function can provide lower and upper prediction limits directly. For weighted NLS fits, the *se.fit = TRUE* option should be used; prediction limits can then be computed from the values of *se.fit* and *residual.scale* that are returned by the function as follows:

$$(UPL_j, LPL_j) = \hat{y}_j \pm t_c \sqrt{s_{e,j}^2 + (s_r^2/w_j)} \tag{15}$$

where the subscript j refers to a new case for which a prediction is to be made, $s_{e,j}$ is the standard error of the fitted value for case j (taken from *se.fit*), s_r is the value of *residual.scale,* and w_j is the weight for case j. For the assumption of multiplicative errors, w_j can be approximated as $1/\hat{y}_j^2$, where $\hat{y}_j$ is computed as ax_j^b. Equation (15) is an approximation that is asymptotically valid for large n (Neter et al. 1996).

Diagnostics

It is conventional to use the residuals from the fitted relationship to assess whether the error terms deviate from normality and constant variance (Fox, 2020). For OLS on log-transformed variables,

the diagnostics refer to the log-transformed relation, not the back-transformed predictions, and are thus not relevant for assessing goodness of fit for the untransformed data.

For NLS regression without weights, the appropriate residuals are as calculated in equation (10). For weighted NLS and GNLS regression, diagnostic procedures should use Pearson residuals, computed as

$$e_{p,i} = e_i\sqrt{w_i} \tag{16}$$

where w_i is the weight applied to observation i during model fitting (Fox, 2020). Normalized residuals, which are the raw residuals scaled by $\sqrt{var(e_i)}$, are also acceptable for diagnostic purposes.

Fox (2020) provided a comprehensive overview of diagnostic procedures and plots for regression models, including approaches to detect influential points and multicollinearity when using multiple predictor variables. A parsimonious approach with a focus on the assumptions of normality, homoscedasticity, and goodness of fit is outlined in the following paragraphs.

The assumption of normality can be graphically assessed by plotting a normal quantile plot of the residuals. For normally distributed, homoscedastic errors, the residuals should fall roughly on a straight line. However, it is not uncommon for a small number of data points at the top and bottom ends of a distribution to deviate from a linear pattern in a normal quantile plot, even for data drawn from a normal distribution.

Plotting the residuals against the predicted values can provide information on both the constancy of the variance and the goodness of fit of the model form. Ideally, the residuals should have relatively even scatter above and below the value of 0, the magnitude of which should not vary with the value of the predictions. If the variance of the residuals deviates from this ideal pattern (e.g., displays a "megaphone" pattern), then the error variance would appear to be heteroscedastic, and a weighting scheme should be explored. If the residuals display a nonlinear trend, such as concave or convex curvature, an alternative model form should be explored.

Assessment of independence of the residuals depends on the type and structure of the data. For example, time series data often exhibit temporal autocorrelation, especially for sub-monthly intervals, which can be assessed using autocorrelation or partial autocorrelation plots (via the *acf* and *pacf* functions in base R). Mapping residuals for spatial data can help evaluate the presence of spatial autocorrelation or trends. Data that can be grouped into categories (e.g., season, region, vegetation, hydrologic regime) often exhibit among-group differences in error distributions, which can be visually assessed by generating box plots by category.

Including more than one predictor variable

It is often appropriate to include more than one predictor variable in a power-law model (McCuen, Leahy, & Johnson, 1990; Bywater-Reyes, Bladon, & Segura, 2018; Northwest Hydraulic Consultants, 2020). For an example with three predictor variables, the model takes the following form when using a multiplicative error model:

$$y_i = \alpha x_{1i}^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} \epsilon_i \tag{17}$$

where x_{1i}, x_{2i}, and x_{3i} represent the values of three predictor variables for case i, and β_1, β_2 and β_3 are the corresponding exponents. Equation (17) could be fit using NLS regression with weights assigned as described earlier. For fitting via OLS, the model would be expressed as:

$$log(y_i) = log(a) + b_1 log(x_{1i}) + b_2 log(x_{2i}) + b_3 log(x_{3i}) + log(\epsilon_i) \tag{18}$$

which has the form of a multiple regression model. As with the case of one predictor variable, the back-transformation by exponentiation results in a bias.

For additive errors, a power-law with three predictors could be expressed as:

$$y_i = \alpha x_{1i}^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} + \epsilon_i \tag{19}$$

The use of multiple predictors introduces the need for more complex diagnostic and model testing procedures, which are beyond the scope of this article. Therefore, the focus here is on the case of one predictor variable.

Description of the data

This analysis focuses on the relationship between a flood quantile and catchment area, which is commonly modelled as a power-law function, in the Canadian portion of the Columbia River basin. The flood quantile of interest is the annual maximum instantaneous flow with a return period of two years, denoted Q_2. Note that this application is intended purely as an illustrative exercise; the resulting models should not be applied in a real-world context.

The analysis includes the set of catchments used in the analysis presented by Moore et al. (2020), which were selected based on the following criteria: 1) currently active gauging station, 2) gauging station established prior to 1977, 3) unregulated flow, 4) drainage area less than 10,000 km^2, and 5) at least 36 years of data from 1977 to 2020. Shape files containing catchment boundaries and a file containing metadata about the gauges and catchments are available via an online repository.[1] The locations of the catchments are shown in Figure 1.

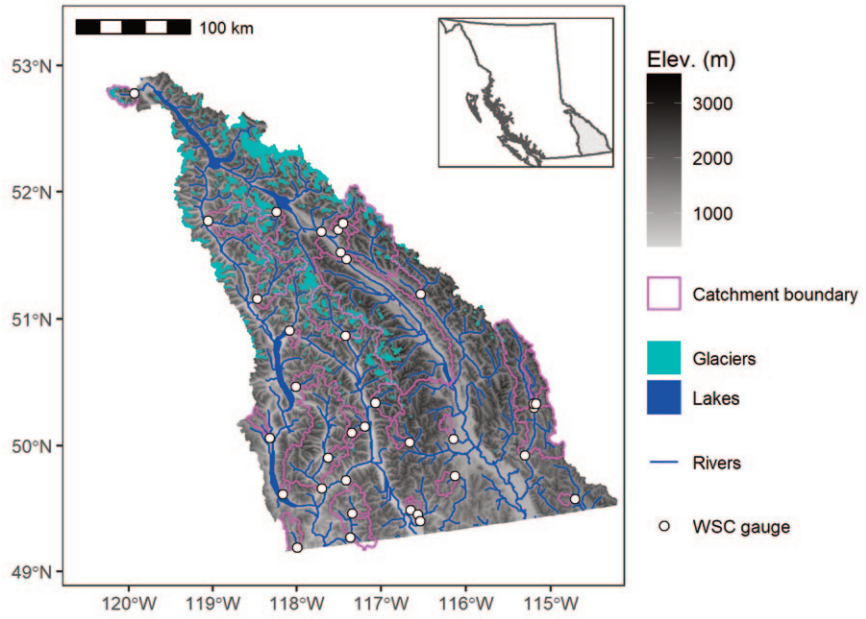

Figure 1. Locations of catchments within the Canadian portion of the Columbia River headwaters. The inset map shows the location of the Columbia River basin within the Province of British Columbia.

Annual maximum instantaneous flows were accessed from the HYDAT hydrometric database using the *hy_annual_instant_peaks* function in the tidyhydat package. Note that instantaneous peak flows were not available for all years of record. The minimum number of years of peak flow values was 18, but the rest had 30 to 43 years of peak flow values.

For each gauge, the two-year flood, denoted Q_2, was computed using the *fevd* and *return.level* functions in the extRemes package for data from 1977 to the end of the record. Flood frequency curves were fit using a generalized extreme value distribution with maximum likelihood estimation.

As seen in Figure 2, the data indicate a nonlinear relationship when plotted using log-log scales, suggesting the presence of a scale break in the relationship between Q_2 and catchment area. The relationship appears to be reasonably linear in log-log space for areas less than 1000 km^2, suggesting that a power-law relationship would be valid over that range; the analysis was accordingly restricted to that range of areas.

Fitted models

As seen in Figure 3(a), the data appear to follow a linear relationship in log-log space for the restricted range of catchment areas. The fit appears to be reasonably tight, with $r^2 = 0.9$. Note, however, that this statistic is for the log-transformed linear relation, not the back-transformed power law, which appears to exhibit substantial scatter, especially in the range from 200 to 600 km^2 (figure 3b)).

Moore. (2024). Fitting Power-Law Relationships in Watershed Science and Analysis, with an Example Using the R Language. http://confluence-jwsm.ca /index.php/jwsm/article/view/53/7 . doi:10.22230/jwsm.2024v7n1a53

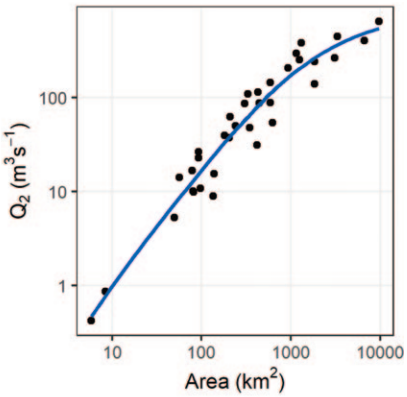

Figure 2: Log-log plot of relationship between two-year flood and catchment area. The fitted curve is a locally weighted scatterplot smoother (loess), with span = 0.75 and degree = 2. For explanation of loess smoothers, see National Institute of Standards and Technology (2012).

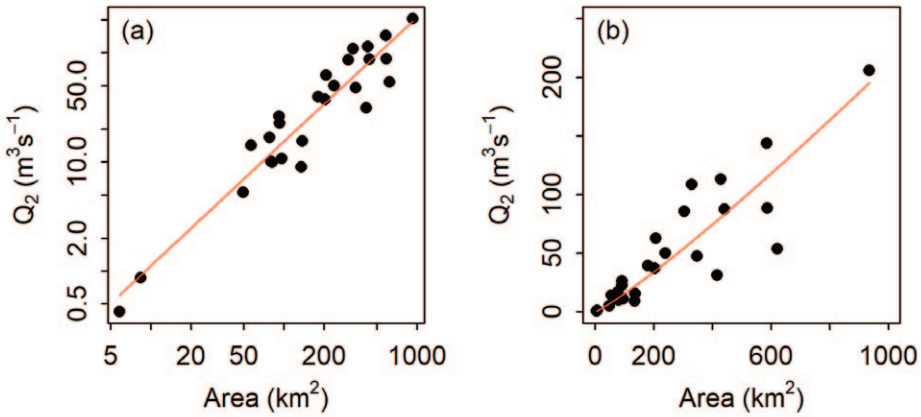

Figure 3: Scatterplots of the two-year flood as a function of catchment area with (a) logarithmic scales and (b) arithmetic scales. The red line in (a) is the linear relationship based on the OLS fit to the log-transformed variables, and the red curve in (b) is the back-transformed power-law relationship.

Transforming the predictor variable modifies the leverage of each point used to fit the data and thus their influence on the model fit. For example, in Figure 3(a), the data points corresponding to the two smallest catchments have leverage values that are more than two times the mean leverage, and thus are potentially influential points (Neter et al., 1996). However, for the untransformed data shown in Figure 3(b), the point corresponding to the largest catchment is the only one with "high" leverage.

The value of the smearing correction is 1.097; that is, the bias-corrected relationship is about 10 percent higher than the uncorrected relationship. As seen in Figure 4, the bias-corrected OLS fit, the weighted NLS fit, and the GNLS fit are visually similar through the full range of catchment areas; however, the unweighted NLS fit deviates from the other three, with higher predictions for smaller catchments and lower predictions for larger catchments.

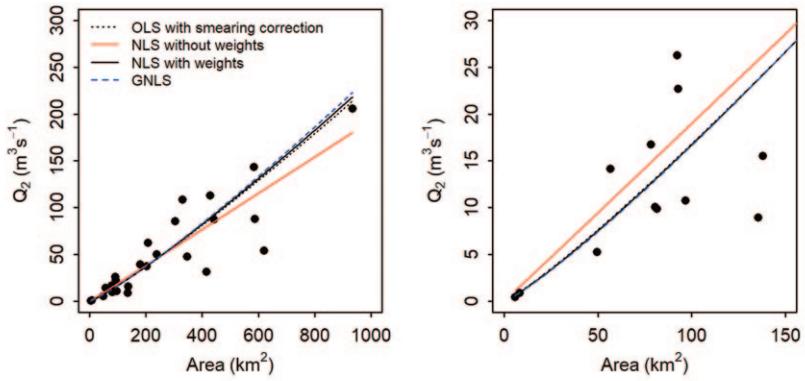

Figure 4: Comparison of the relationships generated by OLS fitting with bias correction, NLS without weights, NLS with weights assuming a multiplicative error model, and GNLS with **weights** $= \hat{y}^{-2t}$*. The left-hand panel shows the full data set while the right-hand panel is restricted to catchment areas less than 150 km^2.*

Table 1 summarizes the fitted coefficients from the various approaches. Note that the bias correction only modifies the value of *a*. The unweighted NLS fit generated a value of *b* that is about 15 percent lower than the other approaches and a value of *a* that is roughly double the other estimates. The GNLS fit with *weights = varPower()* generated a value of t = 1.029, which corresponds to weights computed

Table 1: Summary of fitted values of *a* and *b*.

Method	Weights	*a*	*b*
OLS	none	0.0811	1.138
OLS with smearing correction	none	0.0889	1.138
NLS	none	0.1836	1.007
NLS	$\hat{y}^{-2}$	0.0838	1.150
GNLS	$\hat{y}^{-2t}$	0.0791	1.162
GNLS	$\hat{y}^{-2}$	0.0851	1.147

Notes: GNLS, generalized nonlinear least squares; NLS, nonlinear least squares; OLS, ordinary least squares.

as $1/\hat{y}$ raised to the 2.058 power, which is only 2.9 percent greater than the exponent that is consistent with the multiplicative error model. Interestingly, the bias-corrected OLS fit, the weighted NLS fit, and the GNLS fit generated somewhat different values of *a* and *b* despite the visual similarity of the fitted relationships (Figure 4).

As seen in Figure 5, the fitted relationships and 80 percent prediction limits appear broadly similar for the fits generated by bias-corrected OLS, weighted NLS, and GNLS, although, as noted earlier, the prediction limits are not symmetric around the fitted relationship for the OLS fit. The unweighted NLS fit does not account for heteroscedastic error variance. Accordingly, the prediction limits do not align with the tendency to increasing variance with increasing value of $\hat{y}$. The noisy fits and prediction limits for the GNLS fit represent variability of the bootstrapping method, which was applied for a discrete sequence of x_j values. For all fits, it is important to recognize how wide the limits are, even for a relatively relaxed confidence level of 0.8 (rather than a more commonly used level of 0.9 or 0.95).

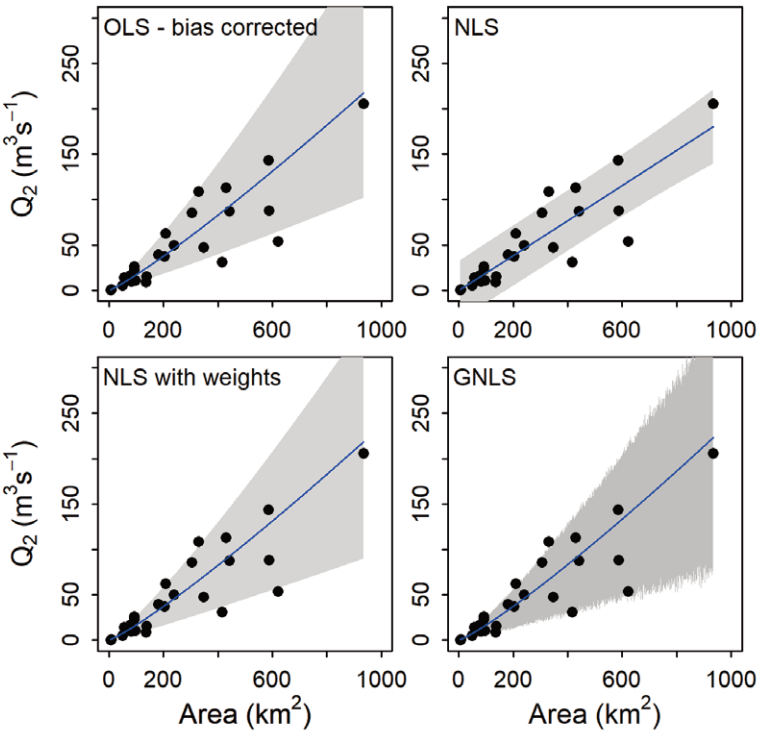

Figure 5: Comparison of the fitted power-law relationships using four approaches, with 80% prediction limits shown as a grey polygon. The GNLS fit is for weights = $\hat{y}^{-2t}$.

Model diagnostics

Figure 6 shows diagnostic plots for NLS regression without weights and for the GNLS fit with *weights = varPower()* (i.e., weights proportional to $\hat{y}^{-2t}$). The variance of the residuals for the NLS fit increase with the predicted value, suggestive of a megaphone shape, which indicates a violation of the constant variance assumption. The normal quantile plot reveals substantial deviations from linearity in both tails, which suggests that the normality assumption is also violated. Therefore, unweighted NLS does not appear to be an appropriate method for fitting a power-law relationship to this data set.

For the GNLS fit, the magnitude of scatter does not exhibit any obvious trends in relation to the predicted value of Q_2. The normal quantile plot indicates that, except for the tails, the distribution appears reasonably consistent with a normal distribution. The diagnostic plots for the weighted NLS fit and the GNLS fit with *t* fixed at a value of 1 were similar to those for the GNLS fit shown in

Figure 6. Compared with the case for the unweighted NLS fit, the diagnostic plots support the use of weights when fitting the power-law relationship.

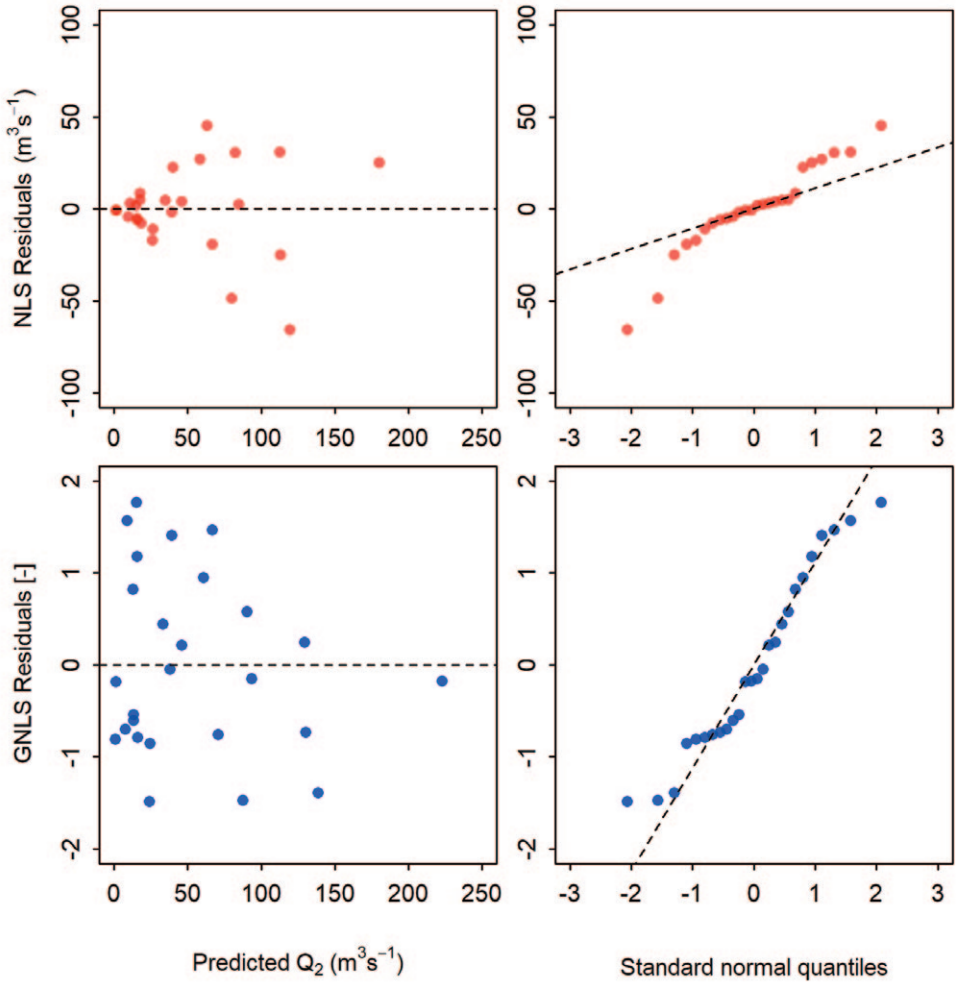

Figure 6: Diagnostic plots for NLS regression without weights (top row, red symbols) and GNLS regression with* weights = $\hat{y}^{-2t}$ *(bottom row, blue symbols). The left-hand column shows the residuals versus fitted values of Q_2; the right-hand column shows normal quantile plots. The bottom row shows Pearson residuals; see text for explanation. The dashed line in the normal quantile plots passes through the 25th and 75th quantiles, and provides a visual reference to assess linearity.

Discussion

As outlined in the introduction, power-law relationships are commonly assumed to hold in many applications in watershed science and analysis. As demonstrated through this case study, the approaches to fitting power-law relationships can differ both in the predicted values and in the estimated prediction limits, and it is important for analysts to examine diagnostic plots to choose the most appropriate method for each data set.

Examination of a scatterplot with logarithmic axes can be a useful first step in fitting a power-law relation. If such a plot indicates a deviation from linearity, as was the case for the example used here, then it may be necessary to restrict the model to a range in which the data points exhibit a linear trend. Alternatively, one could consider a different form of nonlinear model that is more appropriate for the data, perhaps including additional predictor variables.

Many studies that used OLS regression with log-transformed variables to fit a power-law model provided a value of r^2 as a measure of goodness of fit (e.g., Eaton, Church, & Ham, 2002; Adhikari & Marshall, 2012; Bahr, Pfeffer, & Kaser, 2015; Chiu et al., 2016; Gronsdahl et al., 2019; Northwest Hydraulic Consultants, 2020). However, r^2 refers to the goodness of fit in log-log space, and may be misleading about the goodness of fit after back-transforming the fitted linear relation, as was

apparent from the contrast in fit seen in Figure 3, and as was noted by Northwest Hydraulic Consultants (2020).

In the example presented here, the bias correction represented an increase of about 10 percent over the uncorrected relationship. The bias-corrected relationship was visually similar to the weighted NLS and GNLS fits, albeit with different parameter estimates. This observation highlights the need for caution when the focus of a study is to quantify the value of the exponent—for example, to compare empirical estimates with theoretical values—as argued by Mark and Church (1977) based on other considerations.

In the application of NLS regression with weights and GNLS with t fixed at a value of 1, it was assumed *a priori* that a multiplicative error model was appropriate; the suggestion of a megaphone pattern in the plot of residuals versus predicted values for the unweighted NLS fit is consistent with that assumption (Figure 6). However, a multiplicative error model may not be applicable in all cases, and an alternative model may be needed. In some cases, the choice of variance model might be guided by theoretical concerns and/or by interpretion of diagnostic plots. A detailed discussion of diagnosing and modelling non-constant variance is beyond the scope of this work, but further information can be found in Neter et al. (1996) and Fox (2020).

Conclusions

Four main approaches are available to fit power-law relationships to data: 1) application of OLS regression following logarithmic transformation of both the y and x variables, 2) unweighted NLS regression, 3) NLS regression with weights, and 4) GNLS regression with weights computed based on a variance model that is estimated as part of the model fitting. As demonstrated through the case study, these methods can yield different predictive models and prediction limits. Some recommendations for fitting power-law relationships in watershed science and analysis follow.

When fitting a power-law model, it is recommended that OLS regression on log-transformed variables should be used primarily as an initial visual check on the appropriateness of a power-law relationship and to generate starting values for parameter estimates for NLS and GNLS fitting. OLS regression on log-transformed variables is implicitly based on a multiplicative error model, which may not be valid in all cases; in addition, back-transformed predictions are subject to bias, and back-transformed prediction limits will not be symmetric about the predicted value, which means that expression of uncertainty requires explicit specification of both the lower and upper limits. Further, the value of r^2 for the OLS regression on log-transformed variables should not be used as an indicator of goodness of fit for the power-law relationship.

If OLS regression is used to fit a predictive model, back-transformed predictions should be bias-corrected and appropriate diagnostic procedures performed to check that the residuals from the log-transformed fit (equation 2) are consistent with the assumptions of normality and constant variance.

It is recommended that NLS, weighted NLS or GNLS be used to fit power-law relations, depending on the nature of the error variance as determined empirically by regression diagnostics and/or based on theoretical considerations. Where homoscedasticity of the error variance appears to be valid, NLS fitting could be used. Where heteroscedasticity is an issue and the variance can be modelled as a function of the predictors or the fitted values, GNLS may be preferred to weighted NLS because it does not require explicit iteration, and the computation of prediction limits is more straightforward in terms of coding.

For visualizing the precision of the predictive relation, it is recommended that the data be plotted on arithmetic axes, along with prediction limits. Plotting the relationship separately for the lower range of values can be useful for interpreting the fit across the full range of data.

Data availability

All files associated with this study are publicly available via https://zenodo.org/records/10071699.

Acknowledgements

Professor Val Lemay (The University of British Columbia), three anonymous reviewers, and *Confluence* editor Robin Pike provided detailed comments on earlier versions of this manuscript,

which resulted in substantial improvements. However, any errors and/or omissions are the sole responsibility of the author.

Thanks are extended to the developers of contributed R packages that were used in the analysis and production of this document: bcmaps (Teucher, Hazlitt, & Albers, 2021), broom (Robinson, Hayes, & Couch, 2022), dplyr (Wickham et al., 2021), extRemes (Gilleland & Katz, 2016), ggnewscale (Campitelli, 2022), ggplot2 (Wickham, 2016), ggspatial (Dunnington, 2021), here (Müller, 2020), httr (Wickham, 2020), investr (Greenwell & Kabban, 2014), lubridate (Grolemund & Wickham, 2011), magrittr (Bache & Wickham, 2020), nlme (Pinheiro et al., 2021), nlraa (Miguez, 2022), nlstools (Baty et al., 2015), sf (Pebesma, 2018), quarto (Allaire, 2022), raster, readxl (Wickham & Bryan, 2019), rnaturalearth (South, 2017), tidyhydat (Albers, 2017), tidyr (Wickham, 2021) and whitebox (Lindsay, 2016).

Note

1. Available at https://zenodo.org/record/3779279.

References

Adhikari, S., & Marshall, S.J. 2012. Glacier volume-area relation for high-order mechanics and transient glacier states. *Geophysical Research Letters 39*(16):1–6. doi:10.1029/2012gl052712

Albers, S. 2017. Tidyhydat: Extract and tidy Canadian hydrometric data. *The Journal of Open Source Software 2*(20):511. doi:10.21105/joss.00511

Allaire, J.J. 2022. *Quarto: R interface to 'quarto' markdown publishing system*. https://CRAN.R-project.org/package=quarto [February 2, 2024].

Bache, S.M., & Wickham, H. 2020. *Magrittr: A forward-pipe operator for R*. https://CRAN.R-project.org/package=magrittr [February 2, 2024].

Bahr, D.B., Pfeffer, W.T., & Kaser, G. 2015. A review of volume-area scaling of glaciers. *Reviews of Geophysics 53*(1):95–140. doi:10.1002/2014rg000470

Baskerville, G.L. 1972. Use of logarithmic regression in the estimation of plant biomass. *Canadian Journal of Forest Research 2*(1):49–53. doi:10.1139/x72-009

Basu, N.B., Destouni, G., Jawitz, J.W., Thompson, S.E., Loukinova, N.V., Darracq, A., et al. 2010. Nutrient loads exported from managed catchments reveal emergent biogeochemical stationarity. *Geophysical Research Letters 37*(23). doi:10.1029/2010gl045168

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J.-P., & Delignette-Muller, M.-L. 2015. A toolbox for nonlinear regression in R: The package nlstools. *Journal of Statistical Software 66*(5):1–21. doi:10.18637/jss.v066.i05

Boiten, W. 1993. Flow-measuring structures. *Flow Measurement and Instrumentation 4*(1):17–24. doi:10.1016/0955-5986(93)90006-5

Bradu, D., & Mundlak, Y. 1970. Estimation in lognormal linear models. *Journal of the American Statistical Association 65*(329):198–211. doi:10.1080/01621459.1970.10481074

Brardinoni, F., & Hassan, M.A. 2006. Glacial erosion, evolution of river long profiles, and the organization of process domains in mountain drainage basins of coastal British Columbia. *Journal of Geophysical Research 111*(F1). doi:10.1029/2005jf000358

Bywater-Reyes, S., Bladon, K.D., & Segura, C. 2018. Relative influence of landscape variables and discharge on suspended sediment yields in temperate mountain catchments. *Water Resources Research 54*(7):5126–5142. doi:10.1029/2017wr021728

Campitelli, E. 2022. *ggnewscale: Multiple fill and colour scales in "ggplot2."* https://CRAN.R-project.org/package=ggnewscale [February 2, 2024].

Chen, B., Ma, C., Krajewski, W.F., Wang, P., & Ren, F. 2020. Logarithmic transformation and peak-discharge power-law analysis. *Hydrology Research 51*(1):65–76. doi:10.2166/nh.2019.108

Chiu, C.-W., Komatsu, H., Katayama, A., & Otsuki, K. 2016. Scaling-up from tree to stand transpiration for a warm-temperate multi-specific broadleaved forest with a wide variation in stem diameter. *Journal of Forest Research 21*(4):161–169. doi:10.1007/s10310-016-0532-7

Cohn, T.A., Delong, L.L., Gilroy, E.J., Hirsch, R.M., & Wells, D.K. 1989. Estimating constituent loads. *Water Resources Research 25*(5):937–942. doi:10.1029/wr025i005p00937

Duan, N. 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association 78*(383):605–610. doi:10.1080/01621459.1983.10478017

Dunnington, D. 2021. *ggspatial: Spatial data framework for ggplot2.* https://CRAN.R-project.org/package=ggspatial [February 2, 2024].

Eaton, B.C., & Davidson, S.L. 2022. Hydraulic geometry: Empirical investigations and theoretical approaches. In J.F. Shroder (Ed.), *Treatise on Geomorphology*, (2nd ed.). Elsevier, pp. 461–479. doi:10.1016/B978-0-12-818234-5.00002-X

Eaton, B.C., Church, M., & Ham, D. 2002. Scaling and regionalization of flood flows in British Columbia, Canada. *Hydrological Processes 16*(16):3245–3263. doi:10.1002/hyp.1100

Ferguson, R.I. 1986. River loads underestimated by rating curves. *Water Resources Research 22*(1):74–76. *doi:*10.1029/wr022i001p00074

Fox, J.D. 2020. *Regression diagnostics*. SAGE Publications, Thousand Oaks, CA. doi:10.4135/9781071878651

Gardiner, V., & Park, C.C. 1978. Drainage basin morphometry: Review and assessment. *Progress in Physical Geography: Earth and Environment 2*(1):1–35. doi:10.1177/030913337800200102

Gilleland, E., & Katz, R.W. 2016. extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software 72(*8):1–39. doi:10.18637/jss.v072.i08

Gleason, C.J. 2015. Hydraulic geometry of natural rivers. *Progress in Physical Geography: Earth and Environment 39*(3):337–360. doi:10.1177/0309133314567584

Greenwell, B.M., & Kabban, C.M.S. 2014. Investr: An R package for inverse estimation. *The R Journal 6*(1):90–100. https://doi.org/10.32614/RJ-2014-009

Grolemund, G., & Wickham, H. 2011. Dates and times made easy with lubridate. *Journal of Statistical Software 40*(3):1–25. https://www.jstatsoft.org/v40/i03/ [February 2, 2024].

Gronsdahl, S., Moore, R.D., Rosenfeld, J., McCleary, R., & Winkler, R. 2019. Effects of forestry on summertime low flows and physical fish habitat in snowmelt-dominant headwater catchments of the Pacific Northwest. *Hydrological Processes 33*(25):3152–3168. doi:10.1002/hyp.13580

Gumpertz, M.L., & Rawlings, J.O. 1992. Nonlinear regression with variance components: Modeling effects of ozone on crop yield. *Crop Science 32*(1):219–224. doi:10.2135/cropsci1992.0011183x003200010045x

Knapp, J.L.A., von Freyberg, J., Studer, B., Kiewiet, L., & Kirchner, J.W. 2020. Concentration–discharge relationships vary among hydrological events, reflecting differences in event characteristics. *Hydrology and Earth System Sciences 24*(5):2561–2576. doi:10.5194/hess-24-2561-2020

Letcher, R.A., Schreider, S.Yu., Jakeman, A.J., Neal, B.P., & Nathan, R.J. 2001. Methods for the analysis of trends in streamflow response due to changes in catchment condition. *Environmetrics 12*(7):613–630. *doi:*10.1002/env.486

Lindsay, J.B. 2016. Whitebox GAT: A case study in geomorphometric analysis. *Computers and Geosciences 95*:75–84. doi:10.1016/j.cageo.2016.07.003

Mark, D.M., & Church, M. 1977. On the misuse of regression in earth science. *Journal of the International Association for Mathematical Geology 9*(1):63–75. doi:10.1007/bf02312496

Marwick, B., Boettiger, C., & Mullen, L. 2018. Packaging data analytical work reproducibly using R (and friends). *The American Statistician 72*(1):80–88. doi:10.1080/00031305.2017.1375986

McCuen, R.H., Leahy, R.B., & Johnson, P.A. 1990. Problems with logarithmic transformations in regression. *Journal of Hydraulic Engineering 116*(3):414–428. doi:10.1061/(asce)0733-9429(1990)116:3(414)

Miguez, F. 2022. *nlraa: Nonlinear regression for agricultural applications.* https://CRAN.R-project.org/package=nlraa_[February 2, 2024].

Moore, R.D., & Hutchinson, D. 2017. Why watershed analysts should use R for data processing and analysis. *Confluence: Journal of Watershed Science and Management 1*(1). doi:10.22230/jwsm.2017v1n1a2

Moore, R.D., Pelto, B., Menounos, B., & Hutchinson, D. 2020. Detecting the effects of sustained glacier wastage on streamflow in variably glacierized catchments. *Frontiers in Earth Science 8*. doi:10.3389/feart.2020.00136

Müller, K. 2020. *here: A simpler way to find your files*. https://CRAN.R-project.org/package=here_ [February 2, 2024].

National Institute of Standards and Technology. 2012. *e-Handbook of Statistical Methods.* https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm [February 2, 2024].

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. 1996. *Applied linear statistical models, (fourth edition)*. McGraw-Hill/Irwin, New York, NY.

Northwest Hydraulic Consultants. 2020. *British Columbia extreme flood project: Regional flood frequency analysis.* North Vancouver, B.C.

Pebesma, E. 2018. Simple features for R: Standardized support for spatial vector data. *The R Journal 10*(1):439–446. doi:10.32614/RJ-2018-009

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. 2021. *nlme: Linear and nonlinear mixed effects models.* https://CRAN.R-project.org/package=nlme [February 2, 2024].

Robinson, D., Hayes, A., & Couch, S. 2022. *broom: Convert statistical objects into tidy tibbles.* https://CRAN.R-project.org/package=broom [February 2, 2024].

Shook, K., & Gray, D.M. 1997. Synthesizing shallow seasonal snow covers. *Water Resources Research 33(3):*419–426. doi:10.1029/96wr03532

Shook, K., Whitfield, P., Chlumsky, R., Moore, D., & Durocher, M. 2022. CSHShydRology: Canadian hydrological analyses. https://CRAN.R-project.org/package=CSHShydRology [February 2, 2024].

Slater, L.J., Thirel, G., Harrigan, S., Delaigue, O., Hurley, A., Khouakhi, A., Prosdocimi, I., Vitolo, C., & Smith, K. 2019. Using R in hydrology: A review of recent developments and future directions. *Hydrology and Earth System Sciences 23*(7):2939–2963. doi:10.5194/hess-23-2939-2019

South, A. 2017. *Rnaturalearth: World map data from natural earth.* https://CRAN.R-project.org /package=rnaturalearth [February 2, 2024].

Teucher, A., Hazlitt, S., & Albers, S. 2021. *bcmaps: Map layers and spatial utilities for British Columbia.* https://CRAN.R-project.org/package=bcmaps [February 2, 2024].

Wickham, H. 2016. *ggplot2: Elegant graphics for data analysis.* https://ggplot2.tidyverse.org [February 2, 2024].

Wickham, H. 2020. *httr: Tools for working with URLs and HTTP.* https://CRAN.R-project.org /package=httr [February 2, 2024].

Wickham, H. 2021. *tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr [February 2, 2024].

Wickham, H., & Bryan, J. 2019. *readxl: Read excel files.* https://CRAN.R-project.org/package=readxl [February 2, 2024].

Wickham, H., François, R., Henry, L., & Müller, K. 2021. *dplyr: A grammar of data manipulation.* https://CRAN.R-project.org/package=dplyr [February 2, 2024].